¿CÓMO ESTUDIAMOS Y APORTAMOS A LAS LENGUAS INDÍGENAS DESDE LA COMPUTACIÓN?

Anaís M. Almendra C.[1]

En referencia a la relación de la computación con las lenguas indígenas una de las cuestiones en las que se puede coincidir preliminarmente es que la primera es una aliada. En específico, las infraestructuras computacionales constituyen un posible gran apoyo para el estudio de las lenguas indígenas. Sin embargo, y aunque pueda aceptarse esta idea sin mucha discusión, es necesario revisar en concreto cómo la computación puede contribuir en el área de la lingüística de lenguas indígenas. En específico, se abordará el caso de esta afirmación en relación con el mapudungun.

En primer lugar, es importante entender cómo funciona esta lengua. Uno de los fragmentos más esclarecedores respecto a mapudungun es el siguiente:

> El mapudungun (también conocido en la literatura como mapuche o araucano) es una lengua de la periferia andina no clasificada que se habla con distintos grados de vitalidad en el centro y sur de Chile y el sur de Argentina, a ambos lados de la Cordillera de los Andes. Con respecto a sus características tipológicas generales, se trata de una lengua aglutinante y sufijadora con tendencia a la polisíntesis (Golluscio y Hasler, 2017, p. 70).

Desglosemos lo anterior un poco. Tanto Chile como Argentina tienen hablantes de mapudungun, aunque en algunos sectores la vitalidad de la lengua es mayor que en otros. Según Gundermann et. al (2011) la vitalidad del mapudungun es desfavorable, dándose zonas en las que el mapudungun ha desaparecido debido a un gran número de factores contextuales, sociales y similares.

---

[1] Licenciada en Lingüística y Literatura Hispánicas con mención en Lingüística y estudiante de magister en Lingüística con mención en Lengua Española de la Universidad de Chile. Profesora de Educación Media en Lenguaje y Comunicación por la Pontificia Universidad Católica.

Por otro lado, es necesario observar los aspectos morfológicos del mapudungun. Sin embargo, primeramente, es necesario explicar el significado de morfema. Según Escandell (2011), morfema es la unidad mínima de significado y la morfología es el estudio de la estructura interna de las palabras, que es donde están los morfemas. La morfología de una lengua puede ser tipologizada según dos aspectos: su síntesis y su fusión. La síntesis refiere a cuántos morfemas puede incorporar una sola palabra, y la fusión se dictamina según si un morfema tiene un único significado o varios (Pavey, 2010). Como se indica en Golluscio y Hasler (2017), el mapudungun es aglutinante en su fusión y polisintética en su síntesis. Ello quiere decir que, en términos de lo aglutinante de la lengua, esta tiene partes con significados únicos, es decir, cada morfema del mapudungun solo puede indicar una única información (Pavey, 2010). Por otro lado, tiende a ser una lengua polisintética, que quiere decir que muchas de esas partes con significados únicos se pueden unir a una misma raíz (Pavey, 2010). Imaginemos a cada palabra de una lengua como un tren y a cada vagón del tren como un morfema. La naturaleza aglutinante del mapudungun significa que cada vagón tiene un único tipo de carga o mercancía. Por otro lado, la naturaleza polisintética implica que un tren puede llevar muchos carriles al mismo tiempo. Para ejemplificar ello sería interesante revisar la palabra 'ilotukelan' en mapudungun, que en español es una oración completa:

Ilo - tu - ke - la - n

Carne - VERB. - HAB. - NEG. - IND.1.SG

Traducción: yo no como carne.

[2]

[2] Abreviaturas: VERB. Verbalizador, HAB. Habitual, NEG. Negación, IND. Indicativo, 1 Primera persona y SG Singular.

En el ejemplo precedente observamos a la raíz ‘ilo-' cuyo significado es ‘carne’. A esta raíz se le añaden cuatro morfemas: ‘-tu’, ‘-ke’, ‘-la’ y ‘-n’. Cada una de estas unidades gramaticales tiene un significado único y gracias a la unión de todas estas se obtiene el sentido de la palabra.

