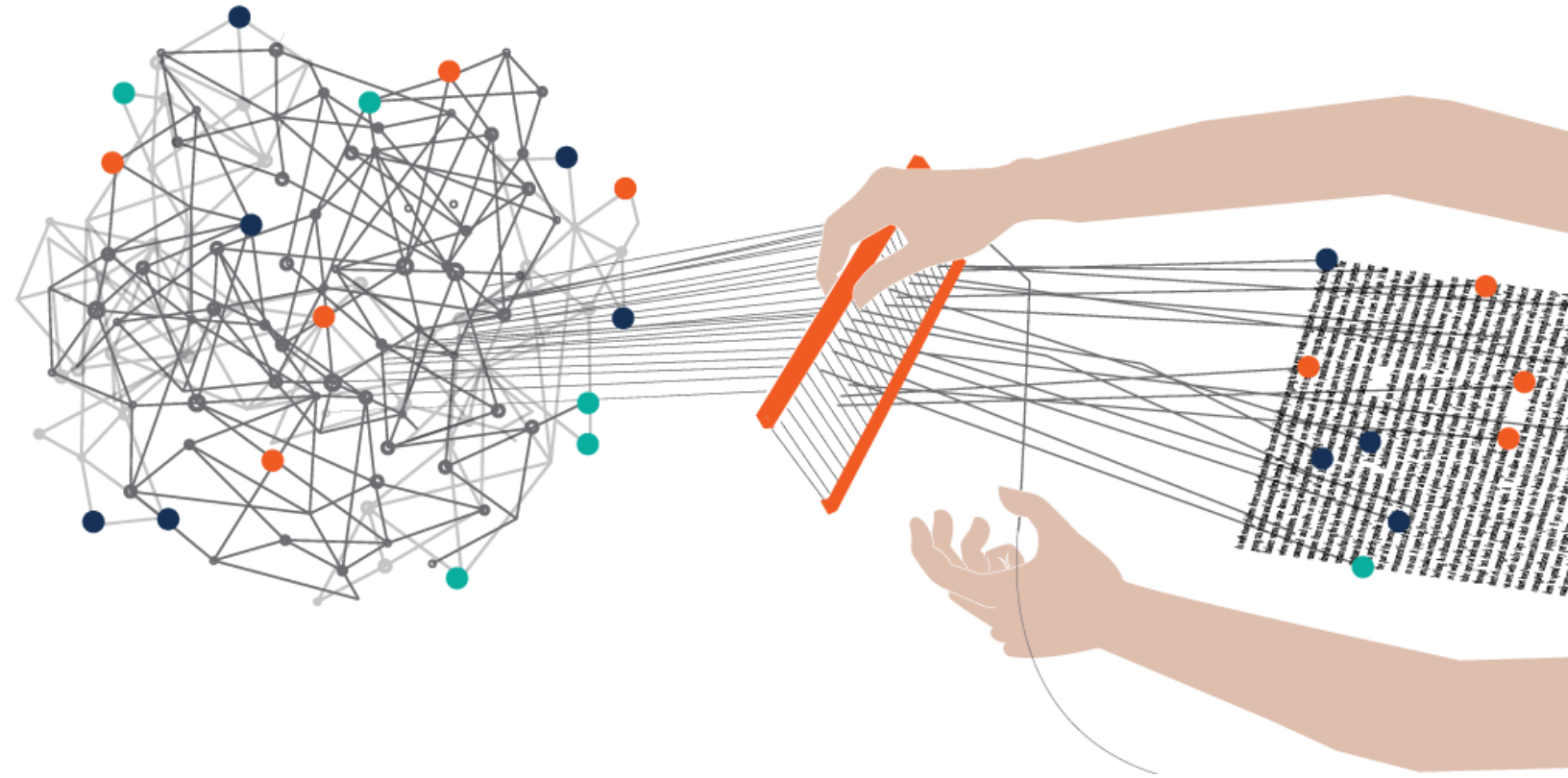

Análisis de la discriminación lingüística en aplicaciones de Procesamiento de Lenguaje Natural: un enfoque en la dominación del inglés frente a otros idiomas.

Gerardo Martínez Cruz

Programa Líderes de LACNIC 2023

Noviembre, 2023.

Gerardo Martínez Cruz

gerardo.martinezcruz@studio.unibo.it

Gerardo es Ingeniero en Telecomunicaciones por la Universidad Nacional Autónoma de México (UNAM). Complementando su formación académica, posee un título de Maestría en Derecho de las TIC por Infotec y actualmente está cursando un Maestría en *Digital Technologies and Innovation Management* en la *Bologna Business School*. Sus áreas de experiencia en investigación de vanguardia abarcan dominios críticos como la Gobernanza de Internet, Inteligencia Artificial, *Machine Learning*, y tecnologías del espectro. Más allá de la academia, Gerardo aporta sus conocimientos y perspectivas al ámbito profesional. Actualmente, trabaja en México como asesor de políticas públicas de ingeniería y tecnología en el Instituto Federal de Telecomunicaciones, donde aplica su experiencia para dar forma y promover el panorama de las telecomunicaciones..

Los puntos de vista y opiniones expresados en esta investigación pertenecen al autor y no necesariamente reflejan la política o posición oficial de LACNIC.

Índice

1. Introducción

En el ámbito de la tecnología moderna, el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) se erige como una disciplina fundamental con amplias implicaciones. Es indispensable por numerosas razones que abarcan amplios dominios, permeando la manera en que los seres humanos interactúan con las computadoras, procesan datos y obtienen conocimientos a partir de las vastas cantidades de texto no estructurado en el mundo digital.

La rápida proliferación de aplicaciones de NLP ha traído consigo indudablemente una nueva era de interacción tecnológica, transformando fundamentalmente la forma en que nos relacionamos en el mundo digital. Estas herramientas sofisticadas, respaldadas por algoritmos complejos y aprendizaje automático, se han integrado perfectamente en nuestra vida cotidiana, permitiéndonos hacer búsquedas en internet, comunicarnos con asistentes de voz y superar las barreras del idioma con facilidad. Sin embargo, bajo la superficie de esta maravilla tecnológica yace un problema significativo: la presencia de una discriminación indirecta perpetuada por estas mismas aplicaciones.

A diferencia de los sesgos más evidentes observados en el NLP, como el sesgo de género o racial, que se manifiestan como resultados evidentes en las aplicaciones de NLP, el sesgo lingüístico se revela como un obstáculo inicial e indirecto que enfrentan las personas que buscan aprovechar el poder del NLP. Esta forma única de sesgo se caracteriza por sus orígenes indirectos, que se derivan del efecto de red poderoso y de la histórica dominación del idioma inglés en el ámbito digital.

Dicho esto, esta investigación tiene como objetivo brindar una exploración exhaustiva del NLP y aclarar su papel fundamental en nuestra era digital contemporánea. Además, profundiza en las complejidades de la discriminación lingüística dentro del ámbito del NLP, atribuida principalmente a la abrumadora dominancia del idioma inglés. Por último, la investigación presenta una serie de estrategias y enfoques destinados a mitigar este sesgo lingüístico en el campo del NLP.

2. ¿Qué es NLP y cómo funciona?

2.1 Definición de NLP

"El lenguaje natural" se refiere a una forma de comunicación utilizada en las interacciones cotidianas entre los seres humanos, como el español, el inglés o el italiano. A diferencia de los lenguajes artificiales, como los códigos de programación y los símbolos matemáticos, los lenguajes naturales se desarrollan y cambian a lo largo de las generaciones, lo que los hace difíciles de definir con reglas precisas [1].

En nuestro contexto digital, el Procesamiento de Lenguaje Natural, o NLP por sus siglas en inglés (*Natural Language Processing*), se utiliza ampliamente para abarcar cualquier manipulación informática del lenguaje natural. Esta manipulación puede abarcar desde tareas simples como contar frecuencias de palabras para comparar estilos de escritura, o hasta esfuerzos más complejos como comprender declaraciones humanas completas lo suficientemente bien como para proporcionar respuestas significativas.

En pocas palabras, NLP puede definirse como una tecnología de *Machine Learning* (que a su vez pertenece al área de la Inteligencia Artificial) que dota a las computadoras de la capacidad de interpretar, manipular y comprender el lenguaje humano, ya sea en forma de texto o lenguaje hablado, de manera muy similar a como lo hacen los seres humanos[2].

NLP integra los principios de la lingüística computacional, que abarca enfoques basados en reglas para el lenguaje humano, con técnicas estadísticas, de *Machine Learning* y de *Deep Learning*. Esta amalgama de enfoques capacita a las computadoras para analizar el lenguaje humano, ya sea en forma de texto o habla, comprendiendo su totalidad, incluyendo las intenciones subyacentes y las emociones transmitidas por el hablante o escritor (comúnmente llamadas sentimientos)[3].

Ejemplos de aplicaciones de NLP son la traducción de texto de un idioma a otro (como *Google Translate* o *Reverso*), el software de dictado de voz a texto, asistentes digitales basados en comandos hablados (como *Alexa*, *Siri* o *Google Assistant*), y la capacidad de resumir grandes volúmenes de texto de manera rápida, incluso en tiempo real. Una tecnología que está a la vanguardia y que combina todas estas técnicas es la famosa aplicación llamada ChatGPT[4].

---

[1] Steven Bird, Ewan Klein, y Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009: ix

[2] IBM. *What is natural language processing (NLP)?* 2023. Disponible en: https://www.ibm.com/topics/natural-language-processing

[3] Idem.

[4] ChatGPT es una variante del modelo de lenguaje GPT (*Generative Pre-trained Transformer*), diseñada específicamente para generar texto similar al humano y participar en conversaciones. Esta tecnología es capaz de generar respuestas coherentes y contextualmente relevantes a entradas de texto, lo que la hace útil para chatbots, asistentes virtuales y otras tareas de procesamiento de lenguaje natural que implican generar texto similar al humano en respuesta a consultas o indicaciones de los usuarios.

En nuestra sociedad digital, ya sea de manera directa o indirecta, nos guste o no, todos estamos inmersos en la ola de desarrollo del NLP, la cual parece ser imparable.

2.2 ¿Por qué NLP importa en la sociedad digital?

En el rápido y cambiante panorama de la tecnología moderna, NLP ha surgido como una disciplina de suma importancia, con profundas y amplias implicaciones que repercuten en diversos campos. Ha remodelado la dinámica fundamental de la interacción entre humanos y computadoras, el procesamiento de datos y la extracción de conocimiento de las vastas cantidades de datos textuales no estructurados.

La prominencia del NLP se puede discernir por su capacidad para forjar una interfaz armoniosa y transformadora entre individuos y máquinas. A diferencia de las interacciones convencionales entre humanos y computadoras que generalmente requieren experiencia en lenguajes de programación intrincados, **el NLP otorga a los usuarios el poder de interactuar con sistemas computacionales a través del lenguaje natural.** Este enfoque, similar a nuestra comunicación interpersonal, cierra la brecha previa que separaba a los humanos de las máquinas, abriendo una nueva era de accesibilidad sin igual y de participación intuitiva en el ámbito de la interacción entre humanos y computadoras.

Las profundas implicaciones de este cambio de paradigma son multifacéticas. Al evitar la necesidad de que los usuarios se adapten a la sintaxis y semántica rígida de los lenguajes de programación, el **NLP democratiza el acceso a la tecnología**. **Empodera a personas** de diversos orígenes y niveles de habilidad para aprovechar sin esfuerzo el potencial de sistemas computacionales complejos. Esta democratización de la tecnología trasciende barreras, **fomentando la inclusión** y ampliando el círculo de beneficiarios para abarcar no solo a tecnólogos expertos, sino también a personas de dominios no técnicos.

Además, la inclinación innata del ser humano hacia el lenguaje natural **convierte al NLP en un canal intuitivo para la comunicación con las máquinas**. Esta interacción simbiótica entre humanos y computadoras, mediada a través del lenguaje, crea un entorno amigable para el usuario en el que la tecnología se convierte en una extensión fluida del pensamiento y la expresión humana. Los usuarios pueden dar órdenes, buscar información y transmitir solicitudes matizadas de manera que refleja sus conversaciones cotidianas, reduciendo así la carga cognitiva asociada a la interacción con la tecnología.

Más allá de mejorar la accesibilidad, el NLP cataliza avances transformadores en el procesamiento de datos y la extracción de conocimiento. La proliferación de fuentes de datos textuales no estructurados, que van desde publicaciones en redes sociales hasta literatura científica, plantea un desafío enorme en la extracción de conocimientos significativos. El NLP nos dota de las herramientas para navegar eficientemente en esta oleada de datos. Proporciona a las máquinas la **capacidad de comprender, categorizar y generar conocimiento** a partir de texto o sonidos no estructurados, un logro que sería

insuperable mediante técnicas convencionales de procesamiento de datos, y que solo con recursos humanos llevaría un tiempo inmedible tan solo para obtener poco conocimeinto[5].

El NLP es un campo versátil con aplicaciones en diversos dominios. Algunas de las áreas en las que se puede aplicar el NLP se muestran en la Tabla 1.

