Coloma, Germán

Working Paper

Correlación entre medidas lingüísticas: Algunas extensiones

Serie Documentos de Trabajo, No. 845

Provided in Cooperation with:
University of CEMA, Buenos Aires

Suggested Citation: Coloma, Germán (2023) : Correlación entre medidas lingüísticas: Algunas extensiones, Serie Documentos de Trabajo, No. 845, Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires

This Version is available at:
https://hdl.handle.net/10419/297774

UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

CORRELACIÓN ENTRE MEDIDAS LINGÜÍSTICAS: ALGUNAS EXTENSIONES

Germán Coloma

Enero 2023
Nro. 845

https://ucema.edu.ar/publicaciones/doc_trabajo.php
UCEMA: Av. Córdoba 374, C1054AAP Buenos Aires, Argentina
ISSN 1668-4575 (impreso), ISSN 1668-4583 (en línea)
Editor: Jorge M. Streb; asistente editorial: Valeria Dowding <ved@ucema.edu.ar>

Correlación entre medidas lingüísticas: algunas extensiones

Germán Coloma*

Resumen

En este documento se vuelven a evaluar algunos resultados que aparecieron originalmente en Coloma (2017) sobre la posible existencia de correlación negativa entre medidas lingüísticas, usando una nueva base de datos de 80 idiomas para los cuales contamos con el mismo texto (que es la fábula de "El viento norte y el sol"). La mayor parte de las conclusiones del trabajo original se ven reforzadas en este nuevo análisis, en especial las que tienen que ver con la existencia de efectos de compensación entre distintos aspectos de la complejidad de los idiomas. Esto es particularmente así cuando se estudian los coeficientes de correlación parcial entre tres cocientes lingüísticos (fonemas por sílaba, sílabas por palabra, y palabras por enunciado), y cuando se emplean métodos de regresión de ecuaciones simultáneas y variables instrumentales.

Palabras clave: efectos de compensación, correlación parcial, cocientes lingüísticos, ecuaciones simultáneas, variables instrumentales.

Correlation between Linguistic Measures: Some Extensions

Germán Coloma

Abstract

In this paper we revisit the results that appeared originally in Coloma (2017) about the possible existence of negative correlation between linguistic measures, using a new database of 80 languages for which we have the same text (which is the fable known as "The North Wind and the Sun"). Most conclusions of the original paper become reinforced, especially the ones related to the existence of language complexity trade-offs. This is particularly clear when we look at partial correlation coefficients between three linguistic ratios (phonemes per syllable, syllables per word, and words per clause), especially when we use simultaneous-equation regression methods and instrumental variables.

Keywords: complexity trade-off, partial correlation, linguistic ratios, simultaneous-equation regression, instrumental variables.

* Universidad del CEMA; Av. Córdoba 374, Buenos Aires, C1054AAP, Argentina. Correo electrónico: gcoloma@cema.edu.ar. El presente documento es básicamente una traducción de Coloma (2022). Agradezco los comentarios de Christian Bentz, Jan Macutek y Alison Tompkins respecto de dicho trabajo. Le agradezco también a Oraimar Socorro, por su ayuda para conseguir muchos de los artículos que fueron usados como fuentes de datos para este documento. Las opiniones expresadas en esta publicación son propias, y no representan necesariamente las de la Universidad del CEMA.

1. Introducción

En Coloma (2017) hay un análisis sobre la posible existencia de correlación negativa entre medidas lingüísticas, usando el texto de una fábula relativamente famosa ("El viento norte y el sol") traducida a 50 idiomas distintos.[1] Con esas traducciones, pudimos construir una base de datos con información acerca de distintas medidas empíricas del texto bajo estudio (fonemas por sílaba, sílabas por palabra, palabras por enunciado). Dicha base de datos incluía también otras variables relacionadas con las características tipológicas de los idiomas, y con algunos factores "no lingüísticos" (p.ej., ubicación geográfica, relaciones filogenéticas, número de hablantes).

La principal conclusión del trabajo mencionado es que las correlaciones negativas existen y son significativas en el contexto bajo estudio. También parecen estar parcialmente ocultas, debido a posibles interacciones entre distintas variables. Como consecuencia de ello, se da que los coeficientes de correlación se incrementan y se vuelven más significativos si se toman en cuenta dichas interacciones. Para ello se utilizaron distintas estrategias alternativas, que implicaron el empleo de coeficientes de correlación parcial, regresiones con ecuaciones simultáneas, variables tipológicas y variables no lingüísticas. Una limitación del análisis, sin embargo, tuvo que ver con la base de datos en sí, que tenía solamente 50 observaciones. Dicha limitación se debió a que, cuando se llevaron a cabo los cálculos, había relativamente pocos ejemplos del texto que se usó para comparar, y muchos de esos ejemplos se referían a lenguas que eran relativamente parecidas en términos de su localización y de su parentesco idiomático.

Como ahora ya han transcurrido varios años, hemos podido construir una base de datos alternativa con 80 idiomas, para los cuales tenemos el texto de "El viento norte y el sol". La fuente utilizada es esencialmente la misma que se usó para la muestra original, es decir, es la colección de "ilustraciones del alfabeto fonético internacional" publicada en IPA (1999) y en el *Journal of the International Phonetic Association*.[2] Como dicha colección es ahora considerablemente más extensa que antes, la nueva base de datos tiene la ventaja de que es más diversa, en el sentido de que contiene idiomas de más familias lingüísticas (y no tantas lenguas de la familia indoeuropea).

En este trabajo, utilizamos nuestra nueva base de datos para llevar a cabo esencialmente los mismos análisis que aparecen en Coloma (2017). Primero, describimos las características básicas de la muestra en términos de su alcance y del valor de sus medidas lingüísticas (sección 2). Luego,

[1] Véase también Coloma (2020).
[2] También incluimos un ejemplo tomado de una fuente diferente (Lichtman y otros, 2010).

en la sección 3, usamos esas medidas para computar coeficientes de correlación, empleando distintas metodologías alternativas. En la sección 4 incluimos algunas variables adicionales en el análisis, a fin de incorporar factores geográficos, filogenéticos y demográficos, en tanto que en la sección 5 introducimos un procedimiento que reemplaza nuestras medidas lingüísticas por "variables instrumentales". Finalmente, en la sección 6, comparamos nuestros nuevos resultados con los originales, y establecemos una serie de conclusiones.