Si comparamos la naturaleza morfológica de la lengua mapuche con el español podemos encontrar que difieren en los dos aspectos ya mencionados. Mientras el mapudungun es una lengua aglutinante y polisintética, el español es una lengua flexiva y sintética, que es el término medio entre una lengua aislante y una polisintética (Pavey, 2010). Esto quiere decir que, al contrario del mapudungun, el español solo puede añadir un par de unidades morfológicas en su estructura —naturaleza sintética—, y que los morfemas que tiene tienden a tener más de un significado —naturaleza flexiva o fusionante—.

Sin embargo ¿por qué es relevante tener en cuenta las diferencias entre el español y el mapudungun si la premisa es que la computación es una aliada? Entender las diferencias entre estas lenguas es esencial ya que muchas de las herramientas actuales de estudio lingüístico están enfocadas en lenguas mayoritarias, como el español y el inglés. Conocer las diferencias entre el español y el mapudungun permite buscar formas de adaptar estas herramientas o saber cómo crear nuevas. A su vez, si gran parte de las personas mapuches tienen como lengua materna el español y buscan aprender mapudungun, es necesario otorgar conocimientos y herramientas que permitan establecer conexiones entre la lengua base y la meta. Entender estas diferencias facilita el camino en el que se puede ver a la computación como una aliada en la lingüística de lenguas indígenas.

¿Qué son las infraestructuras computacionales?

Una vez comprendidos algunos de los aspectos centrales del mapudungun, es posible abordar un concepto que, a primera vista, podría parecer más distante y complejo: las infraestructuras computacionales.

En vista de la tecnología actual, quiero recurrir a una definición creada por una IA que, precisamente, se sustenta en una infraestructura computacional.

Según ChatGPT versión 4o las infraestructuras computacionales se definen y relacionan con la lingüística de la siguiente forma:

> La infraestructura computacional es el conjunto de hardware, software y sistemas que permiten el procesamiento y almacenamiento de datos en entornos digitales. En el contexto lingüístico, incluye herramientas como corpus digitales, analizadores morfológicos y modelos de inteligencia artificial que facilitan el estudio y procesamiento de lenguas. (03 de marzo 2025)

Es decir, una infraestructura computacional es la base, el 'radier', el sustento de las diversas herramientas computacionales que conocemos. Por ejemplo, TikTok opera gracias a una infraestructura computacional compuesta por una serie de dispositivos de hardware y software que sustentan sus bases de datos, videos, algoritmos y otros componentes similares. Ello, en su conjunto, son una infraestructura computacional que concretizan el *scrolleo* incesante de sus usuarios.

Así como la red social ya mencionada, las diversas aplicaciones y herramientas computacionales requieren de una infraestructura específica para su funcionamiento. Estas varían según los recursos, necesidades y objetivos de sus creadores, por lo que una herramienta para el estudio lingüístico puede tener una infraestructura computacional muy parecida a Tiktok o, por el contrario, muy diferente a esta.

Infraestructuras computacionales para el mapudungun

Existe una gran variedad de herramientas que son útiles para el estudio lingüístico, desde plataformas que reproducen el sonido del AFI, hasta aplicaciones que permiten pulir la escritura de artículos. Hay plataformas dedicadas a áreas específicas de la lingüística, a lenguas particulares o un poco de todo. Debido a la amplitud de infraestructuras computacionales disponibles para el estudio lingüístico es que se opta por acotar este campo. Se abordarán solo aquellas infraestructuras dedicadas a nuestra ya comentada lengua, el mapudungun.

Es posible encontrar diversas aplicaciones, plataformas y recursos similares para el estudio del mapudungun, que incluyen diccionarios, juegos o repositorios. Sin embargo, solo nos enfocaremos en aquellas herramientas que buscan ayudar a desentrañar los morfemas de una palabra en mapudungun. Al ser la morfología de la lengua mapuche una fuente muy rica de analizar, este texto se focalizará en estudiar las infraestructuras computacionales dedicadas al análisis morfológico.

En el caso de las infraestructuras computacionales orientadas al análisis morfológico del mapudungun, su tipología aglutinante representa una ventaja significativa. El hecho de que los morfemas en esta lengua posean significados diferenciados y específicos facilita su segmentación y análisis morfológico automatizado. Aunque la variedad de morfemas puede ser amplia, esta no representa una dificultad para los sistemas computacionales, cuya capacidad de procesamiento permite manejar grandes volúmenes de información. Más que la cantidad, el principal desafío en este tipo de análisis es la ambigüedad[3]. En este sentido, la estructura aglutinante del mapudungun puede contribuir a reducirla, convirtiéndose en un factor favorable para el procesamiento computacional.

