

*Eduardo Llanquiman*

*Universidad de Chile*

*Análisis Computacional de Datos Lingüísticos*

*Prof. Javier Vera*

*Variación métrica de palabras con la sexta vocal de la lengua mapuche /ü/ en Sounds Comparison*

Segunda entrega

**Introducción**

Por su distinción y variabilidad, la sexta vocal de la lengua mapuche *ü* ha sido de gran interés, especialmente, para quienes han descrito con diferentes instrumentos metodológicos tanto su manifestación fonética como su representación fonológica. Trabajos como el de Suárez (1959), con una de las primeras descripciones prosódicas de esta vocal, ha sido base para nuevas investigaciones con análisis espectrales cada vez más sofisticados. En el estudio de Soto-Barba, Lara y Salamanca (2016), por ejemplo, se comparan los formantes vocálicos en sectores del Biobío y Queuco. Un resumen histórico de los diversos estudios descriptivos de *ü* se puede encontrar en Dehnhardt, Valenzuela y Villarroel (2015).

Scott Sadowsky quien además de realizar uno de esos estudios (Sadowsky y otros, 2013), fue codirector en la construcción de la base de datos Sound Comparisons, proyecto lingüístico creado por Paul Heggarty, donde se editaron y registraron 37 variedades del mapudungun. Esta base de datos, de libre distribución, contiene 224 significados agrupados en 14 categorías de palabras. Además, en relación a estos significados, la base de datos posee 7.611 transcripciones fonéticas de 7.372 archivos de audio, identificados geográficamente según la zona donde fue tomada la muestra (Sadowsky, Aninao y Heggarty, 2019).

Este tipo de digitalización otorga datos computacionales que permiten calcular distancias entre lenguas, empleadas, principalmente, en investigaciones lingüísticas diacrónicas como: explorar relaciones filogénicas, calcular la relación del tamaño de una población respecto de la evolución de su lengua, estimar fechas de existencia y extinción de una lengua, etcétera (Greenhill, 2011). Asimismo, es posible también calcular distancias de palabras de forma interna en una misma lengua. Recientemente, se ha utilizado esta técnica para corregir, por ejemplo, errores de cadena de textos (Amón, Moreno y Echeverri, 2012) y de reconocimiento de voz (Campos, Campos, Martínez y Uc, 2019).

En este sentido, y con fines más bien sociolingüísticos, observar la variación de la sexta vocal de la lengua mapuche *ü* a partir de la distancia entre palabras que la contengan, establece un objetivo computacional útil que permite comparar los resultados de las investigaciones elaboradas con otras metodologías lingüísticas y, de esta manera, determinar evidencias contundentes de la distribución alofónica y su influencia de variación.

Una distancia como la propuesta por Levenshtein (1966), permite calcular las distancias entre las variaciones alofónicas de una lengua de forma rápida y precisa. La distancia de Levenshtein es una comparación métrica de un *string* que cuenta el número de operaciones necesarias para transformarlo en otro, las posibilidades son: añadir, eliminar o sustituir un elemento (Greenhill, 2011). Por tanto, en una comparación entre dos pronunciaciones diferentes de la palabra *müngu* (esp. cochayuyo), ['mɘ.ŋʊ] y [mɐ̞.'ŋʊ] por ejemplo, la distancia es de 3, y como ambos *strings* están determinados a un lugar geográfico específico de pronunciación, es posible calcular a gran escala dichas distancias, agruparlas y graficarlas.

El objetivo general de este trabajo es, precisamente, medir la distancia fonética, utilizando la distancia de Levenshtein, entre palabras del mapudungun que contengan las vocal *ü* dentro de la base de datos otorgada por Sounds Comparison. Esto con el fin de identificar geográficamente la variación alofónica de dicha vocal y, a su vez, identificar tendencias de pronunciación según las diferentes localidades de habla.

Entonces, como primer objetivo específico, se pretende medir la distancia de Levenshtein de la palabra *adümchefe* (esp. profesor) entre 15 variaciones fonéticas disponibles en la base de datos y, con esto, determinar si la metodología a utilizar es eficiente en cuanto a su aplicación.