2. El viento norte y el sol

La fábula del viento norte y el sol, atribuida a Esopo, es un texto que ha sido utilizado durante muchas décadas por la Asociación Fonética Internacional como un "espécimen" o modelo para ilustrar los sonidos de los distintos idiomas, y también los símbolos fonéticos que se usan para representar dichos sonidos. Es por lo tanto un caso único para el cual numerosos especialistas en la fonética de muchas lenguas distintas han identificado los sonidos, los fonemas, las sílabas y las palabras de dichas lenguas.[3]

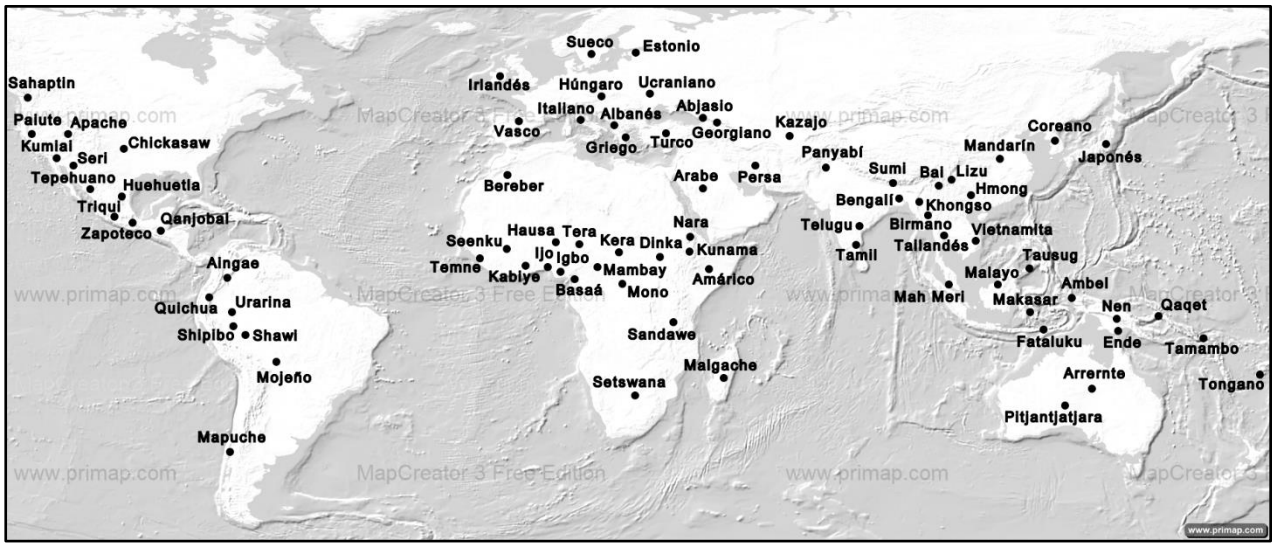
En Coloma (2017) aparece descripta una base de datos basada en "El viento norte y el sol" traducido a los siguientes idiomas: sahaptin, apache oriental, chickasaw, seri, triqui, zapoteco del istmo, quichua ecuatoriano, shiwilu, yine y mapuche (originarios del continente americano); portugués, español, vasco, francés, irlandés, inglés, alemán, ruso, húngaro y griego (de Europa); bereber marroquí, temne, kabiye, igbo, hausa, dinka, nara, amárico, sandawe y bemba (de África); georgiano, turco, hebreo, árabe, persa, tayiko, nepalí, hindi, bengalí y tamil (de Asia occidental); y japonés, coreano, mandarín, cantonés, birmano, tailandés, vietnamita, malayo, tausug y arrernte (de Asia oriental y Oceanía).

Para este trabajo, hemos construido una nueva muestra con 80 idiomas (ver figura 1). Treinta y seis de ellos ya aparecían en la base de datos original: amárico, apache, árabe, arrernte, bengalí, bereber, birmano, chickasaw, coreano, dinka, georgiano, griego, hausa, húngaro, igbo, irlandés, japonés, kabiye, malayo, mandarín, mapuche, nara, persa, quichua, sahaptin, sandawe, seri, tailandés, tamil, tausug, temne, triqui, turco, vasco, vietnamita y zapoteco. Los otros cuarenta y cuatro idiomas son los siguientes: abjasio, aingae, albanés, ambel, bai, basaá, ende, estonio, falaluku, hmong, huehuetla tepehua, italiano, kalabari ijo, kazajo, kera, khongso, kumiai, kunama,

[3] En el apéndice 1 del presente trabajo hemos incluido algunos ejemplos de este texto en distintos idiomas.

lizu, mah meri, makasar, malgache, mambay, mojeño, mono, nen, paiute, panyabí, pitjantjatjara, qanjobal, qaqet, seenku, setswana, shawi, shipibo, sueco, sumi, tamambo, telugu, tepehuano, tera, tongano, ucraniano y urarina.

Figura 1: Ubicación geográfica de los idiomas incluidos en la muestra



Una diferencia obvia entre la base de datos original y la nueva es que esta última tiene treinta idiomas más. Además, su composición es claramente más diversa, ya que incluye ejemplos de 40 familias lingüísticas distintas. La muestra original, en cambio, tenía idiomas pertenecientes a solo 26 familias, y 13 de esos idiomas (26%) eran indoeuropeos. Más aún, había algunas lenguas que eran muy similares entre ellas (como es el caso del español y el portugués, el persa y el tayiko, el mandarín y el cantonés, etc.). Contrariamente, en nuestra nueva base de datos todos los idiomas pertenecen a diferentes "subfamilias".[4]

Los datos básicos que pueden computarse utilizando esta muestra de idiomas provienen de contar el número de fonemas, sílabas, palabras y enunciados incluidos en la traducción de "El viento norte y el sol" para cada uno de dichos idiomas. Con esas cifras, pueden calcularse una serie de cocientes lingüísticos, que básicamente son el cociente entre fonemas y sílabas (*Fon/Sil*), el cociente entre sílabas y palabras (*Sil/Pal*), y el cociente entre palabras y enunciados (*Pal/Enunc*). Estos cocientes pueden verse como medidas de diferentes aspectos de la complejidad de los

[4] Por ejemplo, las nueve lenguas indoeuropeas incluidas en la nueva muestra son: albanés (albánico), bengalí (índico oriental), griego (helénico), irlandés (celta), italiano (latino), panyabí (índico occidental), persa (iranio), sueco (germánico) y ucraniano (eslavo).

idiomas.[5] Por ejemplo, el cociente entre fonemas y sílabas se relaciona con el grado de complejidad fonológica al nivel de la sílaba, mientras que el número de sílabas por palabra puede asociarse a la complejidad morfológica al nivel de la palabra. Por último, el cociente entre palabras y enunciados puede ser visto como una medida empírica de complejidad sintáctica, ya que los enunciados con más palabras tienden a darse en situaciones en las cuales la sintaxis es más compleja.

