



Aproximación computacional a la relación entre distancia fonética y geográfica de palabras con la sexta vocal del mapudungun


 Eduardo Llanquiman y Javier Vera


14 de diciembre de 2021

1. Introducción

Desde tempranas descripciones fonéticas, la sexta vocal *ũ* del sistema fonológico de la lengua mapuche /i/ (Barros, 2005) ha presentado una notoria variabilidad entre las zonas en donde se pronuncia, sin embargo, todavía no existe consenso para una descripción articulatoria ‘estándar’, ni para el contexto de aparición de sus principales alófonos: [ə] y [u] (Mena Sanhueza et al., 2020). En estas investigaciones se han propuesto diferentes criterios para categorizar esta variación, por ejemplo: la posición de la vocal (Salas, 1976), (Lagos Altamirano, 1981) y (Salamanca, 1997), su contexto fonético (Soto-Barba et al., 2016) o acentual (Smeets, 2008).


En la base de datos *Sounds Comparison*,  hay 83 palabras del mapudungun que contienen esta vocal, transcritas cada una de ellas en notación fonética según diferentes localidades de habla. Con fines sociolingüísticos, observando la variación fonética de ciertas conceptualizaciones fonológicas relacionadas con /i/, a partir de la digitalización de estos datos, establece un objetivo computacional útil que permite medir las pronunciaciones en diferentes sectores y, con esto, relacionar las distancias fonéticas y geográficas, agrupándolas en términos de distinción alofónica.


 El objetivo general de este trabajo es, precisamente, identificar las diferencias en las transcripciones fonéticas de palabras del mapuche que contengan *ũ* y, a su vez, identificar tendencias de pronunciación según las diferentes localidades de habla. Mientras que, en específico, se pretende medir la distancia de Levenshtein, entre cuatro palabras del mapudungun que contengan *ũ* en diferentes contextos, atendiendo a los criterios de cantidad de sílabas, de vocales en la palabra y de posición.

 Espera, por una parte, que la relación fonética y geográfica sea directamente proporcional, es decir, mientras más lejana la localidad desde el punto de referencia (transcripción fonética ‘estándar’) más distancia existirá entre las pronunciaciones de una misma palabra. Y por otra, que se presenten correlaciones entre las diferencias fonéticas de las palabras seleccionadas.

2. Metodología

2.1. Descripción de los datos

 *Sounds Comparison* es un 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 15 categorías semánticas. Además, en relación a estos significados, la base de datos posee 7.372 archivos de audio, identificados geográficamente, de los cuales se transcribieron 7.611 notaciones fonéticas (Aninao et al., 2019), a partir del Alfabeto Fonético Internacional (ej. [mõʎ.ˈvəp] y [mõʎ.ˈfʷəp] para /mollfũn/ ‘sangre’) (Rei, 1996).

Para el objetivo general de este trabajo se seleccionaron, en primer lugar, las palabras que contienen la vocal en cuestión (83). En segundo lugar, las palabras con significante único (40), puesto que para algunas equivalencias castellanas existen dos posibilidades (ej. ‘barriga’ es *pũtra* y *angka*); a partir de las 40 palabras resultantes se elaboró  *dataframe* principal. En tercer lugar, se seleccionaron solo aquellas palabras que tienen asociadas el total de localidades, es decir, palabras con 38 diferentes transcripciones fonéticas, una por cada localidad disponible en la base de datos analizada más la

referencia. Estas localidades se dividen en 7 macrozonas del sur de Chile y de Argentina, las cuales se especifican en el cuadro 1.

Chile: Biobío	Chile: Araucanía Norte	Chile: Araucanía Centro	Chile: Araucanía Sur	Chile: Los Ríos	Chile: Los Lagos	Argentina
Santa Bárbara	Angol	Chol-Chol	Nueva Toltén	Mariquina	San Pablo	Chalileo
Cañete	Lumaco	Dollinco	Villarrica	Lanco	Juan de la Costa	Picunches
Tirúa	Ercilla	Vilcún	Curarrehue	Panguipulli		Zapala
Alto Biobío	Galvarino	Puerto Saavedra		Valdivia		Aluminé
	Victoria	Truf Truf				J. de los Andes
	Lonquimay	Icalma				Huiliches
		Freire				Lago Rosario
		Cunco				Jacobacci
						Cushamen
						Futaleufú

Cuadro 1: **37 Localidades divididas en 7 macrozonas de Chile y Argentina.** Se verá a continuación, cada una de estas localidades tiene asociada una ubicación a partir de su latitud y longitud, lo que permite calcular la distancia geográfica exacta entre una muestra y otra.