Tabla 1. Aplicaciones de NLP.	
Manejo de Datos	Una gran cantidad de los datos en el mundo se presenta en forma de texto no estructurado. El NLP ayuda a extraer conocimientos valiosos, sentimientos e información de estos datos, facilitando su análisis y utilización.
Automatización	El NLP es esencial para automatizar diversas tareas que implican el procesamiento de datos de texto. Esto incluye chatbots para el soporte al cliente, generación automatizada de contenido y resumen de texto. Estas aplicaciones ahorran tiempo y recursos.
Búsqueda de Información	Los motores de búsqueda, los sistemas de recomendación y el etiquetado de contenido dependen del NLP para comprender mejor las consultas de los usuarios y las descripciones de contenido, ofreciendo resultados y sugerencias más relevantes.
Traducción	El NLP desempeña un papel crítico en la traducción automática, superando barreras lingüísticas y facilitando la comunicación global. Herramientas como *Google Translate* utilizan técnicas de NLP para proporcionar traducciones entre numerosos idiomas.
Análisis de Sentimiento	El NLP ayuda a las empresas a comprender el sentir público sobre sus productos o servicios a través del análisis de sentimientos en redes sociales, reseñas y comentarios de clientes. Esta información puede informar estrategias de marketing y mejoras en los productos.
Generación de Contenido	El NLP puede generar texto similar al humano, lo que lo hace útil para diversas aplicaciones como la creación de contenido, chatbots y recomendaciones personalizadas.
Salud	El NLP puede ayudar en el procesamiento y análisis de registros médicos, artículos de investigación y datos de pacientes, lo que lleva a mejores diagnósticos, recomendaciones de tratamiento y resultados de investigación.
Finanzas	El NLP es crucial en la industria financiera para analizar noticias, informes y datos en redes sociales, y tomar decisiones de inversión informadas y detectar fraudes financieros.
Cumplimiento legal	Los profesionales del derecho utilizan el NLP para revisar y analizar grandes volúmenes de documentos legales de manera eficiente, lo que ayuda en la investigación, la debida diligencia y el cumplimiento normativo.
Educación	El NLP puede respaldar experiencias de aprendizaje personalizadas mediante el análisis del rendimiento de los estudiantes, la generación de contenido educativo y la retroalimentación en tiempo real.

[5] AWS. *What Is Natural Language Processing (NLP)?* 2023. Disponible en: https://aws.amazon.com/what-is/nlp/?nc1=h_ls

Tabla 1. Aplicaciones de NLP.	
Accesibilidad	La tecnología de NLP ayuda a que el contenido digital sea accesible para personas con discapacidades al proporcionar reconocimiento de voz, texto a voz y otras características de asistencia.
Atención de Desastres	El NLP puede ayudar en la gestión de desastres al analizar publicaciones en redes sociales y noticias para evaluar el impacto de un desastre, identificar áreas afectadas y coordinar los esfuerzos de respuesta.
Seguridad	El NLP se puede utilizar para detectar amenazas en la comunicación en línea, analizar registros de ciberseguridad y mejorar la seguridad de los sistemas al identificar actividades sospechosas
Investigaciones de Mercado	El NLP puede extraer conocimientos de encuestas, reseñas y discusiones en redes sociales para ayudar a las empresas a comprender las tendencias del mercado, las preferencias de los consumidores y el panorama competitivo.

2.3 ¿Cómo funciona el NLP?

Para comprender en pocas palabras cómo funciona el NLP, desde la adquisición de datos, abarcando el procesamiento de datos y los modelos de aprendizaje automático, hasta su aplicación final, podemos generalizarlo en siete pasos que se describen de la siguiente manera:

1. **Recopilación de Datos:** La recopilación de datos representa el paso fundamental en el proceso de NLP. Implica la adquisición de datos textuales de diversas fuentes, como el *scraping web*[6], las redes sociales, la literatura académica o bases de datos específicas de un dominio. La calidad y diversidad del conjunto de datos son críticas ya que impactan directamente en el rendimiento y la generalización de los modelos de NLP.

   Los sistemas de NLP efectivos a menudo requieren recursos lingüísticos extensos, que incluyen corpus[7] de texto y conjuntos de datos etiquetados. Estas colecciones sirven como datos de entrenamiento para los modelos de *Machine Learning*, permitiéndoles **adquirir patrones y asociaciones lingüísticas**. Los corpus abarcan diversos géneros de texto, idiomas y dominios para aumentar su robustez.

2. **Preprocesamiento de Datos:** Los datos textuales crudos recopilados deben someterse a un preprocesamiento para ser adecuados para tareas de NLP. Los pasos de preprocesamiento incluyen la limpieza de texto, que implica la

[6] El *web scraping* es el proceso de utilizar bots para extraer contenido y datos de un sitio web. A diferencia del *screen scraping*, que solo copia píxeles mostrados en la pantalla, el *web scraping* extrae el código HTML subyacente y, con él, los datos almacenados en una base de datos. El rastreador puede replicar todo el contenido del sitio web en otro lugar.

[7] Un corpus se refiere a una colección grande y estructurada de datos de texto o lenguaje hablado que se utiliza para análisis lingüístico, modelado de lenguaje y otras tareas de NLP. El corpus es la base para desarrollar y entrenar varios modelos y algoritmos de NLP.

eliminación de etiquetas HTML, caracteres especiales e información irrelevante. La tokenización (de la cual se hablará más adelante) divide el texto en palabras individuales o tokens, y el *stemming* o lematización reducen las palabras a sus formas base. Esta fase también aborda problemas como errores ortográficos, formato inconsistente y el manejo de datos faltantes.

3. **Extracción de Características[8]:** La extracción de características convierte el texto en representaciones numéricas que los algoritmos de *Machine Learning* pueden procesar de manera efectiva. Las técnicas comunes incluyen incrustaciones de palabras (por ejemplo, Word2Vec, GloVe), que capturan **relaciones semánticas entre palabras**, y TF-IDF (*Term Frequency-Inverse Document Frequency*), que **asigna pesos a las palabras según su importancia** en los documentos[9]. Métodos avanzados como las incrustaciones BERT capturan información contextual.

4. **Selección y Entrenamiento del Modelo[10]:** Elegir el modelo de *Machine Learning* o de *Deep Learning* apropiado es fundamental para el éxito de las tareas de NLP. Aquí se utilizan comúnmente las Redes Neuronales Recurrentes (RNN), las Redes Neuronales Convolucionales (CNN) y los modelos basados en transformadores (por ejemplo, BERT, GPT). El entrenamiento de estos modelos requiere datos etiquetados y las técnicas de optimización como el descenso de gradiente ajustan los parámetros del modelo para desempeñar tareas específicas.

5. **Evaluación del Modelo:** El rendimiento de los modelos de NLP se evalúa rigurosamente mediante métricas de evaluación específicas de la tarea. Las métricas comunes incluyen precisión, exhaustividad, puntuación F1, puntuación BLEU (para la traducción automática) y perplejidad (para la modelización del lenguaje). La validación cruzada o conjuntos de datos de retención son empleados para asegurar que los modelos se generalicen bien a datos no vistos.

6. **Optimización del Modelo:** Optimizar los modelos de NLP implica mejorar su eficiencia, precisión y robustez. Esto se puede lograr mediante la sintonización de hiperparámetros, el aprendizaje por transferencia y métodos de conjunto. Abordar las consideraciones éticas, como la mitigación del sesgo y la equidad, es una parte esencial en la optimización del modelo.

7. **Implementación e Integración:** Una vez que los modelos de NLP están entrenados y optimizados, son implementados en aplicaciones prácticas. Este paso implica el desarrollo de interfaces de usuario, la integración con bases de datos y el garantizar la escalabilidad y confiabilidad. El monitoreo y mantenimiento continuos son

[8] La idea general es entender el objeto de su funcionamiento más no entrar en la particularidad, para más detalle se sugiere al lector investigar dentro de las referencias de la presente investigación.
[9] Raymond Cheng. *Understanding TF-IDF: A Traditional Approach to Feature Extraction in NLP*. Medium. 2023. Disponible en: https://towardsdatascience.com/understanding-tf-idf-a-traditional-approach-to-feature-extraction-in-nlp-a5bfbe04723f
[10] La idea general es entender el objeto de su funcionamiento más no entrar en la particularidad, para más detalle se sugiere al lector investigar dentro de las referencias de la presente investigación.

necesarios para abordar problemas que puedan surgir en la producción, como el cambio de concepto y la degradación del modelo.

2.4 La importancia del procesamiento de datos en el NLP

Los modelos de NLP son fundamentalmente impulsados por datos, dependiendo de patrones, relaciones y asociaciones estadísticas dentro de los datos de entrada para dar sentido al lenguaje. Cuando se enfrentan a datos ruidosos o no estructurados, estos modelos pueden aprender inadvertidamente correlaciones espurias o no capturar matices lingüísticos esenciales. Estas deficiencias pueden manifestarse como predicciones sesgadas, mala generalización o una comprensión inexacta del lenguaje. Por lo tanto, un meticuloso procesamiento de datos sirve como un baluarte de protección contra estos obstáculos.

En este contexto, el procesamiento de datos eficaz se convierte como un pilar fundamental del éxito en el NLP debido a su influencia general en el rendimiento y la interpretabilidad de los modelos de NLP. Esto se enfatiza mediante el antiguo adagio en la informática, "basura entra, basura sale", que encuentra una aplicación especialmente adecuada en el contexto del NLP. Sirve como un recordatorio contundente de que la calidad de los datos de entrada está inexorablemente vinculada a la calidad de los resultados, y esta máxima se amplifica cuando se trata de las complejidades del lenguaje humano[11].

Uno de los principales beneficios de un procesamiento de datos diligente es la mitigación del ruido. Los datos de texto en bruto a menudo albergan inconsistencias, errores tipográficos y contenido irrelevante que pueden distorsionar el proceso de aprendizaje de los modelos de NLP. Al limpiar y estandarizar meticulosamente los datos, los científicos de datos pueden mejorar estos problemas y proporcionar una base más limpia y confiable para el análisis subsecuente.

Además, el procesamiento de datos genera uniformidad en el conjunto de datos. La capitalización inconsistente, las variaciones en la ortografía o las representaciones diversas de una misma entidad pueden llevar a la fragmentación de los datos, obstaculizando la capacidad del modelo para generalizar de manera efectiva. A través de la normalización y estandarización, el procesamiento de datos fomenta un conjunto de datos cohesivo y armonioso, lo que permite a los modelos de NLP identificar y aprender patrones de manera más efectiva.

Esencialmente, **el procesamiento de datos eficaz allana el camino para la extracción de patrones lingüísticos significativos**. El lenguaje, con sus intrincaciones y sutilezas, oculta una gran cantidad de información bajo la superficie. Al reducir las palabras a sus formas base (lematización o *steamming*), eliminar el ruido (palabras vacías) y elucidar estructuras sintácticas (análisis sintáctico), el procesamiento de datos revela información lingüística

[11] Manning, C. D., & Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press. 1999.

valiosa oculta. El procesamiento de datos empodera a los modelos de NLP para discernir sentimientos, extraer entidades, comprender relaciones sintácticas y realizar una multitud de tareas relacionadas con el lenguaje con mayor precisión y profundidad.

Finalmente, el procesamiento de datos en aplicaciones de NLP se adapta al tipo de información de la que se desea obtener conocimientos. Es decir, el tipo de procesamiento que se aplica a un determinado corpus depende de los datos mismos del corpus, que principalmente es el idioma en el que está escrito, en el caso de un corpus de texto, o el idioma que se habla en el caso de los corpus de audio.

2.5 Breve explicación de las técnicas comunes en el procesamiento de datos de NLP

Las técnicas[12][13] más comúnmente empleadas para el procesamiento de datos de texto se describen a continuación:

a. **Normalización de Texto.** La normalización de texto es una técnica fundamental en NLP que engloba un conjunto de operaciones destinadas a estandarizar datos textuales para garantizar la consistencia y la comparabilidad entre documentos. Estas operaciones incluyen:

   Conversión a minúsculas. Convertir todo el texto a minúsculas es un paso común para asegurar un análisis insensible a las mayúsculas. Evita que el modelo trate "Palabra" y "palabra" como entidades distintas.

   Manejo de Abreviaturas. Las abreviaturas y acrónimos pueden plantear desafíos para los modelos de NLP. La normalización de texto puede implicar la expansión de abreviaturas a sus formas completas para mejorar la comprensión, por ejemplo, "Dr." a "Doctor".

   Resolución de Expresiones Numéricas. Resolver expresiones numéricas en sus equivalentes verbales mejora la interpretabilidad del texto. Por ejemplo, convertir "5" a "cinco" hace que el texto sea más legible para los humanos y ayuda en tareas posteriores como el análisis de sentimientos.

b. **Tokenización.** La tokenización es un paso fundamental en NLP que implica descomponer un texto en unidades más pequeñas, conocidas como tokens. Los tokens pueden ser palabras, unidades subpalabras (tokenización de subpalabras) o caracteres (tokenización a nivel de caracteres). Este proceso sirve como base para el análisis y modelado subsecuentes. La tokenización se puede lograr utilizando métodos basados en reglas o modelos de *Machine Learning* sofisticados entrenados específicamente para este propósito.

[12] No necesariamente todos deben ser utilizados al procesar datos, depende del corpus y del programador identificar los que mejor se adapten para limpiar el corpus.
[13] Steven Bird, Ewan Klein, y Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009: 39-73, 80-116.

c. **Eliminación de Palabras Vacías.** Las palabras vacías son palabras ubicuas en un idioma (por ejemplo, "y," "el," "en") que tienen poco contenido semántico. Eliminar las palabras vacías es una técnica común de preprocesamiento de datos en NLP, la cual reduce el ruido y la dimensionalidad de los datos, mejora la eficiencia computacional y enfoca el modelo en palabras y frases que llevan contenido significativo.

d. ***Stemming* y Lematización.** El *stemming* y la lematización son técnicas empleadas para reducir las palabras a sus formas base o de raíz. Estas técnicas ayudan a capturar las variaciones de palabras y mejorar la generalización de los modelos de NLP:

   Stemming. EL *stemming* implica aplicar reglas heurísticas para truncar palabras a sus formas base. Por ejemplo, "terrorífico" se podría truncar a "terror". Si bien este truncad es computacionalmente menos intensivo, puede producir resultados imperfectos.