En la muestra que utilizamos en este trabajo, el cociente entre fonemas y sílabas va desde un mínimo de 1,7115 (para el caso del igbo, idioma nigercongolés hablado en Nigeria) hasta un máximo de 2,9024 (para el caso del kumiai, idioma yumano hablado en la frontera entre México y Estados Unidos), en un contexto en el cual el número promedio de fonemas por sílaba es 2,2491. El cociente entre sílabas y palabras, por su parte, va desde un mínimo de 1 (para el idioma vietnamita) hasta un máximo de 3,6 (para el telugu, lengua dravídica hablada en la India), en un contexto en el cual el número promedio de sílabas por palabra es 2,1541. Por último, el mínimo cociente entre palabras y enunciados es igual a 4,5, y corresponde al paiute (idioma utoazteca hablado en los Estados Unidos), mientras que el máximo cociente entre palabras y enunciados es 21,67, y corresponde al idioma tongano (en un contexto en el cual el número promedio de palabras por enunciado es 11,15).[6]

3. Coeficientes de correlación

Las medidas lingüísticas descriptas en la sección anterior pueden correlacionarse entre sí. Como tenemos tres cocientes (fonemas por sílaba, sílabas por palabra y palabras por enunciado), podemos hallar tres coeficientes de correlación, que son los que aparecen en el cuadro 1.

Cuadro 1: Coeficientes de correlación estándar

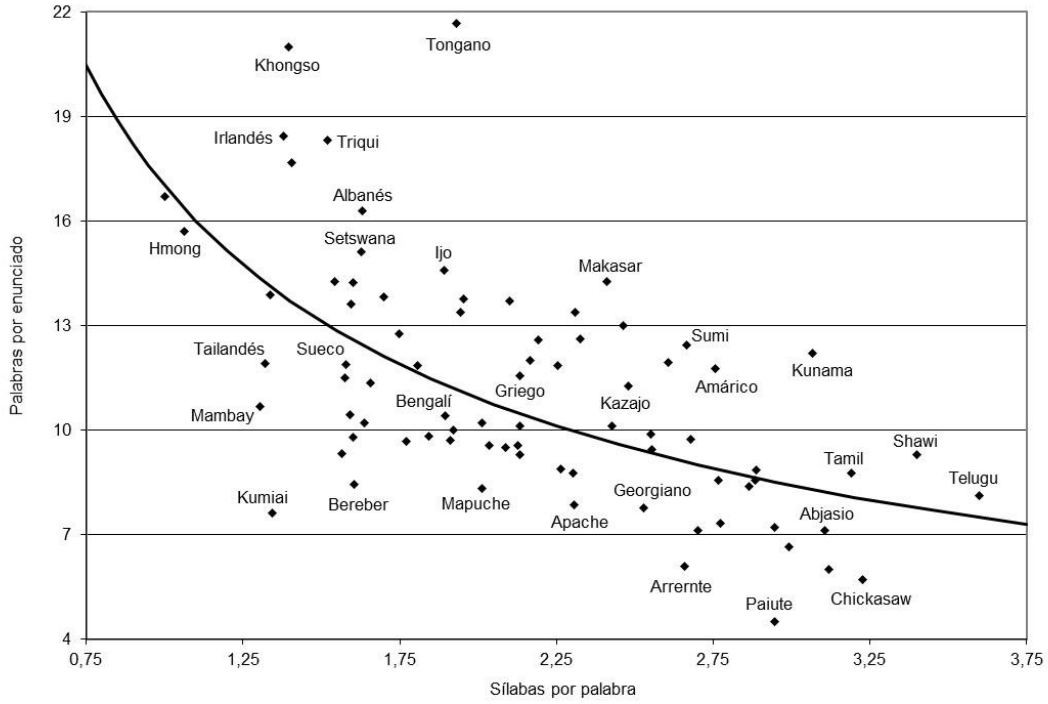
Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Enunciados
Fonemas por sílaba	1,0000		
Sílabas por palabra	-0,2384	1,0000	
Palabras por enunciado	-0,1004	-0,5919	1,0000

El significado básico de estos coeficientes de correlación es que cada cociente lingüístico está negativamente relacionado con los otros dos cocientes. Esto da una pista sobre posibles efectos

[5] Esta interpretación tiene una larga tradición en la literatura sobre lingüística cuantitativa. Véase, por ejemplo, Fenk-Oczlon y Fenk (2008) y Bentz y otros (2023).
[6] La lista completa de los valores de los cocientes lingüísticos aparece en el apéndice 2.

de compensación (*trade-offs*) entre dichas medidas de complejidad idiomática, en el sentido de que, en promedio, una lengua que es más compleja en cierta dimensión tiende a ser más simple en otra dimensión. Por ejemplo, el texto de "El viento norte y el sol" traducido al hmong (idioma de la familia hmong-mien, hablado en China) tiene un promedio de 1,0637 sílabas por palabra, y un promedio de 15,7 palabras por enunciado. En cambio, el mismo texto en chickasaw (de la familia muskogueana, hablado en los Estados Unidos) tiene un promedio de 3,2281 sílabas por palabra, pero solamente 5,7 palabras por enunciado. En este caso, este efecto de compensación detectado parece ser general para toda la base de datos, tal como puede verse en la figura 2 (en la cual cada idioma está representado por un punto en un espacio que relaciona sílabas por palabra con palabras por enunciado, y la línea negra gruesa refleja una tendencia de tipo exponencial).

Figura 2: Relación entre sílabas por palabra y palabras por enunciado



Los valores absolutos de los coeficientes de correlación informados en el cuadro 1 se relacionan también con su significación estadística. Por ejemplo, la correlación entre sílabas por palabra y palabras por enunciado (r = -0,5919) es significativamente distinta de cero a un nivel del 1% de probabilidad, ya que su valor de probabilidad (p = 0,0000) es menor que 1%. En cambio, el coeficiente de correlación entre fonemas por sílaba y sílabas por palabra (r = -0,2384) solo es significativo al 5% de probabilidad, ya que su valor de probabilidad (p = 0,0332) es menor que 5%

pero mayor que 1%. Por último, la correlación entre fonemas por sílaba y palabras por enunciado (r = -0,1004) no es significativa a ningún nivel razonable de probabilidad, puesto que su valor de probabilidad (p = 0,3754) está por encima del 10%.