14 palabras cumplen con los tres requisitos mínimos para un cálculo homogéneo de distancias en cada una de las localidades disponibles. Estas palabras se presentan en el cuadro 2 organizadas según la cantidad de sílabas y la posición de *ü* en la palabra.

último, atendiendo al objetivo específico, se observó la variación de distancias en la comparación de 4 palabras, una por cada categoría fonética mencionada en el cuadro 2: *küla* ‘tres’, *kelü* ‘rojo’, *rüpü* ‘camino’ y *wün* ‘boca’.

Bisílabas 1	Bisílabas 2	Bisílabas 3	Monosílaba
kelü	kelü	ngürü	wün
küten	kurü	rüpü	
müta	kewün		
kütral	kuwü		
	mollfüñ		
	antü		
	pulkü		

Cuadro 2: **palabras asociadas a todas las localidades, divididas por cantidad sílabas y posición de *ü*.** Bisílabas 1 corresponde a palabras de dos sílabas con *ü* en la primera de ella, bisílabas 2 con *ü* en la última, bisílabas 3 con *ü* en ambas y monosílaba corresponde a un solo núcleo *ü*.

2.2. Procesamiento de los datos

La construcción de los datos consistió en la unión de todos los archivos *csv* descargados de *Sound Comparison* en un *dataframe* único, a través del lenguaje de programación *Python*. Por cada palabra se descargó una plantilla con la información, de izquierda a derecha en columnas, correspondiente a: la localidad de la pronunciación, la longitud y latitud de su ubicación, la traducción en español, su notación fonética, la palabra en mapudungun y la categoría semántica. En definitiva, los datos se organizaron en 1.341 filas y 7 columnas (considerando 40 palabras), de las cuales solo se analizaron 152 (38 filas por cada una de las cuatro palabras seleccionadas). Una visión general se puede observar en el cuadro 3, con un ejemplo de 4 filas.

Para manipular los datos se utilizaron los entornos *Spyder* y *Jupyter Notebook*, en los cuales se instalaron las siguientes librerías: *Pandas* (Team, 2020), para la elaboración del *dataframe* y los cálculos de distancia geográfica, *Jellyfish* para medir, en específico, la distancias fonéticas (Turk and Stephens, 2015) y *Scipy* para el *clustering* y la elaboración de los mapas de calor (Virtanen et al., 2020).

2.3. Distancia de Levenshtein

La distancia propuesta por Levenshtein et al. (1966), permite calcular las diferencias entre las variaciones alofónicas de una lengua. Esta distancia es una comparación métrica de un *string* que

Índice	Localidad	Latitud	Longitud	Español	AFI	Mapudungun	Categoría
0	Santa Bárbara	-37.67405	-71.80186	rojo	'kʰë.lə	kelü	Colores
1	Cañete	-37.96722	-73.39282	rojo	'cʰë.lə	kelü	Colores
2	Tirúa	-38.3699	-73.49067	rojo	'cʰë.lë	kelü	Colores
3	Alto Biobío	-38.04457	-71.36344	rojo	kʰë.'lə	kelü	Colores

Cuadro 3: **Base de datos principal.** primeras 4 filas de 152 analizadas.

cuenta el número de operaciones necesarias para transformarlo en otro, las posibilidades son: añadir, eliminar o sustituir un elemento (Greenhill, 2011). Por consiguiente, en una comparación entre dos pronunciaciones diferentes de la palabra *wün* ‘boca’, [wən] y [gwən], por ejemplo, la distancia es 1. Como ya mencionó, estos *strings* están determinados a un lugar geográfico específico de pronunciación según latitud y longitud, por tanto, es posible calcular a gran escala dichas distancias, relacionarlas y graficarlas.