   Lematización. La lematización, por otro lado, mapea palabras a sus lemas o formas de diccionario. Por ejemplo, "mejor" se lematiza como "bien". La lematización generalmente produce formas base lingüísticamente válidas, pero puede ser computacionalmente más intensiva que el truncado.

e. **Análisis Sintáctico.** El análisis sintáctico es una técnica sofisticada para analizar la estructura gramatical de las oraciones. Involucra la identificación de relaciones gramaticales entre palabras y frases, lo que **ayuda a comprender la estructura y la semántica de las oraciones**. Las técnicas de análisis sintáctico comunes incluyen el análisis de dependencias y el análisis de constituyentes, que establecen la jerarquía sintáctica dentro de las oraciones.

Para ilustrar cómo se realiza el procesamiento de datos en NLP, usemos la frase "El Sr. Gerardo está trabajando en una investigación sobre NLP". En la Tabla 2 se muestra cómo se podría aplicar cada paso a esta frase:

Tabla 2. Ejemplo de cómo se hace el procesamiento de datos en NLP.

Técnica del Procesamiento de Datos	Ejemplo
Normalización de Texto	Podemos realizar la normalización de texto convirtiendo todos los caracteres en minúsculas y eliminando los acentos: "el sr. gerardo esta trabajando en una investigación sobre nlp". Esto asegura que el texto sea insensible a las mayúsculas para un análisis posterior.
Tokenización	En nuestra frase, la tokenización podría resultar en los siguientes tokens: ["el", "sr.", "gerardo", "esta", "trabajando", "en", "una", "investigación", "sobre", "nlp"]. Estos tokens sirven como los bloques de construcción para un procesamiento adicional
Eliminación de palabras vacías	Las palabras vacías, que son palabras comunes que a menudo tienen un significado mínimo, se pueden eliminar para reducir el ruido y la dimensionalidad. En nuestra frase, las palabras vacías "esta," "en," y

Tabla 2. Ejemplo de cómo se hace el procesamiento de datos en NLP.	
Técnica del Procesamiento de Datos	**Ejemplo**
	"una" se pueden eliminar, entonces la frase quedaría como: ["sr.", "gerardo", "trabajando", "investigación", "nlp"].
Stemming y Lematización	El *stemming* o la lematización se pueden aplicar para reducir las palabras a sus formas base. En nuestra frase, "trabajando" podría truncarse a "trabajar," e "investigación" puede permanecer sin cambios, ya que ya está en su forma base.
Análisis Sintáctico (opcional)	Para nuestra frase simple este paso podría no ser necesario, pero en oraciones más complejas ayuda a identificar las relaciones entre palabras y frases.

Después de estos pasos de procesamiento de datos, el texto procesado resultante sería: "sr. gerardo trabajar investigación nlp". Como se mencionó anteriormente, el procesamiento de datos en NLP transforma datos textuales no estructurados en un formato que los modelos de NLP pueden comprender y trabajar de manera efectiva, mejorando su capacidad para extraer información y derivar significado del texto.

3. La dominancia del idioma Inglés

Según la metodología del Atlas Mundial de Lenguas, existen **8,324 idiomas**, hablados o señalados, documentados por los gobiernos, instituciones públicas y comunidades académicas. De esos 8,324 idiomas, más de 7,000 todavía siguen en uso[14]. Siguiendo este hecho, según la revista Ethnologue, los idiomas con más hablantes nativos para 2022 son el chino mandarín (920 millones), el español (475 millones), el inglés (373 millones) y el hindi (344 millones) (ver Tabla 3).[15] Sin embargo, en la lista de idiomas con más hablantes, incluyendo hablantes nativos y no nativos, **el inglés ocupa el primer lugar con (1,453 millones)**, seguido por el chino mandarín con (1,119 millones), el hindi (602) y el español (549). Pero, ¿qué ha llevado al inglés a convertirse en un idioma dominante?

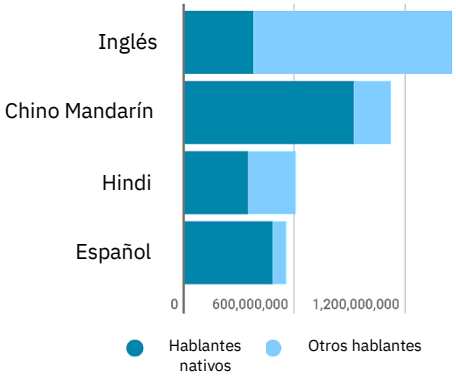

Tabla 3. Lenguajes con más hablantes, 2022.		
País	Hablantes nativos (en millones)	Otros hablantes (in millions)
Chino Mandarín	**920**	199
Español	475	74
Inglés	373	**1080**
Hindi	344	258

[14] UNESCO. *World Atlas Languages*. 2023. Disponible en: https://en.wal.unesco.org/world-atlas-languages https://www.ethnologue.com/browse/names/

[15] Ethnologue, *What are the top 200 most spoken languages?* 2023. Disponible en: https://www.ethnologue.com/insights/ethnologue200/

En la era digital de hoy, el idioma inglés ha surgido como la lengua franca dominante en el mundo analógico y digital. Esta dominancia se extiende a varios aspectos de la comunicación digital, incluyendo la creación de contenido, las redes sociales, los negocios globales y el desarrollo de tecnología. Un factor clave que contribuye a este fenómeno es de carácter histórico.

El inglés, como idioma, ha sido moldeado significativamente por su historia colonial y la expansión del Imperio Británico. Durante el apogeo de este imperio, el inglés se exportó a numerosas regiones de todo el mundo, convirtiéndose en una lengua franca global. Este legado histórico sentó las bases para el uso generalizado del inglés en varios ámbitos, incluyendo la comunicación digital.

Además, durante un período significativo, el idioma de elección en el ámbito de la ciencia ha sido el inglés. Estados Unidos, a partir de finales del siglo XIX, consolidó su estatus como una destacada fuerza económica y política global, con un ascenso notablemente acelerado por los acontecimientos de las dos Guerras Mundiales. Como resultado, el inglés asumió un papel dominante en el discurso internacional y, correspondientemente, en el ámbito de la ciencia. Este desarrollo subyacente desempeñó un papel fundamental en afianzar al inglés como el idioma predominante para la comunicación en ciencia y tecnología, durante su surgimiento en las décadas de 1940 y 1950[16].

En este contexto, otro factor clave es el papel del inglés en los avances tecnológicos e innovación. Una parte significativa de la literatura científica, como lo son los documentos de investigación y la documentación técnica, se publica en inglés. Esto ha convertido al inglés en una herramienta esencial para las personas que trabajan en campos como la informática, la ingeniería y la ciencia de datos, que son fundamentales para el mundo digital.

Además, el inglés tiene una importancia particular en el ámbito de la programación. A pesar de los esfuerzos por desarrollar lenguajes de programación con palabras clave en otros idiomas que no sean el inglés, tales intentos no lograron una adopción generalizada. De hecho, más de un tercio de los lenguajes de programación se desarrollaron en países de habla inglesa[17]. Por si fuera poco, los lenguajes de programación convencionales, de acuerdo con las pautas de estilo, requieren que los programadores utilicen el inglés para escribir comentarios, nombrar variables, funciones y clases[18].

[16] Rainer Enrique Hamel. *The dominance of English in the international scientific periodical literature and the future of language use in science*. AILA Review 20, 2007: 53-71, 56.
[17] Y Studios. *The Language of Codes : Why English is the Lingua Franca of Programming*, 2018. Disponible en: https://ystudios.com/insights-passion/codelanguage
[18] Philip J. Guo, *Non-Native English Speakers Learning Computer Programming: Barriers, Desires*, and Design Opportunities, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018: 2.

Siguiendo esto, la influencia económica de los países de habla inglesa ha desempeñado un papel fundamental en la dominancia del idioma en Internet. Los Estados Unidos, en particular, ha estado a la vanguardia de la innovación digital y el desarrollo tecnológico. Muchas de las empresas tecnológicas más influyentes del mundo como Google, Facebook y Apple, tienen su sede en regiones de habla inglesa. En consecuencia, estas empresas a menudo desarrollan sus productos y servicios en inglés, lo que a su vez moldea el comportamiento y las preferencias de los usuarios en todo el mundo.

Además, el alcance global del inglés se ha visto reforzado por la internacionalización de los negocios. El inglés se utiliza a menudo como el idioma predeterminado para el comercio internacional, las finanzas y los negocios [19]. Como resultado, los profesionales de diversos países necesitan tener un conocimiento práctico del inglés para participar de manera efectiva en la economía global. Esta interdependencia económica fomenta la prominencia continua del inglés en las transacciones comerciales y la comunicación digital.

Es importante destacar que el mundo digital opera en base a **efectos de red,** donde el valor de una red o plataforma aumenta a medida que se unen más usuarios. Este principio también se aplica al idioma en la comunicación digital. **A medida que se produce más contenido en inglés, más usuarios tienden a aprender y usar el inglés para acceder a información, conectarse con otros y participar en comunidades en línea globales** [20]. Este ciclo de retroalimentación adicional solidifica la dominancia del inglés.

Y si todos los hechos anteriores no fueran suficientes, el inglés se utiliza como segundo idioma en países donde la lengua materna no es el inglés. Esto no solo aumenta la universalidad del idioma, sino que también aumenta la demanda del mismo.

3.1 El uso de modelos de NLP frente a otros idiomas

Como se describió anteriormente, el procesamiento de datos en NLP está intrincadamente diseñado para los objetivos específicos y de la naturaleza de la información buscada. La diversidad y adaptabilidad de las metodologías de procesamiento de datos en NLP son emblemáticas de la naturaleza multifacética de esta tecnología. Esta adaptabilidad se basa en el reconocimiento fundamental de que las estrategias de procesamiento empleadas dependen de un factor primordial: el idioma en el que se expresa el texto o los datos sonoros. En consecuencia, **el tipo y el alcance del procesamiento aplicado a un corpus dado están inextricablemente vinculados a los atributos lingüísticos que lo rigen.**

Para ilustrar esta noción con mayor detalle, es necesario destacar que el idioma que subyace en el texto o los datos vocales constituye el punto central en torno al cual gira el procesamiento de datos en NLP. En otras palabras, el idioma en el que se transmite la

[19] Mark Warschauer. *The Changing Global Economy and the Future of English Teaching*. TESOL Quarterly, 2000: 511-535.
[20] Carl Shapiro y Hal R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Review Press, 1999.

información da forma a las técnicas y métodos empleados en su análisis. Esta orientación centrada en el idioma abarca dos aspectos fundamentales: la semántica y la sintaxis.

Para obtener un modelo de NLP que realmente comprenda y genere el lenguaje humano de manera efectiva, es necesario estudiar profundamente cómo funcionan la semántica y la sintaxis. El análisis semántico permite a las máquinas comprender el significado detrás del texto, facilitando tareas como el análisis de sentimientos, la extracción de información y la respuesta a preguntas. La sintaxis, por otro lado, permite a los sistemas de NLP analizar y generar oraciones gramaticalmente correctas, garantizando la coherencia y corrección de la respuesta generada. Tanto la semántica como la sintaxis están interconectadas y constituyen la base para la comprensión y generación de lenguaje de nivel superior en las aplicaciones de NLP. Estos elementos son tan importantes en las aplicaciones de NLP que cada vez más empresas contratan profesionales especializados en lingüística. En la Tabla 4 se presenta la definición de Semántica y Sintaxis para una comprensión más profunda.

Tabla 4. Definiciones de la Semántica y de la Sintaxis.

Semántica	Sintaxis
Se refiere al significado transmitido por palabras, frases, oraciones y discurso. Comprende la **interpretación de elementos lingüísticos con respecto a sus referentes en el mundo real.** La semántica nos permite comprender las **relaciones entre las palabras y cómo transmiten información.** Aborda preguntas como "¿Qué significa esta palabra?" y "¿Cómo se combinan las palabras para crear expresiones significativas?" La semántica desempeña un papel crucial en tareas de NLP como el análisis de sentimientos, la recuperación de información y la traducción automática, donde comprender el significado del texto es fundamental.	Se refiere a la **estructura y disposición de palabras y símbolos dentro de un idioma**. Define las reglas y principios que rigen la formación de oraciones, incluido el orden de las palabras, las categorías gramaticales y las relaciones sintácticas entre los elementos. La sintaxis proporciona el marco para **organizar las palabras en oraciones coherentes y gramaticalmente correctas**. Se ocupa de preguntas como "¿Cómo deben ordenarse las palabras para formar una oraciones válidas?" y "¿Cuáles son los roles gramaticales de las palabras individuales dentro de una oración?" La sintaxis es esencial para tareas como la corrección gramatical, el etiquetado de partes del discurso y el análisis sintáctico en NLP, ya que ayuda a analizar y generar texto correctamente estructurado.

Sin embargo, **la semántica y la sintaxis**, como componentes integrales del análisis lingüístico, **no son constantes universales que se puedan aplicar homogéneamente a todos los idiomas**. Más bien, están intrínsecamente incrustadas en las idiosincrasias y matices de cada idioma específico. Por ejemplo, la semántica de un término o frase puede divergir significativamente entre idiomas, dando lugar a connotaciones distintas, asociaciones culturales y sutilezas lingüísticas. De manera similar, la sintaxis que rige la disposición de palabras y constituyentes en una oración varía notablemente de un idioma a otro, lo que da lugar a reglas gramaticales distintas y configuraciones estructurales diferentes.