En Coloma (2017) hay un resultado empírico interesante referido a la correlación entre medidas de complejidad, que aparece al comparar los coeficientes de correlación estándar (o de Pearson) con los correspondientes "coeficientes de correlación parcial". Los coeficientes estándar (que son los informados en el cuadro 1) se calculan usando información de las variables que uno quiere correlacionar, pero no incorporan ninguna información sobre variables adicionales que puedan influenciar las magnitudes que se están comparando. Por el contrario, los coeficientes de correlación parcial se calculan "controlando por" (o sea, usando información sobre) otras variables que puedan estar ellas mismas relacionadas con las dos variables que queremos analizar.

Para calcular un coeficiente de correlación parcial, es necesario eliminar el posible efecto de otros factores sobre las dos variables que uno quiere correlacionar, usando algún procedimiento estadístico. Una posibilidad es comenzar con una matriz de correlación de todas las variables bajo estudio (que en nuestro caso son solo tres variables) y luego invertir dicha matriz. Una vez que hacemos eso, podemos utilizar la siguiente fórmula:

$$r_{xy} = -\frac{i_{xy}}{\sqrt{i_{xx} \cdot i_{yy}}} \qquad (1)\ ;$$

donde i_{xy} es el coeficiente que corresponde al par de variables x e y en la matriz de correlación inversa, e i_{xx} e i_{yy} son los coeficientes que corresponden a esas variables en la diagonal principal de la misma matriz de correlación inversa.[7] Este proceso de inversión matricial es en rigor una de las posibilidades que existen para calcular coeficientes de correlación parcial. Otra alternativa es correr regresiones para tres ecuaciones, en las cuales cada variable dependa de una constante y de las otras dos variables. Ambos procedimientos tienen el mismo objetivo, que es eliminar los efectos que la variable restante puede tener sobre cada par de variables bajo estudio.

Si aplicamos el procedimiento de regresión, en este caso debemos correr un sistema formado por las siguientes funciones lineales:

$$Fon/Sil = c(1) + c(2)*Sil/Pal + c(3)*Pal/Enunc \qquad (2)\ ;$$

$$Sil/Pal = c(4) + c(5)*Fon/Sil + c(6)*Pal/Enunc \qquad (3)\ ;$$

[7] Para una explicación alternativa del concepto de correlación parcial, véase Rasinger (2013), capítulo 7, o Coloma (2018).

*Pal/Enunc = c(7) +c(8)*Fon/Sil +c(9)*Sil/Pal* (4) ;

donde *Fon/Sil*, *Sil/Pal* y *Pal/Enunc* son nuestros tres cocientes lingüísticos, y *c(1), c(2), c(3), c(4), c(5), c(6), c(7), c(8)* y *c(9)* son los coeficientes a estimar.

Cuando se llevan a cabo estas regresiones utilizando el método de mínimos cuadrados ordinarios (MCO),[8] se obtienen los resultados que aparecen en el cuadro 2. Con ellos, los coeficientes de correlación parcial entre los distintos cocientes lingüísticos pueden calcularse a través de la siguiente fórmula:

$$r_{xy} = -\sqrt{c_{xy} \cdot c_{yx}} \quad (5) ;$$

donde r_{xy} es el coeficiente de correlación parcial entre la variable *x* y la variable *y*, c_{xy} es el coeficiente de regresión de la variable *y* en la ecuación de la variable *x*, y c_{yx} es el coeficiente de regresión de la variable *x* en la ecuación de la variable *y*. Nótese que en esta fórmula estamos suponiendo que, como los dos coeficientes de regresión son negativos, el correspondiente coeficiente de correlación parcial también debe ser negativo.

Cuadro 2: Resultados de la regresión para calcular coeficientes de correlación parcial

Concepto	Coeficiente	Estadístico t	Probabilidad
Ecuación Fonemas/Sílabas			
Constante [c(1)]	2,93436	15,19583	0,0000
Sil/Pal [c(2)]	-0,18285	-3,51006	0,0005
Pal/Enunc [c(3)]	-0,02614	-2,84651	0,0048
R cuadrado	0,1466		
Ecuación Sílabas/Palabras			
Constante [c(4)]	5,07345	9,57201	0,0000
Fon/Sil [c(5)]	-0,75439	-3,51006	0,0005
Pal/Enunc [c(6)]	-0,10969	-7,25756	0,0000
R cuadrado	0,4399		
Ecuación Palabras/Enunciados			
Constante [c(7)]	27,31484	8,20534	0,0000
Fon/Sil [c(8)]	-3,64179	-2,84651	0,0048
Sil/Pal [c(9)]	-3,70326	-7,25756	0,0000
R cuadrado	0,4122		

Aplicando nuestra fórmula a los resultados informados en el cuadro 2, resulta posible obtener los coeficientes de correlación parcial que se muestran en el cuadro 3. Si comparamos esos

[8] Estas regresiones, y todas las otras cuyos resultados se informan en el presente artículo, se llevaron a cabo usando el programa estadístico EViews 10.

resultados con los que aparecen en el cuadro 1, veremos que los tres coeficientes de correlación parcial calculados son más altos que sus correspondientes coeficientes de correlación estándar.[9] Esto implica además que los mismos tienen una significación estadística más alta, que en este caso se percibe en el hecho de que ahora los tres coeficientes son significativos al 1% de probabilidad ("p = 0,0007"; "p = 0,0054" y "p = 0,0000").

Cuadro 3: Coeficientes de correlación parcial (MCO)

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Enunciados
Fonemas por sílaba	1,0000		
Sílabas por palabra	-0,3714	1,0000	
Palabras por enunciado	-0,3086	-0,6373	1,0000

La idea de calcular coeficientes de correlación parcial usando un sistema de ecuaciones lineales puede complementarse con el uso de una metodología conocida como "regresión de ecuaciones simultáneas". La misma consiste en correr las tres ecuaciones al mismo tiempo, aprovechando algunas relaciones que puedan encontrarse entre ellas. Una de esas relaciones es la covarianza entre los resultados de dichas ecuaciones, en particular la de los "residuos" de las mismas (o sea, la covarianza entre las partes de las variaciones de cada variable dependiente que no pueden explicarse a través del procedimiento de regresión).