Para esto, se llevaron a cabo dos procesos: por un lado, se implementó un código con el cálculo de Levenshtein mediante una función que recibiera dos *strings* a comparar y devolviera la medida de distancia. Por otro, se elaboró un código que permitiera realizar esta operación para cada una de las localidades reunidas en pares, de esta manera se asignó un valor para cada uno de los cruces entre la totalidad de las localidades, es decir, 1.444 valores por cada una de las cuatro palabras seleccionadas.

La misma implementación anterior se aplicó para los valores de latitud y longitud: se generaron nuevos datos relacionados con los kilómetros de distancia entre las localidades comparadas y la diferencia de pronunciación descrita arriba. Con estos datos se graficó dicha relación mediante un gráfico de puntos y se analizaron las diferencias, en cuanto al número de sílabas (1 ó 2), la posición de la vocal en la sílaba (inicial o final) y la cantidad de vocales *ü* en la palabra (1 ó 2).

3. Resultados

3.1. Relaciones entre distancias fonéticas y geográficas

La figura 1 muestra la comparación de distancias, en gráficos de puntos, de las cuatro palabras analizadas: *küla*, *kelü*, *rüpü* y *wün*.

Desde la pronunciación de referencia, la palabra *küla* ‘tres’ muestra variación entre las macrozonas de la Araucanía, aunque existe una pequeña relación para las localidades de Argentina, es decir, mientras más lejana la zona (Huilliches, Aluminé, Picunches), más distante son las pronunciaciones. En específico, el segmento vocálico [ə] es articulado más cercano a [u].

En la palabra *kelü* ‘rojo’ se aprecia una relación levemente más notoria en la Araucanía centro-sur en comparación a *küla*, sin embargo, la principal distinción se observa nuevamente en las localidades argentinas donde aparecen puntos [u]. Cabe destacar que hay algunas pocas variaciones con [ɪ] en Mariquina, San Pablo, Valdivia y Ercilla.

En la palabra *rüpü* ‘camino’, se presenta una variación similar a la de *kelü*, pero con grupos de pronunciación marcados: la variación más frecuente se da en la macrozona de la Araucanía norte; más cercana a la referencia se encuentra la macrozona del Biobío y la más lejana en la Araucanía centro. En cuanto a la realización de la vocal *ü*, en casi todas las localidades, se da como [ə] en la primera sílaba, exceptuando únicamente la realización [u] de Cushamén. En cuanto a la última sílaba, se presenta mayor variación no relacionada geográficamente.

Por último, al considerar una palabra monosilábica como *wün*, el resultado es evidentemente más claro: no existe variación. Se ven algunos puntos en específico como Victoria y Puerto Saavedra que varían respecto de la media, pero sin relación geográfica ni vocálica significativa. Lo mismo ocurre en San Pablo, Aluminé y Chalileo. El alófono [ə] se da en casi todas las localidades.

3.2. Relación de distancias mediante mapas de calor

En el afán de observar más gráfica y específicamente los resultados comentados en el apartado anterior, se presentan los datos en mapas de calor a partir de un análisis por *clustering*.

El *clustering* es una herramienta de agrupamiento que asigna, a través de diferentes algoritmos, valores a cada uno de los datos analizados, con el fin de establecer puntos céntricos que asocian valores y forman grupos de datos según criterios determinados (Müller and Guido, 2016). En particular,

