LA INGENIERÍA LINGÜÍSTICA Y EL CORPUS PARALELO DE LENGUAS MEXICANAS: PRESERVACIÓN Y DESARROLLO TECNOLÓGICO

GERARDO SIERRA

Introducción

En México, con una población de más de 126 millones de habitantes, aproximadamente 25 millones se reconocen como indígenas y más de siete millones hablan una de las 68 lenguas indígenas del país. Estas lenguas son fundamentales para la cohesión social, la inclusión y el desarrollo sostenible, según la "Declaración de Los Pinos" para el Decenio de las Lenguas Indígenas (2022-2032). Este reconocimiento no sólo es cultural, también, tecnológico, donde la ingeniería lingüística juega un papel crucial.

La ingeniería lingüística es una disciplina que aplica principios de la ingeniería para resolver problemas relacionados con el lenguaje, incluyendo el desarrollo de tecnologías que facilitan la traducción, la recuperación de información y la preservación de lenguas. Un ejemplo destacado de este ámbito es el desarrollado por Google, actualmente incluye en su traductor automático las lenguas indígenas. Por su parte, el Grupo de Ingeniería Lingüística de la Coordinación de Eléctrica y Computación ha desarrollado el Corpus Paralelo de Lenguas Mexicanas (CPLM), un recurso que alinea el español con seis lenguas indígenas: maya, ch'ol, mazateco, mixteco, otomí y náhuatl.

El CPLM es un corpus paralelo, es decir, una colección de textos en un idioma con sus respectivas traducciones en otro. Este tipo de corpus resultan esenciales para diversas herramientas tecnológicas, como puede ser la traducción automática, la construcción de diccionarios bilingües informatizados y los sistemas multilingües de recuperación de información. El CPLM no solamente facilita el estudio y la preservación de lenguas indígenas, también, promueve su uso en tecnologías modernas.

Desarrollo del CPLM

A continuación, se habla del proceso de construcción del CPLM, pasando por distintas etapas de su desarrollo.

El primer paso en la construcción del CPLM fue recopilar textos en las seis lenguas seleccionadas. Se recopilaron diversos textos, clasificados en siete géneros, lo que permitió obtener una base sólida y diversa de textos para el CPLM: didáctico, expositivo, narrativo, poético, histórico, político y religioso. Los textos didácticos están relacionados con manuales de escritura y aprendizaje; los expositivos explican enfermedades o formas de cultivo; los narrativos cuentan historias tradicionales y cuentos de la vida diaria; los poéticos están escritos en verso; los históricos presentan historias populares de comunidades y fundaciones.

El segundo paso fue la digitalización de los textos físicos con *software* de Reconocimiento Óptico de Caracteres (OCR). Este proceso presentó desafíos debido a caracteres únicos en lenguas indígenas, requiriendo limpieza y corrección manual. Un reto clave fue la representación textual de tonos y caracteres especiales (Tabla 1) como en el mazateco, mixteco y otomí, que no son reconocidos por OCR tradicionales. Para superar esto, se personalizaron caracteres en el *software* ABBYY FineReader y se usó Unicode para asegurar precisión. La variabilidad en la notación de tonos complicó la digitalización, requiriendo metodologías específicas para cada lengua. Entre las innovaciones clave se incluyeron la personalización del OCR y la creación de plantillas específicas, mejorando la eficiencia y precisión del proceso, aunque la verificación manual siguió siendo esencial.