Si ahora introducimos este elemento en la estimación de las ecuaciones simultáneas bajo análisis, estaremos corriendo un sistema de "regresiones aparentemente no relacionadas" (SUR, por su sigla en inglés). Este es un procedimiento relativamente difundido en algunas ciencias sociales tales como la economía, e implica que, cuando uno estima cada ecuación, también usa información de las otras ecuaciones. Dicha información puede mejorar la precisión y la eficiencia estadística de los coeficientes estimados.[10]

Por supuesto, la estimación de las ecuaciones 2, 3 y 4 utilizando SUR produce un conjunto de coeficientes que también puede utilizarse para calcular correlaciones parciales. Para ello, debemos usar la ecuación 5 con los nuevos coeficientes de regresión estimados. Los coeficientes de correlación que surgen de aplicar dicho procedimiento son los que aparecen en el cuadro 4.

[9] Este resultado es equivalente al que aparece en Coloma (2017).
[10] Este procedimiento fue propuesto originalmente por Zellner (1962), y fue utilizado por nosotros en Coloma (2017). Para una explicación de sus propiedades estadísticas, véase Baltagi (2011), capítulo 10.

Cuadro 4: Coeficientes de correlación parcial (SUR)

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Enunciados
Fonemas por sílaba	1,0000		
Sílabas por palabra	-0,6749	1,0000	
Palabras por enunciado	-0,6143	-0,9141	1,0000

Como se ve, los coeficientes de correlación parcial calculados con esta metodología son aún mayores que los informados en el cuadro 3. Todos ellos, además, son estadísticamente significativos al 1% de probabilidad (ya que sus correspondientes valores de probabilidad son todos iguales a 0,0000). Esto puede interpretarse como una señal de que las verdaderas correlaciones negativas entre los distintos cocientes lingüísticos son más altas que las que se obtienen cuando se utiliza un enfoque más simplificado.

4. Variables geográficas, filogenéticas y demográficas

En Coloma (2017, 2020), se incluye una extensión del análisis realizado hasta aquí que explora la posibilidad de calcular coeficientes de correlación parcial controlando por el efecto de otras variables adicionales relacionadas con factores geográficos, filogenéticos y demográficos. Eso se lleva a cabo corriendo un sistema de ecuaciones que incluye dichas variables adicionales, tal como el que se muestra a continuación:

$$\begin{aligned} Fon/Sil = {} & c(1)*Europa + c(2)*Africa + c(3)*Asiaoccidental + c(4)*Asiaoriental \\ & + c(5)*Australasia + c(6)*America + c(7)*Indoeuropeo + c(8)*Afroasiatico \\ & + c(9)*Nigercongo + c(10)*Sinotibetano + c(11)*Austronesio + c(12)*Principal \\ & + c(13)*Sil/Pal + c(14)*Pal/Enunc \end{aligned} \quad (6)\ ;$$

$$\begin{aligned} Sil/Pal = {} & c(21)*Europa + c(22)*Africa + c(23)*Asiaoccidental + c(24)*Asiaoriental \\ & + c(25)*Australasia + c(26)*America + c(27)*Indoeuropeo + c(28)*Afroasiatico \\ & + c(29)*Nigercongo + c(30)*Sinotibetano + c(31)*Austronesio + c(32)*Principal \\ & + c(33)*Fon/Sil + c(34)*Pal/Enunc \end{aligned} \quad (7)\ ;$$

$$\begin{aligned} Pal/Enunc = {} & c(41)*Europa + c(42)*Africa + c(43)*Asiaoccidental + c(44)*Asiaoriental \\ & + c(45)*Australasia + c(46)*America + c(47)*Indoeuropeo + c(48)*Afroasiatico \\ & + c(49)*Nigercongo + c(50)*Sinotibetano + c(51)*Austronesio + c(52)*Principal \\ & + c(53)*Fon/Sil + c(54)*Sil/Pal \end{aligned} \quad (8)\ ;$$

donde *Europa, Africa, Asiaoccidental, Asiaoriental, Australasia* y *America* son variables binarias que toman un valor igual a uno cuando un idioma pertenece a cierta región (y cero en caso contrario); *Indoeuropeo, Afroasiatico, Nigercongo, Sinotibetano* y *Austronesio* son variables que

toman un valor igual a uno cuando un idioma pertenece a cierta familia lingüística;[11] y *Principal* es una variable que toma un valor igual a uno cuando un idioma es hablado por más de 5 millones de personas.[12]

Como puede observarse, este nuevo conjunto de ecuaciones es básicamente equivalente al formado por las ecuaciones 2, 3 y 4, con el agregado de las doce variables binarias que incluyen factores que pueden influenciar la complejidad de los idiomas pero que son esencialmente "no lingüísticos". Como este conjunto de ecuaciones es un sistema de regresiones simultáneas, el procedimiento de regresiones aparentemente no relacionadas descripto en la sección anterior también puede utilizarse aquí. El resultado de ello es el que aparece en el cuadro 5, en el cual se informan los valores de los 42 coeficientes de regresión estimados, junto con sus correspondientes valores de probabilidad y con los respectivos coeficientes R cuadrado.

Cuadro 5: Resultados de la regresión con variables adicionales (SUR)

Variable / Concepto	Ecuación Fon/Sil		Ecuación Sil/Pal		Ecuación Pal/Enunc	
	Coeficiente	Probab.	Coeficiente	Probab.	Coeficiente	Probab.
Europa	3,55344	0,0000	6,85140	0,0000	40,34142	0,0000
Africa	3,44849	0,0000	6,70773	0,0000	39,43959	0,0000
Asiaoccidental	3,56763	0,0000	6,96004	0,0000	40,43655	0,0000
Asiaoriental	3,53805	0,0000	6,61020	0,0000	39,52971	0,0000
Australasia	3,50096	0,0000	6,73787	0,0000	38,92927	0,0000
America	3,57064	0,0000	6,88857	0,0000	39,85420	0,0000
Indoeuropeo	-0,09118	0,3684	-0,29547	0,1281	-0,33895	0,7993
Afroasiatico	0,03267	0,7625	-0,13081	0,5389	-0,22916	0,8736
Nigercongo	-0,25798	0,0145	-0,52889	0,0101	-1,95438	0,1687
Sinotibetano	-0,05191	0,6094	0,05994	0,7669	0,46891	0,7296
Austronesio	-0,02710	0,7886	0,03414	0,8644	1,53415	0,2442
Principal	0,11037	0,0846	0,21668	0,0835	0,48289	0,5712
Fon/Sil			-1,45471	0,0000	-7,55362	0,0000
Sil/Pal	-0,36863	0,0000			-5,40126	0,0000
Pal/Enunc	-0,04277	0,0000	-0,12069	0,0000		
R cuadrado	0,2865		0,5213		0,3663	

Si ahora utilizamos la ecuación 5 con estos nuevos coeficientes de regresión (o sea, con los que corresponden a los cocientes lingüísticos en cada una de las ecuaciones), podemos obtener

[11] Estas son en rigor las cinco familias que tienen seis o más idiomas incluidos en la muestra. Otras ocho familias tienen dos idiomas cada una, y son las siguientes: austroasiática, dravídica, sudánica oriental, otomangueana, pamañungana, túrquica, urálica y utoazteca. Véase apéndice 2.