**Metodología**

* Descripción general de la base de datos

Como se mencionaba más arriba, los datos utilizados en este trabajo son parte de la base de datos Sound Comparisons, la cual consta de 224 palabras del mapudungun divididas en 15 categorías semánticas. Para el objetivo general de este trabajo, la muestra se limita a las 83 palabras que contienen la vocal en estudio *ü*: números (1), colores (3), adjetivos (7), cuerpo (7), personas (3), animales (5), naturaleza (10), fenómenos meteorológicos (4), comida (3), hogar (4), verbos (14), palabras gramaticales (2), tiempo (2), cultura (1) y pares mínimos (16).

Estás palabras están transcritas en notación fonética según la localidad en la que se pronuncia. Las 38 localidades posibles se agrupan en 7 macrozonas del sur de Chile y de Argentina: en la zona del Biobío: Santa Bárbara, Cañete, Tirúa y Alto Biobío; en la zona de la Araucanía norte: Angol, Lumaco, Ercilla, Galvarino, Victoria y Lonquimay; en la zona de la Araucanía centro: Chol Chol, Dollinco, Vilcún, Puerto Saavedra, Truf Truf, Icalma, Freire, Cunco; en la zona de la Araucanía sur: Nueva Toltén, Villarrica, Curarrehue; la zona de Los Ríos: Mariquina, Lanco, Panguipulli, Valdivia; en la zona de Los Lagos: San pablo y San Juan de la Costa; y, por último, en la zona de Argentina: Ranquel, Picunches, Zapala, Aluminé, Dpto Huiliches, Huiliches, Lago Rosario, Jacobacci, Cushamen y Futaleufú.

* Pre-procesamiento de los datos

La construcción de los datos aquí trabajados consistió en la unión de todos los archivos *csv* descargados del sitio web Sounds Comparison en un *dataframe* único a través del lenguaje de programación *Python*. Es decir, por cada palabra (83) se descargó una plantilla con la información, de izquierda a derecha en columnas, referente a: la identificación del idioma, la localidad de la pronunciación, la longitud y latitud de su ubicación, la identificación de la palabra, la traducción en español, la traducción en inglés, el nombre del prototipo de palabra (donde no existe información para el grupo seleccionado), su notación fonética, la palabra en mapudungun y, finalmente, los cognados con palabras de la misma familia (donde tampoco existe información respecto del grupo seleccionado).

Luego de reunir todos los archivos *csv* en un directorio común, se utilizaron 3 librerías de Python para manipularlos en el entorno *Spyder*: *Pandas*, *Glob* y *Os*. Primero, con la función *glob*, se creó una lista con la información de todos los archivos en conjunto, para luego recorrer en un ciclo todas las lecturas de esos archivos y nombrarlas, con la función *os.path.basename*, en otra lista. Seguidamente, se utilizó la función *append* para concatenar toda la información conjunta de esta última lista en un *dataframe*. Por último, se utilizó la función *drop* para eliminar columnas no útiles para el objetivo planteado. En definitiva, los datos se organizaron en 1.341 filas y 7 columnas. Una visión general se puede observar en la tabla 1, con un ejemplo de 4 filas:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Índice | Localidad | Latitud | Longitud | Español | AFI | Mapudungun | Categoría |
| 0 | Santa Bárbara | 37.67405 | 71.80186 | ellos | ˌvë.jë.ˈŋɘn | feyengün | Palabra gramatical |
| 1 | Cañete | 37.96722 | 73.39282 | ellos | cʰɪ.ˈðwɪ.ŋɘn | feyengün | Palabra gramatical |
| 2 | Tirúa | 38.3699 | 73.49067 | ellos | ˈɸë.jë.ɣ̞ɐ̝ | feyengün | Palabra gramatical |
| 3 | Alto Bio-Bío | 38.04457 | 71.36344 | ellos | m̩.ˌvë.jë.ˈŋɘn | feyengün | Palabra gramatical |

Tabla 1: Base de datos principal: primeras 4 filas de 1341 totales.