En vista de lo anterior, han surgido herramientas para el mapudungun que realizan la segmentación de palabras y el etiquetado de cada morfema analizado. Todo ello se logra mediante una interfaz en la que un usuario introduce una palabra en la plataforma y esta entrega de resultado un análisis y/o segmentación de la palabra que introdujo el usuario. A continuación, revisaremos las tres grandes herramientas de análisis y/o segmentación morfológica disponibles para el mapudungun.

Kimalmapuzungun

Kimalmapuzungun es una plataforma en línea[4] desarrollada por Cristian

[3] De hecho, uno de los grandes problemas que identifica el área de PLN (Procesamiento del Lenguaje Natural) es la ambigüedad de las lenguas (Augenstein, 2025). ¿Cómo, por ejemplo, un computador podría entender el significado de: «Juan vio a María con la cámara»?, ¿cómo lo desambigua?, ¿Juan vio a María a través de la cámara o Juan vio a María mientras esta llevaba la cámara? «Lamentablemente, nunca lo sabremos», dijo el sindicato de computadores.

Ahumada como parte de su tesis de magíster en ciencias de la computación. Entre sus distintas funcionalidades, destaca el análisis morfológico del mapudungun, además de herramientas complementarias como la conversión entre grafemarios, la identificación del grafemario utilizado y un contador de palabras. Estas funciones permiten abordar diversos aspectos del procesamiento automático del mapudungun.

Si bien las cuatro pestañas interactivas de *Kimalmapuzungun* pueden ser muy útiles para el aprendizaje del mapudungun, solo se revisará la correspondiente al análisis morfológico y su funcionamiento. El análisis de esta herramienta se realiza a través de la experimentación con la plataforma y la revisión de la tesis de Ahumada (2022), donde otorga detalles respectivos a la infraestructura computacional de este analizador morfológico.

Uno de los primeros aspectos a considerar a la hora de revisar *Kimalmapuzungun* es entender que fue creado con un fin comunicativo, con ello el autor refiere a que, por ejemplo, algunos sufijos son fusionados para facilitar la comprensión de su significado (Ahumada, 2022). En este sentido, no se deben esperar las precisiones lingüísticas que desearía una persona dedicada a esta área, ya que no es el público objetivo de esta herramienta, sino que busca facilitar el uso para personas no expertas en la materia.

Respecto a la construcción del sistema de segmentación y análisis morfológico de *Kimalmapuzungun*, este se puede sintetizar en dos grandes pasos. En primer lugar, Ahumada se encarga de realizar un trabajo de sistematización en expresiones regulares de los diversos morfemas del mapudungun y las respectivas reglas gramaticales asociadas a estas. La información teórica en la que se basa *Kimalmapuzungun* proviene de diversos autores, entre los que destacan Smeets, Catrileo y Cañumil. A partir de esta información, Cristian Ahumada sistematizó las reglas del mapudungun y desarrolló un algoritmo de procesamiento en Python que opera como un

[4] https://kimalmapuzugun.cl

sistema de ramas (Ahumada, 2022). A continuación, se presenta una cita que sintetiza su procedimiento de análisis:

> A partir de estas expresiones regulares se recorre la cadena de una palabra y se genera su árbol. Cada rama se evalúa para ser considerada como válida, y por lo tanto pasar al siguiente paso de "traductor informal". Los morfemas y su orden deben cumplir ciertas restricciones que tienen que ver con la correcta formulación de palabras en mapuzugun, tanto en orden, como se mencionó antes, pero también en la compatibilidad de que dos morfemas estén en la misma palabra. (Ahumada, 2022, p. 33)

En resumen, el analizador morfológico de Cristian Ahumada es una herramienta diseñada específicamente para el procesamiento del mapudungun, que permite segmentar palabras y reconocer los morfemas que las componen. Su funcionamiento se organiza a partir de un sistema de análisis por ramas, el cual consiste en una estructura de decisiones secuenciales que sigue diferentes caminos según las características morfológicas de la palabra analizada. En específico, realiza una identificación del posible morfema y se aplican las reglas gramaticales asociadas a este, de esta manera, va segmentado eligiendo uno u otro camino dependiendo de las reglas morfológicas del mapudungun. Esta lógica fue implementada en Python utilizando expresiones regulares, las cuales permiten identificar patrones específicos de morfemas previamente definidos a partir de un estudio detallado de la gramática del mapudungun.