[12] De acuerdo con esta definición, los "idiomas principales" en nuestra muestra son los siguientes: albanés, amárico, árabe, bengalí, bereber, birmano, coreano, georgiano, griego, hausa, húngaro, igbo, italiano, japonés, kazajo, malayo, malgache, mandarín, panyabí, persa, setswana, sueco, tailandés, tamil, telugu, turco, ucraniano y vietnamita.

nuevos coeficientes de correlación parcial, que son los que se informan en el cuadro 6. Una vez más, estos coeficientes son estadísticamente significativos al 1% de probabilidad, y eso puede ser visto como una señal de que los fenómenos de correlación negativa que encontramos siguen siendo importantes cuando uno controla por la interacción entre medidas de complejidad idiomática y distintos factores de carácter geográfico, filogenético y demográfico.

Cuadro 6: Coeficientes de correlación parcial con variables adicionales (SUR)

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Enunciados
Fonemas por sílaba	1,0000		
Sílabas por palabra	-0,7323	1,0000	
Palabras por enunciado	-0,5684	-0,8074	1,0000

5. Variables instrumentales

Una complicación adicional que puede incluirse en el análisis es el uso de las llamadas "variables instrumentales". Las mismas resultan útiles cuando tenemos una situación de endogeneidad entre las variables que uno quiere correlacionar. En este caso, por ejemplo, estamos corriendo un sistema de regresiones cuya primera ecuación tiene a *Fon/Sil* como variable dependiente, en tanto que *Sil/Pal* es considerada como una variable independiente. Pero dicha situación se invierte en la segunda ecuación, ya que *Sil/Pal* es allí la variable dependiente, mientras que *Fon/Sil* es una de las variables independientes.

Este hecho, y uno similar que ocurre con la tercera ecuación del sistema, se considera en general como una violación de los supuestos estadísticos implícitos en la lógica del método de regresión por mínimos cuadrados, que supone que las variables independientes no deben estar influenciadas por otras variables del sistema (es decir, que deben ser "exógenas"). Si estamos estimando una regresión en la cual algunas variables dependientes son en realidad endógenas, dicha situación debe ser resuelta usando nuevas variables que reemplacen a las originales. Esas son las variables instrumentales que tenemos que crear, y deben ser variables que tengan una relación con las originales pero que, al mismo tiempo, sean exógenas al problema estadístico bajo estudio.[13]

Una forma relativamente directa de resolver en este caso el problema de endogeneidad es usar variables tipológicas que describan algunas características de los idiomas, pero que sean independientes del texto analizado. Estas pueden ser variables relacionadas con la fonología de

[13] Para una explicación más completa de todo esto, véase Baltagi (2011), capítulo 11.

dichos idiomas (p.ej., inventario de consonantes, inventario de vocales, número de tonos) o con su morfosintaxis (p.ej., número de casos, número de géneros, número de categorías de inflexión verbal). También pueden ser variables binarias tales como el nivel de complejidad silábica (simple versus complejo), uso del acento (distintivo versus no distintivo), orden de las palabras (objeto-verbo versus verbo-objeto, adjetivo-sustantivo versus sustantivo-adjetivo), alineamiento morfosintáctico (acusativo versus no acusativo) y grado de síntesis morfológica de la lengua (aislante versus concatenada).

A efectos de crear todas esas variables para los 80 idiomas de nuestra muestra, utilizamos información extraída de las mismas fuentes del texto de "El viento norte y el sol" (o sea, de las correspondientes ilustraciones del alfabeto fonético internacional) y también del Atlas Mundial de Estructuras Lingüísticas (Dryer y Haspelmath, 2013).[14] Luego usamos esas variables para llevar a cabo un procedimiento conocido como "regresión por mínimos cuadrados en tres etapas", que implica reemplazar las variables endógenas originales (en nuestro caso, *Fon/Sil*, *Sil/Pal* y *Pal/Enunc*) por combinaciones lineales de las variables exógenas.[15]

La manera de obtener combinaciones lineales que reemplacen a las variables endógenas originales es correr un nuevo sistema de regresiones por mínimos cuadrados ordinarios, en el cual *Fon/Sil*, *Sil/Pal* y *Pal/Enunc* aparezcan como funciones de las doce variables tipológicas y de las doce variables geográficas, filogenéticas y demográficas descriptas en la sección anterior. Una vez que tenemos los resultados de dichas regresiones, los coeficientes estimados (ver cuadro 7) pueden usarse para crear nuevas variables (que surgen de multiplicar dichos coeficientes por los valores de las correspondientes variables exógenas). Con eso podemos generar tres variables instrumentales (que llamaremos *Fôn/Sîl*, *Sîl/Pâl* y *Pâl/Enûnc*) que sirven para reemplazar a las variables originales de los sistemas de ecuaciones, y que al mismo tiempo son completamente exógenas a dichos sistemas.

El siguiente paso consiste en correr el conjunto de ecuaciones en el cual estamos interesados (por ejemplo, el formado por las ecuaciones 6, 7 y 8) pero usando las variables instrumentales *Fôn/Sîl*, *Sîl/Pâl* y *Pâl/Enûnc* en lugar de las variables originales. Si en esta última etapa introducimos un procedimiento basado en regresiones aparentemente no relacionadas, lo que obtenemos es una estimación por mínimos cuadrados en tres etapas (MC3E) cuyos coeficientes

[14] La lista completa de los valores de las variables tipológicas exógenas aparece en los apéndices 3 y 4.
[15] Este procedimiento fue propuesto originalmente por Zellner y Theil (1962). Nosotros ya lo hemos utilizado anteriormente en Coloma (2016).

pueden a su vez usarse para calcular nuevos coeficientes de correlación parcial. Dichos coeficientes de correlación parcial son los que se muestran en el cuadro 8.