* Técnicas de análisis

Una vez visualizados los datos generales, se tomó una muestra de 1 palabra, *adümchefe*, con 14 variaciones (localidades) fonéticas posibles para calcular, a modo de pilotaje inicial, su distancia. Para esto, se llevaron a cabo dos procesos: en primer lugar, se implementó un código que hiciera el cálculo de la distancia de Levenshtein mediante una función que recibiera dos *strings* a comparar y devolviera la medida de distancia. En segundo lugar, se organizó la información del dataframe en dos ‘diccionarios’ que permitieran, por una parte, contraponer todos los pares de localidades disponibles para la palabra escogida, y por otra, acumular los cálculos de distancia de cada una de las operaciones.

El primer diccionario agrupa la información sobre la variación fonética de la palabra *adümchefe* (en la base de datos ‘personas\_que\_ensenya’) según la localidad en la que se pronuncia (se marca ‘unk’ cuando no existe información para la localidad consultada). Un ejemplo de 2 localidades (de 38 totales) se puede ver en la tabla 2:

|  |  |  |  |
| --- | --- | --- | --- |
| Key | Tipo | Tamaño | Value |
| Alto Bío Bío | Diccionario | 1 | {'persona\_que\_ensenya': 'ˌɐ̝.ðʊm.ˈt͡ʃë.ʋë'} |
| Aluminé | Diccionario | 1 | {'persona\_que\_ensenya': 'unk'} |

Tabla 2: Primer diccionario. En Key se muestran las localidades y en Value la pronunciación de la palabra escogida.

El segundo diccionario contiene la información con la medida de distancia ya calculada. Cada localidad en Key (38) se mide con todas las localidades disponibles (38) en una contraposición de distancia en par, por lo que en Value se genera un diccionario con el valor de Levenshtein para cada localidad comparada. Un ejemplo de la primera fila se puede observar en la tabla 3:

|  |  |  |  |
| --- | --- | --- | --- |
| Key | Tipo | Tamaño | Value |
| Alto Bío Bío | Diccionario | 38 | {‘Futaleufú’: S/I, ‘Villarrica’: S/I, ‘Curarrehue: 6, … (35 más)} |

Tabla 3: Segundo diccionario. En Key se muestra la localidad a comprar con todas las otras localidades y en Value el valor para cada uno de los cálculos. S/I indica sin información.

**Resultados preliminares**

El resultado de las distancias de Levenshtein para la palabra *adümchefe* en 14 de las localidades, se puede observar en la Tabla 4:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A.BB. | Cura. | Frei. | Icalma | L. R. | Lanco | Mariq. | N.Tol. | Pangui. | P.Saa. | S.Bár | T.T | Vald. | Vilcún |
| A. BB. | 0 | 6 | 2 | 7 | 4 | 4 | 6 | 6 | 4 | 5 | 3 | 7 | 5 | 3 |
| Cura. | 6 | 0 | 4 | 8 | 2 | 7 | 5 | 4 | 6 | 3 | 4 | 2 | 6 | 4 |
| Freire | 2 | 4 | 0 | 5 | 2 | 3 | 5 | 4 | 3 | 3 | 4 | 6 | 3 | 4 |
| Icalma | 7 | 8 | 5 | 0 | 7 | 7 | 8 | 8 | 8 | 8 | 9 | 9 | 4 | 9 |
| L.R. | 4 | 2 | 7 | 7 | 0 | 5 | 3 | 3 | 4 | 1 | 2 | 4 | 5 | 2 |
| Lanco | 4 | 7 | 3 | 7 | 5 | 0 | 3 | 5 | 4 | 4 | 6 | 6 | 4 | 6 |
| Mariq. | 6 | 5 | 5 | 8 | 3 | 3 | 0 | 4 | 5 | 2 | 4 | 4 | 6 | 4 |
| N. Tol. | 6 | 4 | 4 | 8 | 3 | 5 | 4 | 0 | 6 | 2 | 5 | 4 | 5 | 5 |
| Pangui. | 4 | 6 | 3 | 8 | 4 | 4 | 5 | 6 | 0 | 5 | 5 | 7 | 6 | 5 |
| P. Saa. | 5 | 3 | 3 | 8 | 1 | 4 | 2 | 2 | 5 | 0 | 3 | 3 | 4 | 3 |
| S. Bárb. | 3 | 4 | 4 | 9 | 2 | 6 | 4 | 5 | 5 | 3 | 0 | 5 | 7 | 0 |
| T. T. | 7 | 2 | 6 | 9 | 4 | 6 | 4 | 4 | 7 | 3 | 5 | 0 | 6 | 5 |
| Valdivia | 5 | 6 | 3 | 4 | 5 | 4 | 6 | 5 | 6 | 4 | 7 | 6 | 0 | 7 |
| Vilcún | 3 | 4 | 4 | 9 | 2 | 6 | 4 | 5 | 5 | 3 | 0 | 5 | 7 | 0 |