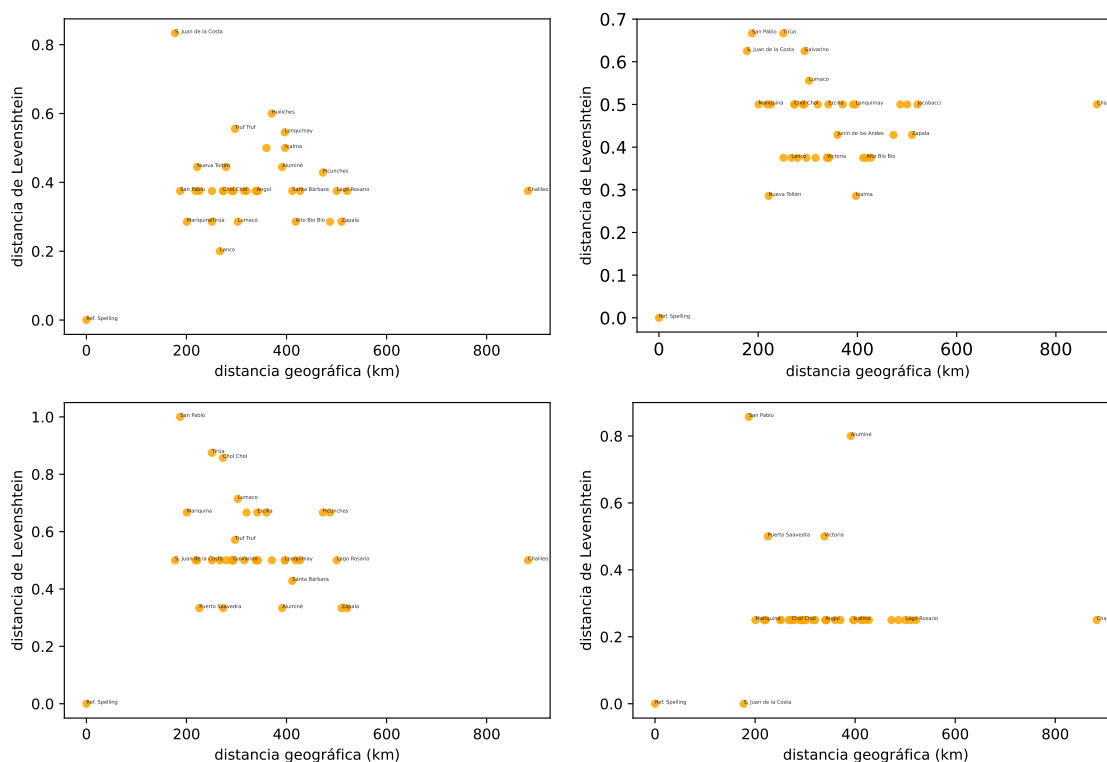


Figura 1: **Relación entre la distancia fonética y geográfica de palabras con \tilde{u} en cuatro diferentes contextos.** Arriba a la derecha la relación para *kũla* 'tres' (bisílaba 1), arriba a la izquierda para *kelũ* 'rojo' (bisílaba 2), abajo a la derecha para *rũpũ* 'camino' (bisílaba 3) y abajo a la izquierda para *wũn* 'boca' (monosílaba).

el *agglomerative clustering* es la suma de dichos algoritmos que, según los criterios de agrupación, jerarquizan distintos clústeres o 'grupos de grupos' asociados a la variabilidad de los datos.

mapas de calor y los dendrogramas permiten graficar la jerarquía de estos agrupamientos, en donde se evidencian relaciones de datos, únicamente, bidimensionales (en este caso fonéticos y geográficos). En la figura 2 se presentan dos mapas de calor por *clustering*, a partir de los dendrogramas al margen, en los cuales se muestran dos grupos de pronunciación.

reafirma que no existe una correlación directa entre la poca variación alfónica de *kũla* y su distribución geográfica. Además, este resultado va en dirección opuesta a lo planteado por Salas (1976) donde se afirma que el contexto inicial de sílaba favorece [u], mientras que en este análisis se observó un evidente uso de [ə] en tal contexto. Es importante mencionar que la vocal varía solo en algunas localidades argentinas, las cuales se marcan en rojo al costado del gráfico.

En *kelũ* la variación es mayor, aunque tampoco existe una correlación evidentemente directa con su distribución geográfica. En este contexto, la variante [ə], muy frecuente en *kũla*, alterna con [u], lo que podría insinuar que la posición de la vocal en la palabra motiva su variación, en particular, cuando \tilde{u} se encuentra en la última sílaba de la palabra.

Cuando se observa el primer gráficos de la figura 2, se pueden corroborar dos afirmaciones: por una parte, no hay una directa relación entre alófonos y localidades, excepto para la macrozona de Argentina y, por otra, la sexta vocal muestra alternancia cuando se encuentra en la última sílaba y presenta resistencia a un alófono en particular cuando está en la primera.

Por último, en el gráfico de *wũn*, es evidente la correlación entre la cantidad de sílabas y la nula variabilidad de la palabra. Además, la correlación entre variación fonética y geográfica se expresa con claridad, entre las localidades de Chile y Argentina.