Düngupeyüm: analizador y generador morfológico para mapudüngun

El analizador morfológico *Düngupeyüm* es una plataforma web[5] desarrollada por Andrés Chandia como parte de su tesis de maestría en Ciencia Cognitiva y Lenguaje. Su función principal es el análisis morfológico del mapudungun, aunque también ofrece herramientas complementarias. Entre sus funcionalidades adicionales destaca la posibilidad de consultar información específica sobre un morfema y acceder a una documentación detallada que

[5] https://www.chandia.net/dungupeyum

describe las etiquetas lingüísticas utilizadas en el proceso de análisis. Esta documentación se basa en diversas fuentes literarias y proporciona información lingüística y gramatical que sustenta el funcionamiento de la herramienta (Chandia, 2024).

La infraestructura computacional que sustenta a *Düngunpeyüm* está casi en su totalidad basada en Python, sin embargo, también utilizan lenguajes de programación como HTML y Java para algunas tareas específicas (Chandia, 2012). Acerca del análisis morfológico, Chandia trabaja con un Transductor de Estado Finito (*Finite State Transducers (FST)*) propuesto por Beesley y Karttunen en su libro *Finite State Morphology* (2003). Los transductores de estado finito se pueden definir, de manera superficial, como una máquina que trabaja con un alfabeto de entrada y uno de salida, donde se establece una correspondencia entre una lengua y otra mediante sus alfabetos. De esta manera, si el usuario le entrega un texto en la lengua A, el FTS otorgará su correspondencia en la lengua B, todo ello basado en una serie de reglas que fueron introducidas anteriormente en la máquina (Chandia, 2012; Parde, 2020). Por tanto, es necesario generar una serie de reglas que le permitan al FTS realizar la tarea propuesta. En el caso de Chandia, creó una serie de expresiones regulares que contenían la información morfológica del mapudungun, basándose para ello en el libro *A Grammar of Mapuche* de Ineke Smeets (2008) y en las entradas léxicas del diccionario del Fray Félix José de Augusta (Chandia, 2012). A continuación, se otorgan más detalles de las reglas utilizadas para el FTS:

> Para generar las reglas morfotácticas que introdujimos al XFST a través de expresiones regulares, generamos un cuadro en el que pudimos detectar las diferentes interrelaciones que tienen los sufijos del paradigma verbal del mapudungun. Se establecieron cuatro niveles de interrelación: prohibición, obligatoriedad, obligatoriedad condicionada y semi-obligatoriedad. (Chandia, 2012, p. 41)

En resumen, el analizador morfológico *Düngupeyüm* permite mapear estructuras lingüísticas de entrada con representaciones de salida, en específico,

es posible realizar una correspondencia de una palabra del mapudungun a una estructura segmentada y analizada gramaticalmente en español. Para su implementación, Chandia desarrolló un conjunto de expresiones regulares que sistematizan las propiedades morfológicas del mapudungun, estableciendo niveles de interrelación entre sufijos verbales como prohibición, obligatoriedad, obligatoriedad condicionada y semi-obligatoriedad (Chandia, 2012). Este trabajo opera a partir de transductores de estado finito, siguiendo el enfoque propuesto por Beesley y Karttunen (2003), y se sustenta en la gramática propuesta por Smeets (2008) y el diccionario de Augusta para la información léxica.

Corpus Histórico del Mapudungun (CHM)

La tercera y última herramienta corresponde al *Corpus Histórico del Mapudungun (CHM),* creado por Benjamín Molineaux. La función principal de esta plataforma web[6] es ser un repositorio de textos en mapudungun, datados entre 1606 y 1930, a través de un formato digital. Este corpus reúne documentos en diversas configuraciones, incluyendo textos transcritos, escaneados y etiquetados. Además de ello, el *CHM* permite buscar morfemas del mapudungun dentro del material recopilado (Molineaux, 2025).