Tabla 1. Caracteres especiales del CPLM

Vocales	Consonantes
í, ì, ĭ, î, ī, i̱, ị, ị̀, ị́, ị̂, ị̄, ḭ, ḭ́, ḭ̀, ḭ̆, ḭ̂, ḭ̄, ɨ, ɨ́, ɨ̀, ɨ̆, ɨ̂, ɨ̄, ɨ̱, ɨ̣, ɨ̣́ é, è, ĕ, ê, ē, e̱, ẹ, ẹ́, ẹ̀, ẹ̆, ệ, ẹ̄, ḛ, ḛ́, ḛ̀, ḛ̆, ḛ̂, ḛ̄, ɛ, ɛ́, ɛ̀, ɛ̆, ɛ̂, ɛ̄, ɛ̱, ɛ̣, ɛ̣́, ɛ̌, ɛ́̌, æ, ǽ, æ̀, æ̆, æ̂, ǣ, ᴂ, æ, ǽ, æ̀, ᴂ́, ᴂ̀, ᴂ̆, ᴂ̂, ᴂ̄ á, à, ă, â, ā, a̱, ạ, ạ́, ạ̀, ặ, ậ, ạ̄, ä, ǟ, æ, ǽ, æ̀, æ̆, æ̂, ǣ, ᴂ, ᴂ́, ᴂ̀, ᴂ̆, ᴂ̂, ᴂ̄ ó, ò, ŏ, ô, ō, o̱, ọ, ọ́, ọ̀, ọ̆, ộ, ọ̄, ö, ȫ, ȯ, ȫ̄ ú, ù, ŭ, û, ū, u̱, ụ, ụ́, ụ̀, ụ̆, ụ̂, ụ̄, ü, ǘ, ǔ, ǖ, ṻ, ü̱, ụ̈, ụ̈́, ụ̈̀, ụ̈̆, ụ̈̂, ụ̈̄	č, Č ȼ ḅ, b̲ J̌, ǰ, Ĵ, ĵ m̲ Nasalización, palatalización, velarización, aspiración ^{n}, ^{J}, ^{W}, ^{h}

Con los datos digitalizados, el siguiente paso fue la alineación de textos, esto es, el proceso de emparejar correspondencias bilingües a nivel de documento, párrafo, oración o palabra. En el CPLM, se utilizó una combinación de algoritmos y un proceso de verificación manual para asegurar la precisión de las alineaciones. Un aspecto crítico de la alineación es la gestión de diferentes niveles de granularidad. Por ejemplo, la alineación a nivel de oración requiere algoritmos precisos para identificar correspondencias exactas entre oraciones en diferentes idiomas.

Una vez que el CPLM fue obtenido, fue necesario buscar una plataforma dónde depositarlo y resguardarlo para su

posterior consulta. Escogimos el gestor de corpus GECO, desarrollado previamente en el Instituto de Ingeniería para corpus no paralelos. La nueva plataforma GECO para corpus paralelos, desarrollada con Django y PostgreSQL, permite a los colaboradores subir, visualizar y buscar textos en el corpus. Utiliza un motor de búsqueda avanzado que maneja la diversidad de caracteres y las variantes lingüísticas, facilitando búsquedas complejas y eficientes. La arquitectura del sistema incluye un *backend* robusto con Django y una base de datos PostgreSQL, capaz de manejar textos de longitud ilimitada y optimizar consultas multilingües. El *frontend*, diseñado con Bootstrap, ofrece una interfaz intuitiva para los usuarios, permitiendo búsquedas detalladas y visualización de concordancias. El sistema de recuperación de información implementado en Python utiliza expresiones regulares para gestionar las búsquedas de palabras y frases, abarcando comodines y distancias entre palabras.

El desarrollo de un sistema de recuperación de información eficiente es crucial para investigadores y usuarios que necesitan explorar el CPLM en profundidad. Para asegurar que los usuarios puedan acceder y utilizar el corpus de manera efectiva, la nueva plataforma GECO implementó búsquedas avanzadas con expresiones regulares y la capacidad de manejar diferentes variantes lingüísticas, lo que permite realizar consultas detalladas y específicas, incluyendo búsquedas de palabras con comodines y búsquedas de frases con distancias específicas entre palabras.

Se creó un sitio *web* donde cualquier usuario puede acceder para hacer consultas (http://www.corpus.unam.mx/cplm). En la cuestión del diseño del logo de la página *web*, se consideraron aspectos importantes de las lenguas indígenas en México, como se muestra en la Imagen 1.

Aportes del CPLM

Las lenguas indígenas de México presentan restricciones experimentales, ya sea de alineación o de cualquier tipo de enfoque, debido a la escasez de contenido digitalizado. No obstante, el desarrollo del proyecto demostró que es posible llevar a cabo distintos tipos de acercamientos a este tipo de lenguas; por ejemplo, la estimación de correspondencias léxicas bilingües con corpus paralelos pequeños mediante análisis estadísticos y las representaciones textuales adaptadas a las características específicas de cada lengua.