Cuadro 7: Resultados de la regresión para crear variables instrumentales (MCO)

Variable / Concepto	Ecuación Fon/Sil		Ecuación Sil/Pal		Ecuación Pal/Enunc	
	Coeficiente	Probab.	Coeficiente	Probab.	Coeficiente	Probab.
Europa	2,22867	0,0000	1,89738	0,0000	12,83559	0,0001
Africa	2,35078	0,0000	2,06875	0,0000	10,01171	0,0016
Asiaoccidental	2,29883	0,0000	2,35648	0,0000	9,32051	0,0026
Asiaoriental	2,44186	0,0000	1,85435	0,0001	8,50132	0,0089
Australasia	2,40828	0,0000	2,02747	0,0000	9,44652	0,0002
America	2,39036	0,0000	2,42498	0,0000	7,77044	0,0009
Indoeuropeo	-0,08629	0,5337	-0,50492	0,0534	2,31852	0,2067
Afroasiatico	0,03238	0,8244	-0,20529	0,4536	0,88456	0,6466
Nigercongo	-0,26640	0,0546	-0,29204	0,2597	0,88102	0,6289
Sinotibetano	-0,11911	0,3544	0,24490	0,3103	0,44635	0,7927
Austronesio	-0,13164	0,3795	-0,00598	0,9830	4,05572	0,0416
Principal	0,09169	0,2698	0,23164	0,1379	-1,07502	0,3276
Consonantes	-0,00065	0,8788	0,00403	0,6114	0,01420	0,7995
Vocales	0,00743	0,3575	0,00074	0,9610	0,05057	0,6354
Tonos	0,00346	0,8963	-0,15910	0,0016	1,15964	0,0011
Acento	0,00305	0,9695	0,33350	0,0268	-0,33533	0,7502
Sílabas Complejas	0,20446	0,0090	-0,29481	0,0437	-0,29802	0,7710
Casos	-0,00573	0,7054	0,05002	0,0798	-0,35728	0,0758
Géneros	-0,00415	0,8700	0,03753	0,4302	-0,18656	0,5777
Inflexiones	-0,02952	0,0654	-0,00801	0,7888	0,06830	0,7459
Objeto-Verbo	-0,05473	0,4992	0,11471	0,4504	0,40110	0,7078
Adjetivo-Sustantivo	0,00439	0,9487	-0,09435	0,4611	1,14516	0,2050
Acusativo	-0,01812	0,8012	0,04354	0,7472	-1,20733	0,2056
Aislante	-0,02651	0,7771	0,09409	0,5924	-1,57612	0,2040
R cuadrado	0,3891		0,6579		0,4720	

Cuadro 8: Coeficientes de correlación parcial con variables adicionales (MC3E)

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Enunciados
Fonemas por sílaba	1,0000		
Sílabas por palabra	-0,9446	1,0000	
Palabras por enunciado	-0,8853	-0,9849	1,0000

Tal como hicimos en las secciones anteriores, ahora resulta posible observar los distintos valores de los coeficientes de correlación y su significación estadística. Vemos así que dichos coeficientes son todos más altos que los obtenidos cuando utilizamos mínimos cuadrados ordinarios y SUR, y que todos ellos son significativos al 1% de probabilidad. Esto puede ser considerado como una señal de que los coeficientes de correlación negativa hallados anteriormente

son robustos a la remoción de los posibles sesgos por endogeneidad que pueden llegar a tener los datos utilizados.

6. Conclusiones y comparación con resultados anteriores

Los resultados hallados en las secciones anteriores, generados a través de una base de datos de 80 idiomas para los cuales tenemos el texto de "El viento norte y el sol", pueden compararse con los obtenidos en Coloma (2017) para la base de datos original de 50 idiomas. Si llevamos a cabo dicha comparación (ver cuadro 9), podemos concluir que muchos de los hechos estilizados que aparecen en el trabajo original siguen vigentes.

Cuadro 9: Comparación de resultados

Concepto	Muestra original		Muestra nueva	
	Coeficiente	Probabilidad	Coeficiente	Probabilidad
Correlación estándar				
Fon/Sil vs. Sil/Pal	-0,2420	0,0905	-0,2384	0,0332
Fon/Sil vs. Pal/Enunc	-0,0522	0,7187	-0,1004	0,3754
Sil/Pal vs. Pal/Enunc	-0,6785	0,0000	-0,5919	0,0000
Correlación parcial (MCO)				
Fon/Sil vs. Sil/Pal	-0,3781	0,0074	-0,3714	0,0007
Fon/Sil vs. Pal/Enunc	-0,3036	0,0340	-0,3086	0,0054
Sil/Pal vs. Pal/Enunc	-0,7132	0,0000	-0,6373	0,0000
Correlación parcial (SUR)				
Fon/Sil vs. Sil/Pal	-0,6730	0,0000	-0,6749	0,0000
Fon/Sil vs. Pal/Enunc	-0,6047	0,0000	-0,6143	0,0000
Sil/Pal vs. Pal/Enunc	-0,9486	0,0000	-0,9141	0,0000
Correl parcial c/var adicionales (SUR)				
Fon/Sil vs. Sil/Pal	-0,5852	0,0001	-0,7323	0,0000
Fon/Sil vs. Pal/Enunc	-0,4163	0,0068	-0,5684	0,0000
Sil/Pal vs. Pal/Enunc	-0,8990	0,0000	-0,8074	0,0000
Correl parcial c/var adicionales (MC3E)				
Fon/Sil vs. Sil/Pal	-0,5325	0,0003	-0,9446	0,0000
Fon/Sil vs. Pal/Enunc	-0,2350	0,1391	-0,8853	0,0000
Sil/Pal vs. Pal/Enunc	-0,9361	0,0000	-0,9849	0,0000

En el cuadro 9 puede observarse, por ejemplo, que los coeficientes de correlación parcial son más altos que los correspondientes coeficientes de correlación estándar en las dos muestras analizadas, y que tienden a incrementarse aún más cuando se utilizan métodos de estimación basados en regresiones aparentemente no relacionadas (SUR) o mínimos cuadrados en tres etapas (MC3E). El "ránking" entre coeficientes de correlación tampoco se altera por el uso de una

metodología u otra. Vemos así que el coeficiente más grande es siempre el que relaciona *Sil/Pal* con *Pal/Enunc*, seguido del que relaciona *Fon/Sil* con *Sil/Pal*, mientras que el coeficiente de correlación entre *Fon/Sil* y *Pal/Enunc* es siempre el que tiene un valor absoluto más bajo.