4. Conclusiones

La posición de la vocal en la palabra, tal como lo habían mencionado otros autores, produce variabilidad alofónica: si la vocal se encuentra en la primera sílaba de la palabra esta no varía notoriamente,

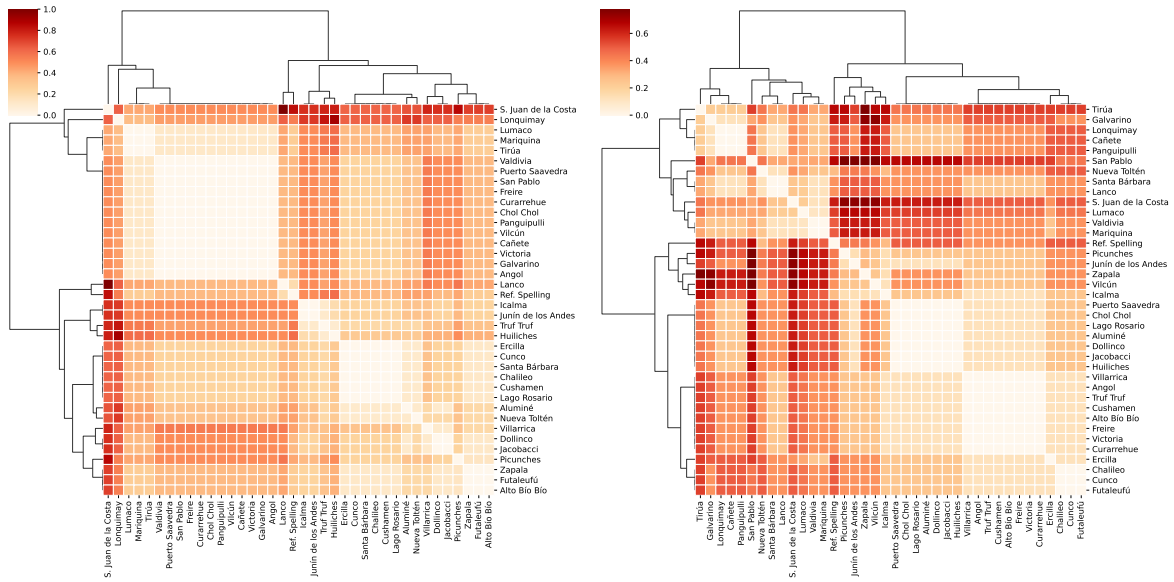


Figura 2: Mapa de calor de las palabras *küla* 'tres' a la izquierda y *kelü* 'rojo' a la derecha. La relación fonética y geográfica no es evidente en ninguno de ambos casos, pero en *kelü* se evidencia mayor variación en las macrozonas centro-sur.