El CPLM contribuye significativamente a la preservación de las lenguas indígenas al proporcionar una herramienta para su estudio, preservación y divulgación. Además, facilita la

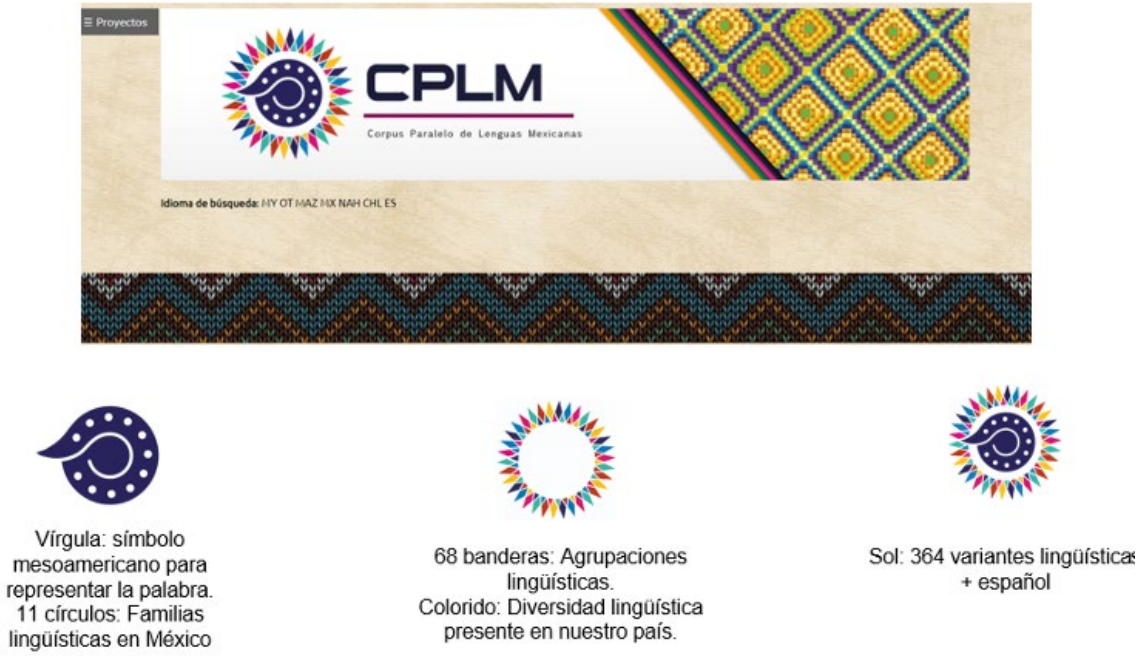

Imagen 1. Explicación del logo de CPLM

creación de tecnologías del lenguaje que pueden ser utilizadas en ámbitos como la justicia y la salud, donde la comunicación precisa es vital.

Otro avance se presentó ante la limitada disponibilidad de textos digitales en lenguas indígenas, lo cual, representa un obstáculo mayúsculo. Para mitigar esto, el equipo del CPLM desarrolló técnicas de análisis estadístico y construcción de modelos que toman en cuenta las particularidades fonológicas, morfológicas, sintácticas y semánticas de cada lengua. Estos modelos compensan la escasez de datos, permitiendo la estimación precisa de correspondencias léxicas.

Un corpus paralelo es una fuente invaluable para el desarrollo de sistemas de traducción automática y herramientas de procesamiento de lenguaje natural. Estos avances no sólo benefician a las comunidades indígenas al proporcionar acceso a servicios en su lengua, también, enriquecen el campo de la ingeniería lingüística. El CPLM es esencial para el desarrollo de sistemas de traducción automática basados en aprendizaje automático, como hoy en día se puede observar con el traductor de Google. Estos sistemas pueden ser entrenados con datos del corpus para aprender las correspondencias entre lenguas indígenas y el español, permitiendo traducciones más precisas y contextualmente adecuadas. Los sistemas de traducción automática que utiliza el CPLM pueden ser especialmente útiles en aplicaciones móviles y en línea, proporcionando traducciones instantáneas y accesibles, facilitando la comunicación entre hablantes de diferentes lenguas.