La herramienta de búsqueda por morfemas permite a los usuarios ingresar un morfema o palabra en mapudungun para encontrar coincidencias dentro del corpus. A través de este proceso, la plataforma segmenta y presenta un análisis de las palabras, lo que puede interpretarse como una forma de análisis morfológico. El funcionamiento de este sistema está basado en el *Extensible Markup Language* (XML), un lenguaje de marcación que permite almacenar información (Ray, 2003). En esta plataforma, Molineaux utiliza el XML "desarrollados por el Text Encoding Initiative para material lingüístico" (Molineaux, 2022, p. 26), es decir, realiza una asignación manual de etiquetas a los textos del corpus, permitiendo así que los usuarios realicen búsquedas de morfemas previamente registrados. Por tanto, si el morfema o palabra consultada no ha sido previamente etiquetada en el corpus, la búsqueda no generará resultados, incluso si dicho elemento existe en la lengua mapuche. En

[6] http://www.amc-resources.lel.ed.ac.uk/CHM/

consecuencia, el alcance del análisis morfológico que ofrece esta herramienta depende estrictamente del material previamente incorporado y etiquetado en la base de datos.

En resumen, las tres herramientas computacionales analizadas realizan, al menos, un proceso de segmentación morfológica del mapudungun. Todos estos procesos de segmentación fueron posibles gracias a un proceso humano de aprendizaje de los aspectos gramaticales de los morfemas de la lengua, los cuales de manera posterior se utilizaron para segmentar y analizar. El *CHM* realiza la tarea a través del sistema XML para guardar la información respectiva a cada morfema del mapudungun. De esta manera, cada vez que un usuario busca un morfema el *CHM* revisa en su base de datos sustentada por XML y entrega aquellos resultados que coinciden. Por otro lado, tanto el analizador morfológico *Düngupeyüm* como el *Kimalmapuzungun* realizan sus análisis a partir de un estudio de los aspectos gramaticales de la morfología del mapudungun. Posterior a ello, la información aprendida por sus desarrolladores se sistematizó en expresiones regulares. Son estas expresiones regulares las que se incluyeron en la programación de cada una de las plataformas, usando *Kimalmapuzungun* un algoritmo de análisis por ramas y utilizando *Düngupeyüm* un Transductor de Estado Finito. De esta forma, cada vez que un usuario ingresa una palabra los programas comienzan un proceso de segmentación morfológica basada en los datos que tienen. Así, se entrega como resultado las distintas posibles segmentaciones que tiene la palabra otorgada por el usuario.

¿Qué es lo que nos otorgan estas infraestructuras computacionales?

Las diversas infraestructuras para el análisis morfológico del mapudungun entregan a los usuarios posibilidades de estudio y aprendizaje de la lengua. Se destaca en esto la automatización del análisis morfológico como una de las principales ventajas que ofrecen estas infraestructuras. Las plataformas estudiadas permiten la segmentación automática de palabras en mapudungun, lo que agiliza los procesos de identificación y validación de morfemas.

Esto permite a hablantes y aprendices de la lengua acceder a información morfológica del mapudungun en distintos grados de tecnicidad, dado que las diversas herramientas tienen su usuario objetivo en torno a este aspecto. Sumado a ello, estos analizadores pueden ser una gran herramienta para lingüistas que estén estudiando el mapudungun, ya que facilita el acceso al significado de las diversas partículas gramaticales de la lengua.

Pongamos un ejemplo para ilustrar todo esto. Imagínese que una persona quiere comenzar a aprender mapudungun y decide tomar un curso para cumplir su objetivo. Como tarea le piden que traduzca un breve texto, pero por mucho que lo lea, no puede identificar qué significan un par de palabras. Decide buscar en Google, preguntarle a ChatGPT, e incluso recurre a Tiktok, pero nada de lo que le aparece le hace sentido en el contexto del texto. En vista de esto, se acuerda que leyó un ensayo respecto a infraestructuras computacionales para el análisis morfológico del mapudungun, y ¡click!, su cerebro comienza a andar. Decide utilizar los diversos analizadores morfológicos y su tarea se comenzó a hacer sola, cada traducción se fue haciendo más y más sencilla con la ayuda de estas herramientas. ¿Esto cambió la vitalidad del mapudungun? Es muy probable que no, pero al menos facilitó a una persona su camino en el trayecto a ser hablante de la lengua, y he allí la clave.