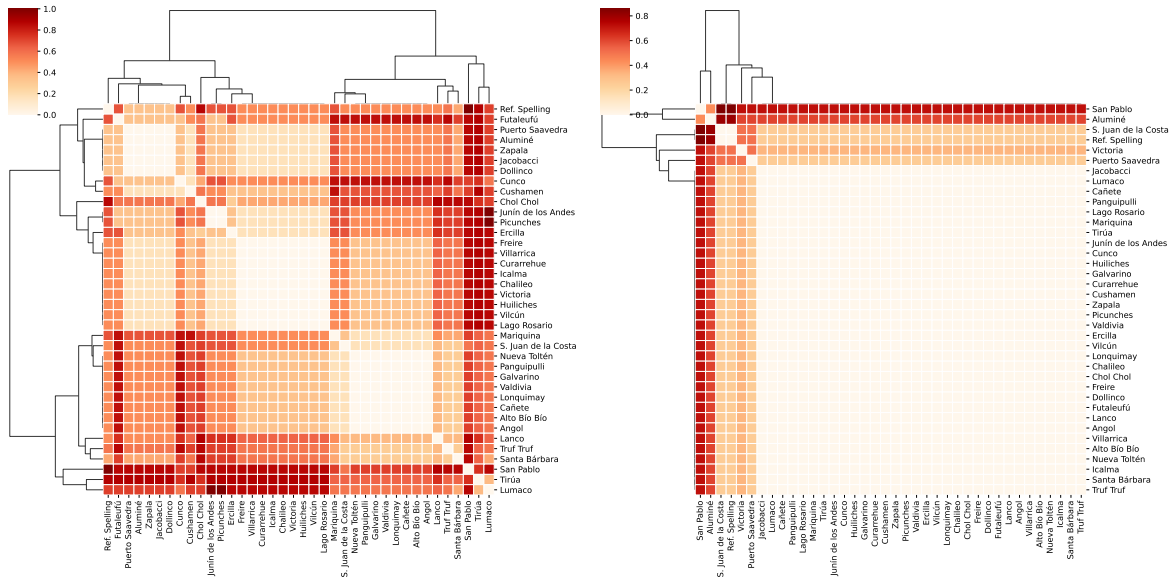



Figura 3: Mapa de calor de las palabras *rüpü* 'camino' a la izquierda y *wün* 'boca' a la derecha. En el primer caso existe variabilidad sujeta a la distancia geográfica, principalmente en Argentina, mientras que en el segundo caso la correlación es evidente para todas las localidades.

en cambio,  hace cuando la vocal se encuentra en la última.

La cantidad de sílabas de la palabra también podría indicar contexto de variación: mientras que para las 3 palabras con dos sílabas la variación es más compleja de interpretar, para la monosilábica es bastante más clara la correlación fonética y geográfica.

La cantidad de vocales (una o dos) dentro de la palabra, y en concordancia con el comentario anterior, puede indicar que el contexto vocálico influye en la relación alofónica de la sexta vocal, no obstante, es importante recalcar que la palabra *rüpü* tiene dos vocales idénticas (no hay contacto con otras vocales), por ende, es imposible de momento afirmar que este contexto motive la variación.

La distancia fonética no está directamente relacionada con las macrozonas ubicadas en Chile, no obstante, se presenta una distinción para las localidades de Argentina, las cuales varían más respecto del punto de referencia. En este último caso la localidad de Challeo es una excepción.

En futuros trabajos complementarios es preciso considerar el contexto fonético consonántico y vocálico de las palabras estudiadas, puesto que también pueden ser motivo de variación vocálica. Además, estudiar otras vocales del mapudungun en los mismos contextos aquí considerados puede dar luces de comportamientos fonológicos más amplios del mapudungun.

Referencias

- Aninao, M. J., Sadowsky, S., and Heggarty, P. (2019). Sound Comparisons: Mapudungun.
- Barros, J. P. V. (2005). *Voces en el viento: raíces lingüísticas de la Patagonia: lingüística comparativa de las lenguas aborígenes del sur del continente americano*. Mondragón.
- Greenhill, S. J. (2011). Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.
- Lagos Altamirano, D. (1981). El estrato fónico del mapudungu(n). *Nueva Revista del Pacífico*, 19(20):42–66.
- Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Union*, 10(8):707–710.
- Mena Sanhueza, D. A. et al. (2020). *Resolución de aspectos controversiales de la fonética y fonología del mapudungun mediante métodos de fonética acústica y estadística inferencial*. Universidad de Concepción, Facultad de Humanidades y Arte, Programa de Doctorado en Lingüística.
- Müller, A. C. and Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. O’Reilly Media, Inc.
- Rei, F. (1996). Tipa: A system for processing phonetic symbols in latex. *TUGBoat*.
- Salamanca, G. (1997). Fonología del pehuenche hablado en el alto bío bío. *RLA. Revista de lingüística teórica y aplicada*, 35:113–124.
- Salas, A. (1976). Esbozo fonológico del mapudungun, lengua de los mapuches o araucanos de chile central. *Estudios filológicos*, 2(11):143–154.
- Smeets, I. (2008). *A grammar of Mapuche*. De Gruyter Mouton.
- Soto-Barba, J., Lara, I., and Gutiérrez, G. F. S. (2016). Descripción fonético-acústica de la sexta vocal en el chedungun hablado en alto bío-bío. *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, 34(14):229–241.
- Team, T. (2020). Pandas development pandas-dev/pandas: Pandas. *Zenodo*, 21:1–9.
- Turk, J. and Stephens, M. (2015). Jellyfish: approximate and phonetic matching of strings.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.