Para ilustrar este punto, consideremos la errónea utilización de un modelo de NLP que ha sido exclusivamente entrenado con datos en inglés cuando se enfrenta a un corpus compuesto por texto en español. La falta de armonía semántica y sintáctica entre el inglés y el español, arraigada en sus disparidades lingüísticas, hace que una aplicación sea

inviable. Intentar usar un modelo de NLP basado en inglés en un corpus en español conlleva el riesgo de producir resultados inexactos o sin sentido, debido a la incongruencia entre el conocimiento lingüístico del modelo y las características inherentes del idioma español.

En este contexto, la adaptación de modelos de NLP a idiomas que no son el inglés plantea tres desafíos notables[21], cada uno de los cuales presenta obstáculos únicos para lograr un rendimiento efectivo en NLP:

- **Falta de Recursos.** Uno de los principales desafíos en la adaptación de modelos de NLP a idiomas diferentes al inglés es la **escasez de recursos lingüísticos esenciales**. Los conjuntos de datos extensos y de alta calidad son un requisito fundamental para entrenar modelos robustos de NLP. En muchos casos, estos conjuntos de datos son abundantes en inglés, pero pueden ser difíciles de encontrar o crear para idiomas diferentes a éste. Sin una cantidad suficiente y diversa de datos, el desarrollo y ajuste fino de modelos de NLP se vuelve significativamente más desafiante.

  Además, la disponibilidad de modelos preentrenados y los recursos específicos de idioma a menudo está limitada para idiomas que no son el inglés. Estos modelos preentrenados pueden servir como un punto de partida crucial para tareas de NLP, pero su escasez puede obstaculizar el proceso de desarrollo.

- **Complejidad del idioma.** La complejidad lingüística inherente de los idiomas que no son el inglés puede plantear un desafío significativo. Idiomas como el chino y el árabe, conocidos por sus intrincados caracteres y escrituras, requieren técnicas de procesamiento especializadas. A diferencia del inglés, que utiliza un sistema de escritura basado en el alfabeto, estos idiomas involucran caracteres complejos que a menudo requieren enfoques diferentes de tokenización y de procesamiento de texto. El manejo efectivo de estas complejidades demanda modificaciones en los algoritmos y herramientas de NLP.

- **Diferencias Culturales.** Los modelos de NLP están profundamente influenciados por las sutilezas culturales y las idiosincrasias específicas del idioma de sus usuarios objetivos. En consecuencia, adaptar modelos desarrollados originalmente para un idioma, con su respectiva cultura, a otro idioma, con otro tipo de cultura, para que funcionen de manera efectiva puede ser difícil. Las variaciones culturales en expresiones, modismos, sentimientos y convenciones lingüísticas requieren una personalización extensa y de un ajuste fino del modelo para asegurarse de que se alinee con el contexto cultural específico y las expectativas de la audiencia objetivo.

[21] Anastassia Kornilova and April Guo. *Adapting language-based models beyond English.* Snorkel, 2023. Disponible en: https://snorkel.ai/adapting-language-based-models-beyond-english/

3.2 Spicy y NLTK: las mejores herramientas de Python para NLP... ¿Sólo para el idioma Inglés?

Actualmente, Python ha sido el lenguaje de programación más común para desarrollar modelos de NLP[22] debido a su simplicidad, extensas librerías[23] y un ecosistema vibrante para la ciencia de datos y *Machine Learning*. Muchas librerías y marcos de trabajo de NLP están construidos en Python, lo que lo convierte en el lenguaje preferido para el desarrollo de NLP.

Python ofrece varias librerías para desarrollar modelos de NLP, siendo las más comúnmente utilizadas *Natural Language Toolkit* (NLTK) y spaCy. Con éstas, los programadores tienen la capacidad de crear chatbots, herramientas de resumen automatizado y sistemas de reconocimiento de entidades. Aunque ambas librerías teóricamente pueden manejar cualquier tarea de NLP, cada una destaca en situaciones específicas [24].

NLTK ofrece una amplia gama de librerías de procesamiento de texto, recursos y modelos preentrenados. Su popularidad se debe, en parte, a que es fácil de usar y se utiliza a menudo con fines educativos [25]. Por otro lado, spaCy es conocido por sus modelos preentrenados listos para producción, y es preferido por los desarrolladores para construir aplicaciones de NLP escalables, ganando popularidad por su velocidad y eficiencia [26].

Sin embargo, como se mencionó anteriormente, la efectividad de las aplicaciones de NLP (incluidas las herramientas y librerías) puede variar significativamente cuando se aplican a diferentes idiomas. Entonces, ¿podemos generalizar que estas librerías son adecuadas para idiomas diferentes al inglés?

En 2018 se realizó un análisis llamado "Características básicas de spaCy: comparación de rendimiento para portugués, francés e inglés"[27], en el que se demostró que spaCy era

---

[22] Turing. *Which Language Is Useful for NLP and Why?* 2023. Disponible en: https://www.turing.com/kb/which-language-is-useful-for-nlp-and-why

[23] Las librerías están compuestas por módulos incorporados (escritos en C) que brindan acceso a la funcionalidad del sistema, como la entrada/salida de archivos, así como módulos escritos en Python que proporcionan soluciones estandarizadas para muchos de los problemas que surgen en la programación del día a día. Algunos de estos módulos están diseñados específicamente para fomentar y mejorar la portabilidad de los programas de Python, ya que abstraen los detalles específicos de la plataforma en API neutrales a la Plataforma. https://docs.python.org/3/library/index.html

[24] Swaathi Kakarla, *Natural Language Processing: NLTK vs spaCy*. Active State, 2019. Disponible en: https://www.activestate.com/blog/natural-language-processing-nltk-vs-spacy/

[25] NLTK. *Natural Language Toolkit*. 2023. Disponible en: https://www.nltk.org/

[26] Dominik Kozaczko. *9 best Python Natural Language Processing (NLP) libraries.* Sunscrapers, 2023. Disponible en: https://sunscrapers.com/blog/9-best-python-natural-language-processing-nlp/

[27] Wilame. *spaCy basic features: comparing performance for Portuguese, French and English*. Medium, 2018. Disponible en: https://medium.com/@wila.me/spacy-basic-features-comparing-performance-for-portuguese-french-and-english-bb2edab49b4

competente para realizar tareas fundamentales en textos bien estructurados en inglés, incluyendo comunicados de prensa y obras literarias. Sin embargo, se reconoció que se requieren esfuerzos adicionales al abordar idiomas distintos al inglés.

Con resultados similares, en 2019 se llevó a cabo un análisis denominado "Preprocesamiento de texto en diferentes idiomas para el procesamiento del lenguaje natural en Python", en el que se evaluaron los efectos de la eliminación de palabras vacías, la eliminación de palabras extremadamente frecuentes y poco frecuentes, y el *stemming* para los idiomas inglés, alemán, húngaro y rumano, utilizando NLTK[28]. En general, los resultados del análisis demostraron que los idiomas pueden diferir sustancialmente a las técnicas de preprocesamiento de texto. En el caso de la eliminación de palabras vacías, ésta eliminó más palabras en idiomas donde no se utilizaban sufijos extensamente, mientras que el *stemming* afectó más a los idiomas ricos en sufijos. En general, el inglés fue más adecuado para esas técnicas de procesamiento, lo que puede llevar a resultados no deseados en aplicaciones de NLP para idiomas diferentes al inglés.

Los resultados de estos análisis comparativos indican que spaCy y NLTK funcionan mejor en inglés en diversas tareas de NLP, en comparación con idiomas diversos al inglés. Las causas de esta discrepancia es atribuida a diversos factores:

- **Modelos específicos en idioma inglés.** SpaCy y NLTK cuentan con modelos y recursos de idioma más sólidos disponibles para el inglés. Esta abundancia de recursos contribuye a un rendimiento superior en tareas en inglés, como el reconocimiento de entidades nombradas, donde SpaCy sobresale. De hecho, spaCy solo ofrece modelos para 24 idiomas[29], de los cuales solo 10 idiomas[30] se benefician de un conjunto de 4 paquetes cada uno, mientras que los idiomas restantes cuentan con un conjunto de 3 paquetes cada uno.

- **Disponibilidad de datos.** La disponibilidad de grandes conjuntos de datos de entrenamiento de alta calidad en inglés facilita un mejor rendimiento para los modelos de procesamiento del lenguaje natural. Otros idiomas a menudo carecen de corpus limitados o inexistentes, lo que conduce a tasas de precisión y recuperación más bajas.

- **Complejidad del idioma.** Los idiomas no ingleses a menudo exhiben una mayor complejidad lingüística que incluye diversas estructuras gramaticales y variaciones morfológicas. Esta complejidad plantea desafíos para los modelos de procesamiento del lenguaje natural entrenados principalmente con datos en inglés.

[28] Mor Kapronczay. *Text preprocessing in different languages for Natural Language Processing in Python. Part II – Case of Study.* Medium, 2019. Disponible en: https://medium.com/starschema-blog/text-preprocessing-in-different-languages-for-natural-language-processing-in-python-fb106f70b554

[29] Consultado el 28 de septiembre de 2023 en: https://spacy.io/usage/models

[30] El catalán, el chino, el danés, el inglés, el francés, el alemán, el japonés, el esloveno, el español y el ucraniano.

Este análisis resalta los desafíos que enfrentan las librerías de NLP en aplicaciones multilingües y enfatiza la importancia de los esfuerzos dedicados para mejorar las capacidades de NLP en idiomas diferentes al inglés. Reducir esta brecha de rendimiento es fundamental para el avance del procesamiento del lenguaje natural en diversos contextos lingüísticos.

4. La disparidad lingüística en las aplicaciones de NLP: la Discriminación Indirecta

El auge de las aplicaciones de NLP ha revolucionado sin duda alguna la forma en que interactuamos con la tecnología. Sin embargo, bajo la superficie se encuentra un problema sutil pero profundo: la discriminación indirecta perpetuada por estas aplicaciones. Esta discriminación no es intencional, sino más bien una consecuencia de la dominancia del inglés en el desarrollo tecnológico.

A diferencia de otras formas de sesgos observados en NLP, como el sesgo de género o el sesgo racial, que se manifiestan como resultados discernibles dentro de las aplicaciones de NLP, el sesgo lingüístico se presenta como un impedimento inicial e indirecto para las personas que buscan utilizar aplicaciones de NLP. Este tipo particular de sesgo se caracteriza como indirecto debido a su origen en el efecto de red y la prevalencia histórica de la dominancia del inglés. Es importante señalar que el sesgo lingüístico no es el resultado de acciones deliberadas o negligentes por parte de programadores o de cómo se recopilaron los datos, si no que **surge de manera orgánica a partir del desarrollo histórico y la influencia del idioma inglés.**

Como se mencionó anteriormente, el efecto de red es una fuerza tremenda que sostiene la dominancia del idioma inglés en el mundo digital. A medida que se crea y consume más contenido en inglés, a medida que las empresas lo adoptan dentro de sus operaciones, a medida que la tecnología y la innovación siguen centradas predominantemente en el inglés y a medida que las plataformas de redes sociales promueven la interacción en inglés, el efecto de red se fortalece. Como consecuencia, otros idiomas enfrentan barreras gigantescas al intentar desbancar al inglés de su posición de dominio digital.

En este contexto, la dominancia del inglés en el desarrollo tecnológico es innegable. La gran mayoría de conjuntos de datos, modelos preentrenados y recursos de NLP están principalmente disponibles en inglés. Así, los desarrolladores a menudo priorizan el inglés al diseñar y entrenar aplicaciones de NLP. Si bien esto puede no estar impulsado por una intención discriminatoria, **tiene consecuencias discriminatorias**. Esto fomenta indirectamente un sistema técnico y social donde el inglés se convierte en el estándar de facto para obtener resultados de calidad en aplicaciones de NLP.

Como resultado de esta dominancia, los hablantes de inglés, ya sean nativos o no, disfrutan de un estatus privilegiado en el ámbito digital. Cuando éstos utilizan aplicaciones de NLP obtienen mejores resultados simplemente porque estas aplicaciones están mejor

optimizadas para el inglés. **Este favoritismo no intencional los sitúa en una posición de primera clase en comparación con los hablantes no nativos de inglés.** En esencia, si deseas obtener resultados de primera clase de las aplicaciones de NLP, eres empujado a usar el inglés en tus consultas. Este sesgo inherente refuerza la idea de que el inglés es el idioma del mundo digital.

Las consecuencias de este sesgo lingüístico son particularmente severas para idiomas como el español o el italiano, y aún más para idiomas indígenas como el náhuatl o el guaraní. Los usuarios de estos idiomas enfrentan un doble desafío: en primer lugar, deben superar la brecha digital existente, que incluye problemas de acceso a Internet y la división digital. En segundo lugar, deben superar una brecha lingüística, ya que las aplicaciones de NLP funcionan de manera deficiente o son inexistentes en estos idiomas en comparación con el inglés.

Ahora bien, pongamos nuestra atención a la población mundial. Según datos del 22 de noviembre 22 de 2022, la población mundial aumentó a una asombrosa cifra de 8000 millones de personas[31]. Sin embargo, en el panorama lingüístico, una cifra se destaca: existen aproximadamente 1453 millones de hablantes de inglés, tanto nativos como no nativos. Por lo anterior, existe un intrigante fenómeno: más del 80% de los habitantes del mundo se encuentran fuera de la jugada de NLP donde el idioma inglés domina.

La discriminación indirecta perpetuada por las aplicaciones de NLP es una preocupación alarmante que debe ser abordada. Si bien la dominancia del inglés en el desarrollo tecnológico puede no estar motivada por intenciones maliciosas, da como resultado un mundo digital en el que los hablantes de inglés ocupan una posición privilegiada. Los hablantes no nativos de inglés, especialmente se ven en desventaja aquellos que hablan idiomas menos hablados o lenguas indígenas. Superar esta brecha lingüística en las aplicaciones de NLP no es solo una cuestión de diversidad lingüística, sino también de inclusión digital y de equidad. Entonces, ¿qué se puede hacer para crear un entorno digital más equitativo e inclusivo en el ámbito de NLP?