Tabla 4: Resultado para la distancia de Levenshtein de la palabra adümchefe en 14 localidades.

La primera impresión con este ejercicios es de arbitrariedad: se tendería a pensar que mientras más cercanas las localidades menos distancia tendrían las pronunciaciones, lo que no es así. Vilcún está a 108 km de Icalma y la distancia es alta, de 9. En cambio, Lago Rosario está a 836 km de esta última y la distancia es solamente de 2.

**Conclusiones preliminares**

Con este primer ejercicio se pueden hacer las siguientes observaciones:

1. El código implementado para realizar los cálculos es eficiente, no obstante, crea datos no útiles que entorpecen el análisis de los resultados. Esto debido a que no siempre hay información fonética de las palabras escogidas, lo que genera en algunas ocasiones, un cálculo entre el tamaño de un *string* y cero, dando como resultado el tamaño completo de dicho *string* y no la distancia entre dos distintos. De esto se desprende que se debe eliminar, de cada palabra seleccionada, las localidades que no contengan la información solicitada, mediante un código que las identifique.
2. Si se quiere lograr una comparación de todas las palabras seleccionadas en 1.444 contraposiciones de localidades, es necesario crear otro diccionario que permita ordenar los resultados de las distancias de forma operacional y de esta manera poder graficar parte de los datos que vayan siendo de interés en los análisis de las distancias.
3. Por último, es importante determinar la forma concreta de la distancia, es decir, todos los alófonos de /ü/, de esta manera no solo se mediría la distancia entre los *strings* sino que también se podría observar la influencia de la vocal en esa distancia.

**Referencias bibliográficas**

AMÓN, I. MORENO, F. Y ECHEVERRI, J. (2012). Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español. *Revistas Ingenierías Universidad de Medellín*, 11(20), 127-138.

CAMPOS, D. CAMPOS, M. MARTÍNEZ, I. Y UC, V. (2019). Corrección de errores del reconocedor de voz de Google usando métrica de distancia fonética. *Physical Engineering in Biomedicine*, 57.

DEHNHARDT, M. VALENZUELA, A. Y VILLARROEL, N. (2015). *Adaptación de la sexta vocal /ü/ del mapudungun al español. Una revisión del diccionario etimolójico de Rodolfo Lenz.* Tesis de Licenciatura en Lengua y Literatura Hispánica mención Lingüística. Universidad de Chile.

GREENHILL, S. (2011). Levenshtein distances fail to identify language relationships accurately. *Computacional linguistic* 37(4), 689-698.

LEVENSHTEIN, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10, 707-710.

SADOWSKY, S. Y OTROS (2013). “Mapudungun”. *Journal of the International Phonetic Association* 43(1), 87-96.

SADOWSKY, S. ANINAO, M. Y HEGGARTY, P. (2019). Sound Comparisons: Mapudungun. (disponible online en https://soundcomparisons.com/Mapudungun, consultado el 2021-10-31.)

SOTO-BARBA, J. LARA, I. Y SALAMANCA, G. (2016). Descripción fonético-acústica de la sexta vocal en el chedungun hablado en Alto Bio-Bío. *Onomázein* 34, 229-241.

SUÁREZ, J. (1959). The phoneme of an Araucanian dialect. *IJAL* 25, 77-181.

SADOWSKY, S. ANINAO, M. Y HEGGARTY, P. (2019). Sound Comparisons: Mapudungun. (disponible online en https://soundcomparisons.com/Mapudungun, consultado el 2021-10-31.)