Volviendo a nuestra premisa

Entonces, ¿es la computación una aliada en el estudio de lenguas indígenas? Sí. Los ejemplos de herramientas de análisis morfológico para el mapudungun pusieron en evidencia que el avance en infraestructuras computacionales es esencial no solo para el estudio lingüístico, sino también para promover la enseñanza y preservación de la lengua. En la actualidad, tanto *Düngupeyüm*, como *Kimalmapuzungun* y el *Corpus Histórico del Mapudungun* están disponibles para cualquier persona que quiera utilizarlas, ya sea con fines de segmentar morfemas y revisar el análisis de estos, buscar palabras o cualquier otro elemento relacionado. En este sentido, están disponibles para ayudar a suplir necesidades de una persona interesada en el mapudungun y funcionar como una gran colaboración lingüístico-computacional para esta lengua indígena.

Conocer sobre las infraestructuras computacionales y sus posibilidades abre una ventana en la lingüística computacional, y también, al estudio de las lenguas indígenas. Permite visualizar otros caminos posibles y observar cómo estos pueden amoldarse a nuestras necesidades, representando un gran paso en el desarrollo de herramientas útiles para el estudio de las lenguas indígenas.

Referencias bibliográficas

Augenstein, I. (10 March 2025). *Natural Language Processing Fundamentals*. [Conferencia principal]. KHIPU 2025 Latin American Meeting in Artificial Intelligence. Santiago, Chile.

Ahumada, C. (2022). *Diseño y desarrollo de una infraestructura computacional básica para el aprendizaje del mapuzugun* [Tesis para optar al grado de magíster en ciencias, mención computación].

Chandia, A. (2012). *Dungupeyem1_alfa2_v0.1: un prototipo de analizador morfológico para el mapudungun a través de transductores de estados finitos* [Máster interuniversitario de Ciencia Cognitiva y Lenguaje]

Chandia.net. (26 de abril de 2024). *Düngupeyüm: analizador y generador morfológicos para mapudüngun.* Herramientas lingüístico-computacionales para la preservación y difusión del mapudüngun. Recuperado el 5 de agosto del 2024: https://www.chandia.net/dungupeyum

Escandell, M. (2011). Capítulo 4: Morfología. *Invitación a la lingüística*. (pp. 127-151). Editorial Universitaria Ramón Areces.

Golluscio, L. y Hasler, F. (2017). Jerarquías referenciales y alineamiento inverso en mapudungun. *RASAL Lingüística,* 2017, 69-93.

Gundermann, H., Canihuan, J., Clavería, A., & Faúndez, César. (2011). El mapuzugun, una lengua en retroceso. *Atenea* (Concepción), (503), 111-131. https://dx.doi.org/10.4067/S0718-04622011000100006

Molineaux, B. (2021). El Sermón en lengua de Chile en el Corpus Histórico del Mapudungun de texto colonial a texto digital. *Revista de Lenguas y Literatura Indoamericanas*, 23(2), 1-27. http://portal.amelica.org/ameli/journal/608/6083356001/

Molineaux, B. (2023). The Corpus of Historical Mapudungun: Morpho-phonological parsing and the history of a Native American language. *Corpora,* 18(2), 175-191. https://doi.org/10.3366/cor.2023.0281

Parde, N. [Natalie Parde] (5 de septiembre del 2020). Finite State Transducers. [Video]. Youtube. https://www.youtube.com/watch?v=QIFIvFiRi2I

Molineaux, B. (6 de marzo del 2025). *Corpus Histórico del Mapudungun*. http://www.amc-resources.lel.ed.ac.uk/CHM/

OpenAI. (2025). ChatGPT (versión 4o) [Modelo de lenguaje de gran tamaño]. https://chat.openai.com/chat

Ray, E. (2003). Chapter 1. Introduction. *Learning XML*. (pp. 14-61). O'Reilly.