5. Barreras que bloquean los desarrollos de NLP para idiomas diferentes al Inglés

En esta investigación se demostraron varias barreras que obstaculizan el desarrollo de NLP en idiomas que no son el inglés:

i. La dominancia del inglés.

[31] ONU Noticias. *La población mundial alcanzó hoy 15 de noviembre de 2022 las 8000 millones de personas de acuerdo al informe Perspectivas de la Población Mundial, que también prevé que India superará a China como el país más poblado del mundo en 2023. El bebé 8 mil millones nació en República Dominicana*. ONU Habitat, 2023. Disponible en: https://onuhabitat.org.mx/index.php/ya-somos-8-mil-millones-de-personas#:~:text=La%20poblaci%C3%B3n%20mundial%20alcanz%C3%B3%20hoy,millones%20naci%C3%B3%20en%20Rep%C3%BAblica%20Dominicana.

ii. La falta de corpus en idiomas diferentes al inglés.
iii. La falta de códigos multilingües y herramientas para aplicaciones de NLP.
iv. La falta de interés genuino en producir aplicaciones de NLP para idiomas diversos al inglés.

5.1 La dominancia del inglés

The idea of multi-language inclusivity in the digital context is not radically to topple the English language in the digital world nor make it in the NLP realm. As was described, the English dominance is something that has prevailed since many decades ago and in a world immerse in the globalization it becomes non-sense task. Furthermore, the idea of eradicating English in the NLP realm would imply eradicating it in other sectors such as economy or digitalization, so it is clear that the answer is not “deleting” English locally in the NLP realm. Besides, eradicating locally English in NLP applications or locally stop using it would be a synonym of blocking digital rights for the people that speak English.

La idea de la inclusividad multilingüe en el contexto digital no tiene como objetivo derrocar radicalmente el idioma inglés en el mundo digital ni en el ámbito del NLP. Como se mencionó anteriormente, el dominio del inglés es algo que ha prevalecido desde hace muchas décadas y, en un mundo inmerso en la globalización, sería una tarea sin sentido. Además, la idea de erradicar el inglés en el ámbito del NLP implicaría erradicarlo en otros sectores, como la economía o la digitalización, por lo que está claro que la respuesta no es "borrar" el inglés en el ámbito del NLP. Además, erradicar el inglés en las aplicaciones de NLP, o dejar de usarlo localmente, sería sinónimo de bloquear los derechos digitales de las personas que hablan inglés.

5.2 La falta de corpus en idiomas diferentes al inglés

El recurso principal para que los científicos de datos desarrollen cualquier aplicación de *Machine Learning*, incluido el NLP, es contar con grandes cantidades de datos (corpus). Cuanto más extenso sea el corpus, más fiable y efectiva será la aplicación. En palabras coloquiales, los datos necesarios para crear una aplicación de NLP sólida deben ser bastante extensos para que los modelos de procesamiento del lenguaje natural sean resistentes a los cambios, sensibles a las entradas y que tengan una baja probabilidad de cometer errores. Sin embargo, la sola idea de crear un corpus enfrenta algunos desafíos:

- **Costos en la creación del corpus.** En términos generales, con las herramientas digitales actuales, hacer corpus no es una tarea difícil; simplemente implica "recopilar datos" de una clase, como "cuarenta mil tuits", "treinta y cinco mil grabaciones de voz" o "veinticinco mil libros". Sin embargo, la creación de corpus representa una limitación económica porque solo las empresas tecnológicas medianas y grandes o las instituciones académicas sólidas pueden costear los gastos asociados con esta tarea, como contratar programadores humanos para recopilar datos, o los costos relacionados con el propio corpus, como el almacenamiento de datos, su actualización y su mantenimiento.

- **Inexistencia de datos para crear el corpus.** Por otro lado, el desarrollo de un corpus puede ser un objetivo alcanzable en idiomas comunes como el español, francés o portugués, dado que la mayoría de estos datos ya están disponibles y previamente se han digitalizado y procesado y, en algunos casos, solo es cuestión de recopilar los datos mismos. Sin embargo, esta labor es difícil de lograr en lenguas indígenas. Esto se debe a algunos factores como: algunas comunidades indígenas tienen una presencia limitada en el mundo digital o, en la mayoría de los casos, no tienen presencia; la mayoría de los datos (información) que las comunidades indígenas pueden tener no han sido digitalizadas (en el caso de sonidos o libros) o pertenecen únicamente a la tradición oral (aún sin ningún registro digital); y algunas comunidades han adoptado una posición reacia con respecto a la apropiación de tecnologías digitales.

- **Barreras regulatorias para la recopilación de datos.** Las regulaciones en todo el mundo tienen diferentes matices, pueden ser *ex post* o *ex ante*, lo que indica el momento en que se aplica la regulación, y cada una puede ser suave o estricta, lo que indica el nivel de severidad que tiene para aquellos que no han seguido las reglas. Al crear corpus, dos regulaciones principales desempeñan un papel clave: la regulación de privacidad de datos y la regulación de propiedad intelectual.

  Por un lado, si bien es cierto que las regulaciones de privacidad de datos protegen a los usuarios finales de que terceros utilicen indebidamente sus datos personales, también se convierten en una barrera infranqueable cuando se necesita recopilar una gran cantidad de datos. Además, compartir corpus es un factor importante para maximizar y promover modelos de NLP multilingües, como sucede actualmente con los modelos de NLP basados en inglés. Una vez más, es importante tener en cuenta que las aplicaciones de procesamiento del lenguaje natural se basan principalmente en inglés no solo porque existen corpus en inglés, sino también porque han alcanzado un alto nivel de difusión gracias a que pueden compartirse fácilmente [32].

  Es importante destacar que los corpus en inglés principalmente se han creado en Estados Unidos, el cual carece de una ley integral que aborde la privacidad de todos los tipos de datos. En cambio, su regulación se basa en una colección de leyes aplicadas por diferentes instituciones. Debido a la ausencia de una regulación federal de privacidad, en Estados Unidos la iniciativa privada tiene una libertad considerable para manejar datos, a menos que un estado en específico haya implementado su propia regulación de privacidad de datos [33]. Esta regulación suave

[32] Por ejemplo, la Universidad de Pensilvania, a través del Consorcio de Datos Lingüísticos (LDC), se ha convertido en una referencia destacada para mantener, actualizar y agregar corpus disponibles de forma gratuita para todos. El LDC ha crecido y se ha convertido en una organización que crea y distribuye una amplia variedad de recursos lingüísticos, y también respalda programas de investigación patrocinados y evaluaciones de tecnología basada en el lenguaje al proporcionar recursos y aportar experiencia organizativa. Disponible en: https://www.ldc.upenn.edu/about

[33] Thorin Klosowski , *The State of Consumer Data Privacy Laws in the US (And Why It Matters)*, The New York Times, 2023, Aailable at: https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/

que tiene Estados Unidos ha facilitado la recopilación de datos, su adopción generalizada, y el intercambio de corpus. Por otro lado, las regulaciones de privacidad en América Latina, cuyos elementos principales se basan en el GDPR de la Unión Europea, se desarrollaron con un enfoque pro persona, lo que no facilita en primera instancia la creación y el intercambio de corpus a terceros.

Por otro lado, las regulaciones de propiedad intelectual también desempeñan un papel importante en la recopilación y el intercambio de datos. Deben aclararse cuestiones relacionadas respecto quién sería el titular del corpus, quién tendría derecho a exigir un interés legítimo para monetizar el corpus y si la acción de recopilación de datos no infringe los derechos de propiedad intelectual de otros.

Un ejemplo famoso que ilustra este punto fue la acción emprendida por *The New York Times* (NYT). Este periódico ha implementado medidas proactivas para evitar la utilización de su contenido para el entrenamiento de modelos de inteligencia artificial (por ejemplo, ChatGPT). Según el informe de Adweek, el 3 de agosto de 2023, el NYT modificó sus Términos de Servicio para prohibir expresamente el uso de su contenido, que incluye texto, imágenes, videos, audio, apariencia visual, metadatos o colecciones, en la creación de cualquier programa de software, incluido, entre otros, el entrenamiento de sistemas de aprendizaje automático o inteligencia artificial [34].

5.3 La falta de códigos multilingües y herramientas para aplicaciones de NLP

Bastante relacionado con el punto anterior, la ausencia herramientas para el procesamiento de datos en lenguajes de programación de NLP (como Python) es una mera consecuencia de la falta de corpus. Como se describió anteriormente, las librerías ayudan a los programadores a facilitar el procesamiento de datos para los corpus, y se personalizan según el idioma del corpus. Cuantos más corpus en inglés existan, más interés habrá por parte de los programadores en desarrollar librerías especializadas para procesar datos y, consecuentemente, estas librerías estarán más afinadas y especializadas.

5.4 La falta de interés genuino en producir aplicaciones de NLP para idiomas diversos al inglés

El interés de las personas en producir aplicaciones de NLP en idiomas diferentes al inglés está bastante relacionado con la supremacía del idioma inglés en diferentes esferas sociales. Si el mercado demanda que los programadores desarrollen aplicaciones de NLP basadas en el inglés, entonces la existencia de corpus en inglés estará justificada y, como consecuencia, el interés de los programadores en desarrollar aplicaciones de NLP en inglés. Los programadores son más propensos a invertir su tiempo y esfuerzo en la creación

[34] Jess Weatherbed, *The New York Times prohibits using its content to train AI models*, The Verge, 2023, Disponible en: https://www.theverge.com/2023/8/14/23831109/the-new-york-times-ai-web-scraping-rules-terms-of-service

de herramientas que se adapten a idiomas con corpus robustos debido a que será mayor la base de usuarios, así como las herramientas para la manipulación del corpus.

La propensión de los desarrolladores a invertir sus recursos y experiencia en la creación de aplicaciones de NLP está fundamentalmente entrelazada con la existencia de un mercado existente. El principal incentivo que guía tales esfuerzos es la eventual monetización de la aplicación de NLP, impulsada por el principio de que cuanto mayor sea el número de hablantes del idioma en la base de usuarios prevista, mayor será la demanda del mercado.

La motivación fundamental para que los desarrolladores de terceros emprendan proyectos de aplicaciones de NLP es la promesa de los retornos de inversión. Estos desarrolladores suelen invertir un tiempo, esfuerzo y recursos significativos en el desarrollo de aplicaciones con la anticipación de una remuneración económica. Por lo tanto, la existencia de un mercado receptivo y sustancial para la aplicación de NLP es un factor determinante en la decisión de iniciar el desarrollo de la aplicación misma. Esto es, el mercado proporciona el impulso económico que justifica el proceso de desarrollo.

Siguiendo esto, un determinante crítico del potencial de mercado para las aplicaciones de NLP es el número de personas que hablan el idioma en el que se basa la aplicación. Una población que habla el idioma con más hablantes, significaría una base de usuarios potencialmente más grande para la aplicación de NLP. Si la utilidad y la relevancia de la aplicación para una audiencia aumenta, aumentará también su demanda en el mercado. Por lo tanto, los idiomas con un mayor número de hablantes resultan naturalmente atractivos para los desarrolladores debido a su mayor capacidad de mercado.

Este principio se aplica no solo al idioma inglés, sino que se extiende a otros idiomas oficiales[35], como el español, el francés o el italiano, reflejando un principio económico universal en el contexto del desarrollo de aplicaciones de NLP.

6. Potenciales soluciones para mitigar la discriminación lingüística en los desarrollos de NLP

Es necesario realizar esfuerzos para desarrollar y optimizar aplicaciones de NLP para una gama más amplia de idiomas, asegurando que todos, independientemente del idioma que hablen, puedan acceder a resultados de alta calidad y participar plenamente en el mundo digital. Mucho más allá de simplemente usar o no el inglés en el ámbito de NLP, existe una necesidad urgente de democratizar el acceso a la tecnología, haciéndola disponible para todos, principalmente (pero no exclusivamente) en el campo de la Inteligencia Artificial. Como se mencionó anteriormente, actualmente vivimos en una era de cambios digitales abruptos y la demanda de tecnologías de vanguardia (como NLP) se difunde fácilmente en

[35] Aquellos reconocidos en la constitución de cada país.

una sociedad digital. Sin embargo, cada vez que aparece un nuevo ícono tecnológico, trae consigo una brecha de apropiación que solo unos pocos pueden superar.

Para ilustrar esto, tomemos el famoso caso de los teléfonos móviles (ahora *smartphones*). Su llegada generó una brecha importante en su apropiación que, combinada con brechas generacionales y económicas, actualmente todavía existe. Los celulares evolucionaron a *smartphones* sin esperar a que las personas cerraran realmente la brecha de apropiación de uno a otro, y como es de observarse esta tecnología sigue evolucionando, dejando atrás a quienes no pueden costearla ni digitalmente apropiársela. La Inteligencia Artificial, donde se encuentra NLP, está teniendo un fenómeno similar: esta tecnología está evolucionando rápidamente y cuanto más no podamos asegurar una inclusión real para todos, más grande será la brecha y más difícil será superarla. Con estos sesgos lingüísticos en NLP se crea una nueva brecha para que las personas se apropien plenamente de esta tecnología. En la imagen 1 se ilustran las brechas por las que una persona debe pasar para apropiarse completamente de la tecnología de NLP.