El CPLM también tiene aplicaciones en la educación, proporcionando recursos para el aprendizaje de lenguas indígenas. Las plataformas educativas pueden utilizar el corpus para desarrollar materiales didácticos y herramientas interactivas que faciliten el aprendizaje de estas lenguas, tanto para hablantes nativos como para aquellos interesados en aprenderlas. Las aplicaciones educativas basadas en el CPLM pueden incluir cursos en línea, aplicaciones móviles de aprendizaje de lenguas y recursos interactivos. Asimismo, la adquisición léxica automática que utiliza el corpus para construir diccionarios bilingües, es esencial para el desarrollo de aplicaciones educativas y de asistencia lingüística. Estas herramientas pueden ayudar a revitalizar lenguas en peligro de extinción y promover mayor interés en la diversidad lingüística de México.

En el ámbito de la salud, el CPLM puede ser utilizado para desarrollar herramientas de asistencia lingüística que permitan a los profesionales de la salud comunicarse eficazmente con pacientes que hablan lenguas indígenas. Esto es crucial para asegurar que los pacientes reciban una atención adecuada y comprendan plenamente las instrucciones médicas. Las herramientas de asistencia lingüística, como las aplicaciones de traducción en tiempo real y los diccionarios médicos bilingües, pueden mejorar significativamente la calidad de la atención médica para hablantes de lenguas indígenas. Estas herramientas pueden ser integradas en sistemas de salud pública para proporcionar un servicio más inclusivo y equitativo.

La preservación de lenguas indígenas a través del CPLM no sólo mantiene vivo el patrimonio cultural, también empodera a las comunidades al ofrecerles herramientas tecnológicas que respetan y reflejan su identidad lingüística. Regresando al ámbito jurídico, el acceso a documentos legales en lenguas indígenas puede mejorar significativamente la equidad y la justicia.

Una de las principales oportunidades para el CPLM es su expansión para incluir más lenguas indígenas y mayor variedad de textos. Esto no sólo aumentará la riqueza del corpus, también, proporcionará más datos para mejorar los algoritmos de traducción y el procesamiento de lenguaje natural. La inclusión de más lenguas y textos requiere una metodología robusta para la recopilación y digitalización de datos. Esto puede implicar colaboraciones con comunidades indígenas y organizaciones culturales para asegurar la disponibilidad de textos y la representatividad de diversas lenguas.

Conclusiones

El desarrollo continuo de algoritmos para la alineación y el procesamiento de textos multilingües es esencial. Los avances en técnicas de aprendizaje automático y procesamiento de lenguaje natural pueden mejorar significativamente la precisión y eficiencia del CPLM. La integración de algoritmos de aprendizaje profundo, como redes neuronales recurrentes y transformadores, puede ofrecer mejoras sustanciales en la alineación y traducción de textos. Estos algoritmos, que aprenden de grandes volúmenes de datos, pueden manejar mejor las complejidades y variaciones lingüísticas presentes en el CPLM.

El Corpus Paralelo de Lenguas Mexicanas representa un avance significativo en la ingeniería lingüística y la preservación de las lenguas indígenas. Su desarrollo no sólo contribuye al conocimiento y a la preservación de estas lenguas, también, abre nuevas oportunidades para su uso en tecnologías avanzadas. Al integrar el conocimiento ancestral con innovaciones tecnológicas, el CPLM demuestra que la preservación cultural y el avance tecnológico pueden coexistir y enriquecerse mutuamente.

La ingeniería lingüística, a través del CPLM, está preservando las lenguas indígenas de México, además, está creando un puente entre el pasado y el futuro, entre la tradición y la innovación. Este proyecto es un testimonio del poder de la tecnología para promover la diversidad y la inclusión; es un ejemplo inspirador de cómo la ingeniería puede contribuir a un mundo más justo y equitativo.