En este contexto, está claro que se necesita un enfoque de inclusión multilingüe para mitigar la brecha lingüística que ha prevalecido durante décadas, no solo en el contexto social, sino también en el digital, incluyendo el ámbito de NLP. Es una *mitigación* porque no se puede deshacer lo que durante muchos años ha traído el dominio del inglés, y sería

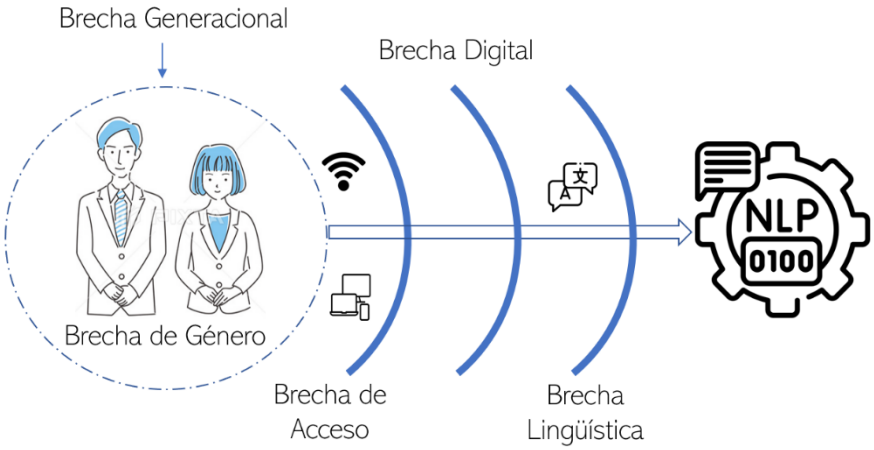

Imagen 1. Diferentes brechas que las personas tienen que enfrentar para usar las aplicaciones de NLP.

una tarea sin sentido cortar “de tajo” el uso del inglés: esta dominancia abarca un amplio espectro en áreas como la educación, la economía, la política, los negocios y la sociedad digital, por lo que pensar en erradicarlo o dejar de usarlo implicaría también erradicarlo en otras áreas diferentes al NLP.

Partiendo del punto de que el idioma inglés está ahí y su presencia continuará para siempre, lo que realmente se debe hacer es promover localmente el uso de idiomas diferentes al inglés en el desarrollo de NLP para que el 80% de la población mundial pueda tener acceso a aplicaciones de NLP de alta calidad, al igual que lo hacen los hablantes de

inglés. Sin embargo, ¿cuál sería el enfoque que realmente puede mitigar el dominio del inglés al brindar inclusividad en el ámbito de NLP?

En primer lugar, es bien sabido que si algo en el mundo digital quiere ser cambiado, se requiere **un enfoque de múltiples partes interesadas**: es imposible abordar problemas digitales desde un solo lado, la problemática debe ser abordada en forma periférica. Por lo tanto, los gobiernos locales y nacionales, las instituciones públicas, el sector privado y las universidades, cada uno en su área de competencia, deben promover el uso de su idioma/idiomas locales para desarrollar herramientas, códigos y corpus sus propios idiomas. Los enfoques de las partes interesadas para fomentar aplicaciones de NLP multilingües se discuten a continuación.

6.1 Gobierno

6.1.1 La diversidad lingüística en las agendas digitales. Las instituciones públicas desempeñan un papel importante en la eliminación de barreras cuando se necesita adoptar una nueva tecnología. A través de sus agendas digitales, regulaciones o la adopción de estándares, ya sea directa o indirectamente, los gobiernos fomentan que las nuevas tecnologías se desplieguen en diferentes horizontes sociales y económicos.

Las agendas digitales suelen hacer referencia a un plan estratégico o a un conjunto de políticas que un gobierno, organización o institución adopta para promover y gobernar el uso de tecnologías digitales en diversos aspectos de la sociedad. Los objetivos específicos y las áreas de enfoque de una agenda digital pueden variar, pero su objetivo principal es aprovechar el potencial de las tecnologías digitales en beneficio de individuos, empresas y la economía en general [36].

Un elemento crucial que debe ser la piedra angular de cualquier agenda digital efectiva es la **Inclusión Digital.** Este principio imperativo tiene como objetivo reducir la brecha digital asegurando que cada segmento de la población, independientemente de su estatus socioeconómico, edad o ubicación geográfica, no solo tenga acceso a las tecnologías digitales, sino que también posea las habilidades necesarias para usarlas de manera efectiva [37].

En este panorama digital en rápida evolución, donde la Inteligencia Artificial y el NLP han encontrado su lugar en la vanguardia de los avances tecnológicos, los gobiernos deben adaptar y ampliar sus agendas digitales. Un aspecto fundamental de esta adaptación es el compromiso con la **promoción y preservación de los idiomas locales**, incluidas las lenguas indígenas, en el ámbito de las tecnologías de IA y NLP.

---

[36] UNESCO, *Agenda Digital Digital Agenda*, 2023, Disponible en: https://es.unesco.org/creativity/policy-monitoring-platform/agenda-digital-digital-agenda

[37] DISCOVER, *La inclusión digital: ¿qué es y por qué es importante?,* 2023, Disponible en: https://www.discoverdigital.eu/lms-es/courses/discover-digital/online-training/lessons/la-inclusion-digital-que-es-y-por-que-es-importante/

Para aprovechar plenamente el potencial de la IA y NLP, los gobiernos no solo deben incluir los idiomas locales y lenguas indígenas en sus agendas digitales, sino que también deben **apoyar activamente la investigación, el desarrollo y la adopción** de estas tecnologías con un enfoque en la diversidad lingüística. Además, este enfoque se alinea con los esfuerzos internacionales para promover la diversidad lingüística, la preservación del patrimonio cultural y los principios de equidad digital[38].

6.1.2 Generación de corpus. Como se ha explicado anteriormente, el proceso de generación de un corpus, un recurso fundamental para las aplicaciones de NLP, conlleva cargas técnicas y económicas que pueden ser sustanciales, a menudo planteando desafíos para individuos u organizaciones interesadas en el desarrollo de aplicaciones de NLP.

Sin embargo, los gobiernos tienen la capacidad de asignar financiamiento y recursos para iniciativas de investigación y desarrollo. Cuando se trata de la creación de corpus para aplicaciones de NLP, el financiamiento gubernamental puede aliviar significativamente la carga financiera. Al proporcionar subvenciones, subsidios de investigación o financiamiento específico para proyectos, los gobiernos pueden empoderar a instituciones académicas y organizaciones de investigación para llevar a cabo proyectos integrales de desarrollo de corpus sin las limitaciones de presupuestos limitados. Este apoyo financiero, a su vez, fomenta la producción de corpus de alta calidad.

Dicho esto, los gobiernos pueden desempeñar un papel crucial para abordar este obstáculo al proporcionar apoyo financiero, colaborar con instituciones académicas o establecer asociaciones con entidades externas para facilitar la creación de corpus. Esta intervención estratégica no solo facilita las cargas asociadas con la generación de corpus, sino que también fomenta la innovación y avances en el campo.

Las colaboraciones entre gobiernos e instituciones académicas representan una relación simbiótica con un gran potencial. Las instituciones académicas a menudo albergan experiencia en lingüística y NLP, lo que las convierte en socios ideales para proyectos de desarrollo de corpus. Así, los gobiernos pueden entablar asociaciones con estas instituciones para aprovechar sus conocimientos y capacidades de investigación. Esta sinergia puede llevar a una creación de corpus más eficientes, ya que combina conocimientos académicos con recursos y apoyo gubernamentales.

Además, los gobiernos también pueden establecer asociaciones estratégicas con organizaciones externas, incluidas empresas de tecnología o entidades enfocadas en el lenguaje. Estas asociaciones pueden estructurarse para compartir los costos, los recursos y la experiencia necesarios para el desarrollo de corpus. Al involucrar a la iniciativa privada, los gobiernos pueden aprovechar una amplia gama de habilidades y tecnologías, lo que lleva a la creación de corpus más completos y sofisticados.

[38] UNESCO, *The protection and promotion of linguistic diversity addressed by UNESCO*, 2019, Disponible en: https://www.unesco.org/en/articles/protection-and-promotion-linguistic-diversity-addressed-unesco

No obstante, el papel crítico que los gobiernos pueden desempeñar para facilitar la creación de corpus va más allá del mero apoyo técnico y financiero. Pueden contribuir activamente al proceso de construcción del corpus aprovechando sus vastos repositorios de materiales de archivo. Estos repositorios abarcan una gran cantidad de textos digitalizados y archivos de sonido, meticulosamente curados para preservar el patrimonio nacional y la memoria histórica [39].

Los gobiernos son custodios de un tesoro de información histórica y cultural, lo que los convierte en una fuente fundamental para la generación de corpus. Al brindar acceso y compartir estas bases de datos invaluables, los gobiernos no solo pueden impulsar sus propias iniciativas, sino también empoderar a terceros, incluyendo empresas privadas e instituciones académicas, para contribuir al desarrollo de corpus.

Este enfoque colaborativo tiene una importancia particular cuando se trata de la preservación y propagación de las lenguas indígenas. A menudo, las lenguas indígenas están subrepresentadas en los corpus existentes, a pesar de su importancia cultural y lingüística. Cuando los gobiernos abren proactivamente sus archivos y se asocian con partes interesadas, fomentan un entorno en el que las lenguas indígenas pueden integrarse de manera más efectiva en los corpus. Esta inclusividad no solo sirve para salvaguardar la diversidad lingüística, sino que también garantiza que estas lenguas reciban la atención y la preservación que merecen en el ámbito de la Inteligencia Artificial.

6.1.3 Desbloqueo regulatorio

Privacidad de datos. Los gobiernos locales tienen la autoridad para dar forma y regular políticas de protección de datos dentro de sus jurisdicciones. Estas políticas, cuando se elaboran cuidadosamente, pueden ser fundamentales para flexibilizar la recopilación de datos, en particular para facilitar la recopilación de datos con el propósito de generar corpus a través de empresas privadas o de instituciones académicas.

Las regulaciones de protección de datos generalmente buscan encontrar un equilibrio entre la protección de la privacidad del individuo y la promoción de la innovación. Los gobiernos locales desempeñan un papel fundamental en la calibración de este equilibrio para fomentar la recopilación de datos para los corpus de NLP. Al **introducir flexibilidad en las regulaciones**, reconocen la importancia de la investigación y el desarrollo de NLP, al tiempo que aseguran que las preocupaciones respecto a la privacidad de los datos se aborden adecuadamente.

[39] James B. Rhoads, *The Role of archives and records management in national information systems: a RAMP study*, General Information Programme and UNISIS, United Nations Educational, Scientific and Cultural Organization, France, 1989. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000084735

Para prevenir riesgos en la privacidad, las regulaciones pueden incorporar disposiciones que promuevan la anonimización y desidentificación de datos[40]. Esto permite la generación de corpus, al tiempo que mitiga las amenazas a la privacidad. Los gobiernos locales también pueden establecer estándares para el manejo de datos que fomenten la eliminación de información de identificación personal, garantizando que los datos utilizados para los corpus permanezcan anónimos.

Además, los gobiernos locales pueden establecer políticas que especifiquen los períodos de retención y actualización de datos, permitiendo la recopilación de datos necesaria para el desarrollo del corpus en plazos razonables. Estas políticas deben definir durante cuánto tiempo se pueden almacenar los datos con fines de investigación y en qué condiciones deben eliminarse de manera segura.

Finalmente, la transparencia es un principio que debe seguirse para aquellos que generan, mantienen y administran el corpus. Los gobiernos locales deben exigir a las organizaciones e instituciones que revelen sus actividades de recopilación de datos, sus propósitos y las salvaguardias implementadas. También pueden establecer mecanismos de responsabilidad para supervisar el cumplimiento de las regulaciones. Estas medidas proporcionarían certeza a todas las entidades relacionadas en la generación del corpus.

Propiedad intelectual. Por otro lado, el establecimiento de regulaciones de propiedad intelectual bien definidas por parte de los gobiernos es crucial para determinar la propiedad y los derechos de uso de los corpus. Estas regulaciones proporcionan claridad sobre varios aspectos vitales, incluido quién tiene derecho a monetizar el uso del corpus, si el corpus debe estar abierto al público o reservado para entidades privadas, y cómo se gestionan los derechos de propiedad intelectual cuando el gobierno es el creador del corpus.

Las regulaciones de propiedad intelectual deben establecer explícitamente quién posee los derechos de propiedad sobre el corpus. En caso de que el gobierno sea el creador del corpus, éste puede tener la propiedad sobre los datos e incluso abarcar el derecho a monetizar su uso. Esto puede implicar otorgar licencias del corpus a entidades privadas, instituciones académicas o empresas comerciales a cambio de un costo. Por otro lado, cuando el gobierno colabora con entidades externas en la creación del corpus, estas regulaciones deben definir si el gobierno, terceros o ambos tienen la autoridad para monetizar el uso del corpus y del cómo se compartirán los ingresos, si corresponde.

Además, los gobiernos deben tomar una decisión sobre si el corpus debe estar abierto al público o reservado para uso privado, incluso si el gobierno es el creador del corpus o si debería colaborar con una entidad externa. El acceso abierto implica que los datos están disponibles para el público, a menudo bajo licencias permisivas, lo que puede promover la

[40] Amy Isard, *Approaches to the Anonymization of Sign Language Corpora*, Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 95–100, Language Resources and Evaluation Conference (LREC2020), France, 2020.

innovación, la investigación y la inclusión[41]. Por otro lado, restringir el acceso al público puede ser un medio para que quienes tienen derechos sobre el corpus (gobierno, empresas o instituciones académicas) generen ingresos o controlen cómo se utiliza la información. La elección entre acceso abierto y cerrado debe estar en línea con los objetivos y políticas del gobierno, los objetivos de las organizaciones de terceros y las consideraciones de política pública.

Es importante destacar que los corpus con acceso abierto al público son uno de los principales factores que han permitido que las aplicaciones de NLP basadas en el inglés se dispersen en todo el mundo y mantengan la dominancia del idioma inglés en el campo de la Inteligencia Artificial. Sin embargo, se necesita un enfoque caso por caso para definir si el corpus debe estar cerrado o abierto al público, ya que podría infringir los derechos de propiedad intelectual de otros o contener información sensible.

6.1.4 Generación de necesidades de mercado

Para los principales idiomas oficiales, como el español o el francés, existen incentivos evidentes para que las empresas privadas desarrollen aplicaciones NLP para satisfacer una necesidad de mercado específica. Esto se debe principalmente al hecho de que las personas que hablan el idioma en el que se creó la aplicación pueden convertirse en clientes potenciales. Por lo tanto, existe una estrecha correlación directa entre la cantidad de hablantes por idioma y el interés de terceros en desarrollar una aplicación de NLP.

Por ejemplo, Alexa, que es el asistente virtual de Amazon, ha sido desarrollada para interactuar con clientes en inglés (con variantes de EE. UU., Reino Unido y Canadá, con acento australiano o hindi), alemán, japonés, francés (con variantes canadienses y francesas), italiano, español (con variantes mexicanas, estadounidenses y españolas), hindi, portugués y árabe[42]. Esto garantiza que Amazon tenga millones de posibles clientes en diferentes regiones globales.

Sin embargo, este incentivo puede no estar completamente definido en idiomas en los que el número de hablantes es solo unos cientos. Al respecto, según la Revista Ethnologue, los idiomas indígenas del Pacífico (de América del Norte y del Sur), que constituyen el 18.5% de los idiomas del mundo, son hablados por un número tan pequeño de personas que tienen un promedio de solo 1,000 hablantes cada uno[43]. Sin embargo, cuando se consideran colectivamente, representan más de un tercio de los idiomas del mundo. Además, estas comunidades indígenas a menudo enfrentan desafíos económicos y pueden caracterizarse por bajos ingresos o niveles de pobreza. Muchos miembros de estas

[41] Peter Suber, *What Is Open Access?,* MIT Press, 2019. Disponible en: https://openaccesseks.mitpress.mit.edu/pub/6y6fc8k5/release/2

[42] Alexa, *AVS International*, Alexa, 2023, https://developer.amazon.com/en-US/alexa/devices/alexa-built-in/development-resources/international#:~:text=Alexa%20can%20interact%20with%20customers,Portuguese%2C%20and%20SA%2DArabic.

[43] Ethnologue, *What continents have the most indigenous languages?*, 2023. Disponible en: https://www.ethnologue.com/insights/continents-most-indigenous-languages/

comunidades (o toda la comunidad misma) no tienen siquiera acceso a dispositivos digitales, lo que crea tanto una brecha económica como una brecha de acceso. Esta falta de acceso se traduce a que estas comunidad no pueden usar aplicaciones de NLP incluso si estuvieran disponibles. Esto, a su vez, resulta en una brecha de apropiación, donde hay poco o ningún interés en adquirir aplicaciones de NLP por partes de las comunidades.

Como resultado, desde la perspectiva de los desarrolladores y la iniciativa privada, puede haber poca atracción para crear aplicaciones de NLP para idiomas con un potencial de mercado tan limitado. La falta de demanda, combinada con las brechas económicas y de acceso, hace que sea menos viable para el sector privado invertir en el desarrollo de aplicaciones de NLP para estos idiomas. Un fenómeno tangible que ilustra esta situación es lo que sucede actualmente en el sector de las telecomunicaciones: no hay incentivos para que los operadores de telecomunicaciones lleguen a zonas remotas donde se encuentran pequeñas comunidades, coincidentemente donde se encuentran las comunidades indígenas[44]. El costo de desplegar infraestructura para que los operadores de telecomunicaciones lleguen a estas zonas no es económicamente viable porque no hay un retorno de inversión o éste es muy bajo. Dicho esto, ¿cuál podría ser el enfoque de los gobiernos locales para generar necesidades de mercado e interés para que terceros desarrollen aplicaciones de NLP?

Para abordar este problema, los gobiernos podrían adoptar un enfoque de incentivos para la iniciativa privada, a fin de promover el desarrollo de aplicaciones de NLP en sus idiomas locales. En el centro de esta estrategia se encuentra la propuesta de reducir las obligaciones fiscales para las empresas privadas que emprendan el desarrollo de aplicaciones de NLP en sus idiomas locales. La reducción, expresada como un porcentaje respecto a los impuestos totales pagados por las empresas, sirve como un incentivo motivador para que las entidades privadas inviertan en la diversidad lingüística. El atractivo de la reducción de impuestos actúa como un incentivo persuasivo para la participación corporativa en el desarrollo de tecnología lingüística.

Para garantizar la implementación equitativa y responsable de este programa de incentivos, es fundamentar contar con transparencia y políticas bien definidas. Estas políticas deben elaborarse en estrecha colaboración con las autoridades locales y los actores de la industria. La eficacia de tales políticas depende de su claridad, sin dejar margen para la ambigüedad en los criterios de elegibilidad, los requisitos de cumplimiento y los procedimientos de evaluación.

Además, los gobiernos locales, en el marco del cumplimiento de sus agendas digitales, podrían utilizar el instrumento de **licitaciones públicas** para atraer la participación del sector privado en el desarrollo y la propagación de aplicaciones de NLP. Esta estrategia, si se utiliza de manera estratégica, podría dar lugar a una multitud de aplicaciones, como aquellas de asistentes virtuales diseñadas para personas mayores. El atractivo de las

[44] Utilities One, *The Challenges of Telecommunications Infrastructure Development in Rural Areas*, Utilities One, 2023, Disponible en: https://utilitiesone.com/the-challenges-of-telecommunications-infrastructure-development-in-rural-areas

licitaciones públicas radica en su potencial para atraer a empresas privadas con el incentivo de contratos gubernamentales, alineando así los intereses comerciales con el bienestar de la comunidad en general.

Es importante destacar que el éxito de estos enfoques depende de un sólido marco de evaluación. Las empresas privadas deben proporcionar pruebas concretas del desarrollo y la implementación efectivos de aplicaciones de NLP en lenguas locales. Esto requiere el establecimiento de métricas de rendimiento precisas y puntos de referencia, que servirían como criterios para evaluar la utilidad y relevancia de las aplicaciones para satisfacer las necesidades lingüísticas y de comunicación dentro de la comunidad.

Además, estos enfoques deben alentar a las empresas privadas a adaptar sus aplicaciones para que se ajusten a las idiosincrasias culturales y lingüísticas de la comunidad local. Esto no solo mejora la relevancia cultural de las aplicaciones, sino que también fomenta una conexión más profunda con la comunidad. Es importante recordar que estas comunidades lingüísticas de tamaño modesto pueden no tener una presencia destacada en el escenario mundial, pero protegen una parte significativa de nuestro patrimonio lingüístico compartido[45].

Por otro lado, para garantizar la vitalidad de las lenguas indígenas, los gobiernos locales podrían fomentar su presencia en el ámbito digital dentro del ámbito del gobierno electrónico. Este cambio de paradigma rompería, en parte, con los problemas de brecha digital existentes y buscaría crear aplicaciones de NLP que no solo serían accesibles, sino que también específicamente serían relevantes para las comunidades locales. Las lenguas indígenas, que a menudo corren el riesgo de caer en el olvido, se benefician significativamente de este enfoque integrado.

Un ejemplo ilustrativo de esta estrategia es la creación de traductores de NLP diseñados para superar las barreras lingüísticas entre lenguas indígenas, como el náhuatl[46], y el español. Estos traductores basados en NLP, cuando se integran como interfaces de programación de aplicaciones (API) en aplicaciones digitales públicas, se convierten en herramientas poderosas para facilitar la comunicación y el intercambio cultural. Estas iniciativas son ejemplos de un uso consciente y culturalmente sensible de la tecnología para preservar el patrimonio lingüístico.

Finalmente, los gobiernos locales deben promover de manera proactiva y difundir información sobre este programa de incentivos entre los posibles participantes del sector privado. Estas campañas informativas despertarían el interés y la participación dentro del sector empresarial.

6.1.5 Promoción de las habilidades digitales

[45] Ethnologue, *What continents have the most indigenous languages?*, 2023. Disponible en: https://www.ethnologue.com/insights/continents-most-indigenous-languages/
[46] Nahuatl is the most spoken indigenous language in Mexico.

Los gobiernos, conscientes del potencial transformador de las aplicaciones de NLP, deben emprender activamente una misión para equipar a sus ciudadanos con las habilidades digitales necesarias. Estas habilidades no solo facilitan la utilización efectiva de las aplicaciones de NLP, sino que también empoderan a las personas para participar en una economía y sociedad cada vez más digitalizadas. Dado que las habilidades digitales se han convertido en una competencia fundamental, su desarrollo constituye un mandato gubernamental crucial.

Si bien toda la población se beneficia de unas habilidades digitales mejoradas, los estudiantes representan un grupo esencial para el empoderamiento digital. Las instituciones educativas son los principales lugares donde se cultivan estas habilidades, y los jóvenes estudiantes son más adaptables y **receptivos a los avances tecnológicos**[47]. Equipar a los estudiantes con habilidades digitales, incluida la competencia en aplicaciones de NLP, garantiza no solo que la futura fuerza laboral esté bien preparada para navegar por el entorno digital, sino también que haya personas capaces de desarrollar aplicaciones de NLP en su propio idioma y no solo en inglés.

Para cumplir con esta tarea, los gobiernos deben **colaborar estrechamente con las instituciones educativas** para integrar la Inteligencia Artificial en el plan de estudios. Las aplicaciones de NLP, que sustentan una amplia variedad de interacciones digitales, deben ser un componente fundamental de este marco educativo. Esta alineación entre la educación y la tecnología es crucial para preparar a los estudiantes para los desafíos y oportunidades digitales que se avecinan.

Además, cuando los estudiantes están equipados con el conocimiento para desarrollar aplicaciones de NLP en sus propios idiomas, se fomenta la innovación en lingüística y tecnología. Éstos pueden identificar deficiencias en las herramientas de procesamiento de lenguaje y crear soluciones que estén perfectamente ajustadas a las particularidades de su propio idioma.

Por último, los gobiernos locales no deben limitarse a desarrollar habilidades digitales, sino que también deben fomentar la exploración y la innovación digitales. Las aplicaciones de NLP ofrecen un terreno fértil para la creatividad y la resolución de problemas. Al fomentar un entorno que promueva la experimentación y la innovación, los gobiernos nutren a una generación de estudiantes que no solo son usuarios hábiles, sino también **posibles creadores de aplicaciones de NLP**.

6.2 Sector Privado

Como se discutió anteriormente, la falta predominante de incentivos para que las empresas privadas inviertan en aplicaciones de NLP en idiomas distintos al inglés, particularmente en lenguas indígenas, plantea un desafío profundo. Las empresas privadas a menudo operan dentro de un paradigma orientado hacia las ganancias económicas, lo

[47] Inge Molenaar, Anne Horvers y Rick Dijkstra Young, *Learners' Regulation of Practice Behavior in Adaptive Learning Technologies*, Frontiers, 2019, https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02792/full

cual es totalmente comprensible, ya que sin éstas, las empresas no podrían mantenerse por sí mismas. Por lo tanto, el desarrollo de aplicaciones de NLP en lenguas indígenas, que pueden tener una viabilidad comercial limitada, ofrece incentivos financieros mínimos, disuadiendo con frecuencia la inversión en estos dominios lingüísticos.

Sin embargo, en estos casos excepcionales, surge una **obligación moral y ética** para que las empresas privadas prioricen la búsqueda del **bienestar social**. Esto se refiere a la responsabilidad que las empresas privadas tienen hacia la sociedad, especialmente en casos en los que el bien común supera con creces los retornos financieros inmediatos. La preservación de las lenguas indígenas no es únicamente una preocupación lingüística, sino una obligación moral que resuena con los principios de respeto cultural e inclusión.

Además, la tarea de democratizar la tecnología y promover la inclusión digital ya no es exclusiva del ámbito de los gobiernos locales. En una era en la que el sector privado ejerce un inmenso poder e influencia tecnológica, es imperativo reconocer que el alcance e impacto de la tecnología se extienden mucho más allá de la jurisdicción del sector público. Las empresas privadas, en particular aquellas en el sector tecnológico, cuentan con recursos económicos sustanciales, capacidad innovadora y alcance global. Están posicionadas para ser actores clave en la promoción de la inclusión. Su papel trasciende las actividades puramente económicas; **tienen un impacto más amplio en la sociedad** y, por lo tanto, tienen una responsabilidad ética y social.

Al perseguir aplicaciones de NLP en lenguas indígenas, las empresas privadas contribuyen a la causa más amplia del bienestar social. Esta misión encaja perfectamente con los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, que abarcan diversos aspectos del bienestar humano, la diversidad cultural y las sociedades inclusivas. Al desarrollar aplicaciones de NLP en idiomas distintos al inglés, las empresas privadas dan pasos para:

- **Preservar y promover el patrimonio de las lenguas indígenas,** manteniendo la diversidad cultural. Esto se alinea con el ODS 11 (Ciudades y Comunidades Sostenibles), que vislumbra espacios urbanos culturalmente ricos e inclusivos[48].
- **Fortalecer el acceso a la educación e información**, mejorando las tasas de alfabetización y cerrando las brechas digitales. Tales esfuerzos respaldan directamente el ODS 4 (Educación de Calidad) y el ODS 9 (Industria, Innovación e Infraestructura)[49].

[48] United Nations, *Goal 11: Make cities inclusive, safe, resilient and sustainable*, Sustainable Development Goals, 2023, Disponible en: https://www.un.org/sustainabledevelopment/cities/
[49] United Nations, *Goal 9: Build resilient infrastructure, promote sustainable industrialization and foster innovation*, Sustainable Development Goals, 2023, Disponible en: https://www.un.org/sustainabledevelopment/education/

- **Reducir las desigualdades,** abordando el ODS 10 (Reducción de las Desigualdades) al garantizar **la inclusión lingüística y digital**[50].

6.3 La Academia

Las instituciones académicas desempeñan un papel fundamental en el desarrollo de corpus y aplicaciones de NLP, especialmente en el contexto de los idiomas locales. Su función es crucial en el fomento de la innovación y el avance en el estado del arte en NLP. Esto se debe a que las instituciones educativas y los centros de investigación son espacios naturales de innovación, que constantemente empujan los límites del conocimiento y la tecnología. Poseen experiencia de vanguardia en NLP y a menudo colaboran con otras instituciones académicas para avanzar **colectivamente en el campo de la inteligencia artificial**. Su capacidad para impulsar la investigación y la innovación es inigualable.

A pesar de su papel crucial, existe un sesgo innegable dentro de las instituciones académicas hacia los corpus y aplicaciones de NLP basados en inglés. Esta preferencia ha acelerado la dominancia del inglés en el ámbito del NLP. Si bien la importancia global del inglés es innegable, debe ser preservada la diversidad lingüística del mundo.

Para cerrar esta brecha lingüística, las instituciones educativas pueden adoptar dos enfoques estratégicos: modificar los planes de estudio para alinearse con la inteligencia artificial y promover la creación de corpus y aplicaciones de NLP en idiomas locales.

6.3.1 Modificación de los planes de estudio para la compatibilidad con la Inteligencia Artificial (IA): Para abordar la brecha lingüística, las instituciones educativas pueden comenzar por modificar sus planes de estudio, especialmente en bachillerato y en los cursos de pregrado para que sean compatibles con la IA. Esto implica integrar el NLP y la IA en el núcleo de sus programas, con un enfoque en los idiomas locales. De esta manera, las instituciones educativas equiparían a los estudiantes con las habilidades necesarias para trabajar en aplicaciones de NLP adaptadas a los idiomas locales:

- Diversificación de cursos. Ofrecer una amplia variedad de cursos que incluyan NLP, IA y lingüística computacional puede alentar a los estudiantes a explorar aplicaciones en idiomas locales.
- Oportunidades de investigación: Fomentar oportunidades de investigación en idiomas locales puede estimular aún más el interés y la innovación de los estudiantes en NLP.

6.3.2 Promoción de corpus y aplicaciones de NLP en idiomas locales: Las instituciones educativas pueden participar activamente en la creación de corpus y aplicaciones de NLP en sus idiomas locales. Esto implica tanto la investigación como el desarrollo de

[50] United Nations, *Goal 10: Reduce inequality within and among countries*, Sustainable Development Goals, 2023, Disponible en: https://www.un.org/sustainabledevelopment/education/

aplicaciones en colaboración con otros actores, como comunidades locales, organismos gubernamentales y empresas privadas:

- Iniciativas de investigación. Fomentar a profesores y estudiantes para llevar a cabo investigaciones en idiomas locales y su aplicación en NLP. Deben recopilar datos y construir sus propios conjuntos de herramientas para procesar sus propios idiomas.

- Involucramiento comunitario. Colaborar con comunidades locales y hablantes de idiomas indígenas para comprender sus necesidades lingüísticas y crear en conjunto aplicaciones de NLP.

- Asociaciones intersectoriales. Establecer asociaciones con agencias gubernamentales y empresas privadas para financiar y respaldar proyectos de NLP en idiomas locales.

7. El camino hasta ahora: hacia un ambiente NLP multilingüe e inclusivo

El camino hacia un panorama de NLP más inclusivo y multilingüe ha estado marcado por numerosos hitos y conocimientos críticos. Se han emprendido esfuerzos para garantizar que las personas, independientemente del idioma que hablen, puedan acceder a aplicaciones de NLP de alta calidad y participar plenamente en el mundo digital. Más allá de la mera elección de usar o no el inglés en el NLP, ha surgido una necesidad apremiante de democratizar el acceso a la tecnología, especialmente en el ámbito de la Inteligencia Artificial.

Actualmente nos encontramos en una era de rápida transformación digital, donde la demanda de tecnologías de vanguardia como el NLP es omnipresente. Sin embargo, con cada nueva tecnología que surge, aparece una brecha de apropiación, dejando atrás a muchas personas. Para ilustrar esto, el caso de la tecnología móvil, que evolucionó en smartphones, ejemplifica cómo el progreso tecnológico puede superar la capacidad de ciertas poblaciones para adaptarse y acceder a estas innovaciones. El mundo de la Inteligencia Artificial, donde reside el NLP, está experimentando una evolución similar. A medida que esta tecnología avanza, se vuelve imperativo asegurar que la inclusión no sea simplemente una idea de “buena voluntad”, sino una parte integral de su desarrollo.

Los sesgos lingüísticos en el NLP han contribuido a la predominancia del inglés, pero erradicar el uso del inglés en diversos ámbitos no es práctico ni productivo. En cambio, el enfoque debe cambiar hacia la promoción del uso de los idiomas diferentes al inglés en los desarrollos de NLP, garantizando así que la gran mayoría de la población mundial pueda acceder a aplicaciones de NLP de alta calidad, al igual que los hablantes del inglés. Para abordar este desafío, es esencial un enfoque de múltiples partes interesadas. Los

gobiernos locales y nacionales, las instituciones públicas, las entidades del sector privado y las universidades desempeñan cada uno un papel único en fomentar aplicaciones de NLP multilingües.

El camino hacia un ámbito de NLP multilingüe e inclusivo está en curso. El éxito en este esfuerzo depende de políticas claras, transparencia, colaboración y de un compromiso de preservación de la diversidad lingüística. A medida que avanzamos, debemos reconocer la importancia de nuestro idioma como parte fundamental del patrimonio humano y la identidad cultural. El multilingüismo en el NLP no es solo una cuestión técnica; es un reflejo de nuestro compromiso de garantizar que la tecnología sea accesible y benéfica para todos, sin importar el idioma que hablemos.

Este camino está lejos de terminar, pero es un viaje que vale la pena emprender para asegurar que el mundo digital realmente pertenezca a todos.

8. Referencias

- [Bird, Klein, Loper, 2009] Steven Bird, Ewan Klein, y Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009: ix, 39-73, 80-116.

- [Carpenter, 1997] Bob Carpenter. *Type-Logical Semantics*. MIT Press, 1997.

- [Cheng, 2023] Raymond Cheng. *Understanding TF-IDF: A Traditional Approach to Feature Extraction in NLP*. Medium, 2023. Disponible en: https://towardsdatascience.com/understanding-tf-idf-a-traditional-approach-to-feature-extraction-in-nlp-a5bfbe04723f

- [Chierchia and McConnell-Ginet, 1990] Gennaro Chierchia and Sally McConnell-Ginet. *Meaning and Grammar: An Introduction to Meaning*. MIT Press, Cambridge, MA, 1990.

- [Chomsky, 1965] Noam Chomsky. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA, 1965.

- [Chomsky, 1970] Noam Chomsky. *Remarks on nominalization*. In R. Jacobs and P. Rosenbaum, editors, Readings in English Transformational Grammar. Blaisdell, Waltham, MA, 1970.

- [Chomsky and Halle, 1968] Noam Chomsky y Morris Halle. *The Sound Pattern of English*. New York: Harper and Row, 1968.

- [Church and Patil, 1982] Kenneth Church y Ramesh Patil. *Coping with syntactic ambiguity or how to put the block in the box on the table.* American Journal of Computational Linguistics, 8:139–149, 1982.

- [Cohen and Hunter, 2004] K. Bretonnel Cohen y Lawrence Hunter. *Natural language processing and systems biology.* In Werner Dubitzky and Francisco Azuaje, editors, Artificial Intelligence Methods and Tools for Systems Biology, page 147–174. Springer Verlag, 2004.

- [Cole, 1997] Ronald Cole, editor. *Survey of the State of the Art in Human Language Technology*. Studies in Natural Language Processing. Cambridge University Press, 1997.

- [Copestake, 2002] Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA, 2002.

- [DISCOVER, 2023] DISCOVER. *La inclusión digital: ¿qué es y por qué es importante?* 2023. Disponible en: https://www.discoverdigital.eu/lms-es/courses/discover-digital/online-training/lessons/la-inclusion-digital-que-es-y-por-que-es-importante/

- [Enrique Hamel, 2007] Rainer Enrique Hamel. *The dominance of English in the international scientific periodical literature and the future of language use in science.* AILA Review 20, 2007: 53-71, 56.

- [Ethnologue, 2023a] Ethnologue. *What are the top 200 most spoken languages?* 2023. Disponible en: https://www.ethnologue.com/insights/ethnologue200/

- [Ethnologue, 2023b] Ethnologue. *What continents have the most indigenous languages?* 2023. Disponible en: https://www.ethnologue.com/insights/continents-most-indigenous-languages/

- [Guo, 2018] Philip J. Guo. *Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities.* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018: 2.

- [Hamel, 2007] Rainer Enrique Hamel. *The dominance of English in the international scientific periodical literature and the future of language use in science.* AILA Review 20, 2007: 53-71, 56.

- [IBM, 2023] IBM. *What is natural language processing (NLP)?* 2023. Disponible en: https://www.ibm.com/topics/natural-language-processing

- [Kakarla, 2019] Swaathi Kakarla. *Natural Language Processing: NLTK vs spaCy.* Active State, 2019. Disponible en: https://www.activestate.com/blog/natural-language-processing-nltk-vs-spacy/

- [Kapronczay, 2019] Mor Kapronczay. *Text preprocessing in different languages for Natural Language Processing in Python. Part II – Case of Study*. Medium, 2019. Disponible en: https://medium.com/starschema-blog/text-preprocessing-in-different-languages-for-natural-language-processing-in-python-fb106f70b554

- [Klein, 2009] Ewan Klein y Edward Loper. *Natural Language Processing with Python.* O'Reilly, 2009: ix, 39-73, 80-116.

- [Klosowski, 2023] Thorin Klosowski. *The State of Consumer Data Privacy Laws in the US (And Why It Matters).* The New York Times, 2023. Disponible en: https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/

- [Kornilova and Guo, 2023] Anastassia Kornilova y April Guo. *Adapting language-based models beyond English.* Snorkel, 2023. Disponible en: https://snorkel.ai/adapting-language-based-models-beyond-english/

- [Manning and Schütze, 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

- [Molenaar, Horvers, Dijkstra, 2019] Inge Molenaar, Anne Horvers, and Rick Dijkstra Young. *Learners' Regulation of Practice Behavior in Adaptive Learning Technologies.*

Frontiers, 2019. Disponible en: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02792/full

- [Niebel, 2021] Crispin Niebel. *The impact of the general data protection regulation on innovation and the global political economy.* Computer Law & Security Review, ELSEVIER, 2021. Disponible en: https://www.sciencedirect.com/science/article/abs/pII/S026736492030128X

- [NLTK, 2023] NLTK. *Natural Language Toolkit*. 2023. Disponible en: https://www.nltk.org/

- [ONU Noticias, 2023] ONU Noticias. *La población mundial alcanzó hoy 15 de noviembre de 2022 las 8000 millones de personas de acuerdo al informe Perspectivas de la Población Mundial, que también prevé que India superará a China como el país más poblado del mundo en 2023. El bebé 8 mil millones nació en República Dominicana.* ONU Habitat, 2023. Disponible en: https://onuhabitat.org.mx/index.php/ya-somos-8-mil-millones-de-personas#:~:text=La%20población%

- [Shapiro and Varian, 1999] Carl Shapiro y Hal R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Review Press, 1999.

- [Suber, 2019] Peter Suber. *What Is Open Access?* MIT Press, 2019. Disponible en: https://openaccesseks.mitpress.mit.edu/pub/6y6fc8k5/release/2

- [Turing, 2023] Turing. *Which Language Is Useful for NLP and Why?* 2023. Disponible en: https://www.turing.com/kb/which-language-is-useful-for-nlp-and-why

- [United Nations, 2023a] United Nations. *Goal 11: Make cities inclusive, safe, resilient and sustainable*. Sustainable Development Goals, 2023. Disponible en: https://www.un.org/sustainabledevelopment/cities/

- [United Nations, 2023b] United Nations. *Goal 4: Quality Education. Sustainable Development Goals, 2023*. Disponible en: https://www.un.org/sustainabledevelopment/education/

- [United Nations, 2023c] United Nations. *Goal 9: Build resilient infrastructure, promote sustainable industrialization and foster innovation*. Sustainable Development Goals, 2023. Disponible en: https://www.un.org/sustainabledevelopment/education/

- [United Nations, 2023d] United Nations. *Goal 10: Reduce inequality within and among countries*. Sustainable Development Goals, 2023. Disponible en: https://www.un.org/sustainabledevelopment/education/

- [Warschauer, 2000] Mark Warschauer. The Changing Global Economy and the Future of English Teaching. TESOL Quarterly, 2000: 511-535.

- [Wilame, 2018] Wilame. *spaCy basic features: comparing performance for Portuguese, French and English*. Medium, 2018. Disponible en: https://medium.com/@wila.me/spacy-basic-features-comparing-performance-for-portuguese-french-and-english-bb2edab49b4

- [Y Studios, 2018] Y Studios. *The Language of Codes: Why English is the Lingua Franca of Programming*, 2018. Disponible en: https://ystudios.com/insights-passion/codelanguage