En lo que se refiere al uso de variables instrumentales, los resultados no son exactamente los mismos. Para la base de datos original utilizada en Coloma (2017), el coeficiente de correlación entre fonemas por sílaba y palabras por enunciado (r = -0,2350) no es estadísticamente significativo cuando se lo calcula utilizando MC3E.[16] Ese no es el caso cuando se utilizan los datos de la nueva muestra, ya que ahora dicho coeficiente es "r = -0,8853", y esa cifra es estadísticamente significativa a cualquier nivel razonable de probabilidad. Este resultado es importante, porque los coeficientes obtenidos a través de procedimientos de mínimos cuadrados en tres etapas tienen la propiedad de que son consistentes e insesgados, y son por lo tanto menos sensibles a las influencias de posibles shocks estocásticos. Como la nueva muestra de idiomas, además, es más grande y diversa que la muestra original, las nuevas cifras halladas son probablemente más representativas que las anteriores, y muestran que los efectos de compensación que están detrás de los coeficientes de correlación negativos son en realidad más robustos que los que aparecen utilizando una muestra más reducida.

Después de llevar a cabo distintos tipos de cálculos y estimaciones, y de compararlos con los resultados que se habían obtenido con la base de datos original, resulta posible derivar algunas conclusiones y comentarios adicionales. La conclusión más importante es que los efectos de compensación entre medidas de complejidad idiomática que se detectaron en el trabajo original también aparecen en este nuevo estudio. Las principales diferencias entre los resultados originales y los actuales, sin embargo, parecen ser las siguientes:

a) Los coeficientes de correlación entre fonemas por sílaba y palabras por enunciado son siempre más altos con la nueva base de datos que con la original.

b) Los coeficientes de correlación entre fonemas por sílaba y sílabas por palabra también son mayores con la nueva base de datos, siempre que controlemos por el efecto de ciertas variables no lingüísticas. Si no lo hacemos, en cambio, los coeficientes obtenidos son prácticamente idénticos para la muestra original y para la nueva muestra.

c) Por el contrario, los coeficientes de correlación entre sílabas por palabra y palabras por

[16] Este resultado no aparece en el artículo original, el cual no incluía un análisis del problema de endogeneidad a través del empleo de variables instrumentales.

enunciado son en general menores en las estimaciones realizadas con la nueva base de datos, excepto para el caso en el cual utilizamos variables instrumentales a fin de remover posibles sesgos relacionados con la endogeneidad de las variables correlacionadas.

Referencias bibliográficas

Baltagi, Badi (2011) *Econometrics*, 5ta. edición. Berlín, Springer.

Bentz, Chistian, Ximena Gutiérrez, Olga Sozinova y Tanja Samardzic (2023) Complexity Trade-Offs and Equi-Complexity in Natural Languages: A Meta-Analysis. *Linguistics Vanguard*, en prensa.

Coloma, Germán (2016) A Simultaneous-Equation Regression Model of Language Complexity Trade-Offs, Documento de Trabajo Nro. 497. Buenos Aires, Universidad del CEMA.

Coloma, Germán (2017) The Existence of Negative Correlation between Linguistic Measures across Languages. *Corpus Linguistics and Linguistic Theory* 13: 1-26.

Coloma, Germán (2018) Relaciones entre medidas de complejidad lingüística, Documento de Trabajo Nro. 658. Buenos Aires, Universidad del CEMA.

Coloma, Germán (2020) Language Complexity Trade-Offs Revisited, Documento de Trabajo Nro. 721. Buenos Aires, Universidad del CEMA.

Coloma, Germán (2022) Correlation between Linguistic Measures: An Extended Analysis. *Studies in Linguistics and Literature* 6(4): 109-132.

Dryer, Matthew y Martin Haspelmath (2013) *The World Atlas of Language Structures Online*. Leipzig, Max Planck Institute for Evolutionary Anthropology.

Fenk-Oczlon, Gertraud y August Fenk (2008). Complexity Trade-Offs Between the Subsystems of Language. En M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change*, 43-65. Amsterdam, John Benjamins.

IPA (1999) *Handbook of the International Phonetic Association*. Cambridge, Cambridge University Press.

Rasinger, Sebastian (2013) *Quantitative Research in Linguistics*, 2da. edición. Londres, Bloomsbury (hay versión en castellano: *La investigación cuantitativa en lingüística*; Madrid, Akal).

Zellner, Arnold (1962) An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias. *Journal of the American Statistical Association* 57: 348–368.

Zellner, Arnold y Henri Theil (1962) Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica* 30: 54–78.

Fuentes de los datos utilizados

Anderson, Samuel, Bert Vaux y Zihni Pysipa (2023) Cwyzhy Abkhaz. *Journal of the International Phonetic Association*, en prensa.

Anonby, Erik (2006) Mambay. *Journal of the International Phonetic Association* 36: 221-233.

Arnold, Laura (2022) Ambel. *Journal of the International Phonetic Association* 52: 368-388.

Arvaniti, Amalia (1999) Standard Modern Greek. *Journal of the International Phonetic Association* 29: 167-172.

Ashkaba, John y Richard Hayward (1999) Kunama. *Journal of the International Phonetic Association* 29: 179-185.

Asu, Eva y Pire Teras (2009) Estonian. *Journal of the International Phonetic Association* 39: 367-372.

Babel, Molly, Michael Houser y Maziar Toosarvandani (2012) Mono Lake Northern Paiute. *Journal of the International Phonetic Association* 42: 233-243.

Bennett, W. G., Maxine Diemer, Justine Kerford, Tracy Probert y Tsholofelo Wesi (2016) Setswana (South African). *Journal of the International Phonetic Association* 46: 235-246.

Bhaskararao, Peri y Arpita Ray (2017) Telugu. *Journal of the International Phonetic Association* 47: 231-241.

Breen, Gavan y Veronica Dobson (2005) Central Arrernte. *Journal of the International Phonetic Association* 35: 249-254.

Chirkova, Katia y Yiya Chen (2013) Lizu. *Journal of the International Phonetic Association* 43: 75-86.

Clynes, Adrian y David Deterding (2011) Standard Malay (Brunei). *Journal of the International Phonetic Association* 41: 259-268.

Coretta, Stefano, Josiane Riverin, Enkeleida Kapia y Stephen Nichols (2023) Northern Tosk Albanian. *Journal of the International Phonetic Association*, en prensa.

Dawd, Abushush y Richard Hayward (2002) Nara. *Journal of the International Phonetic Association* 32: 249-255.

Eaton, Helen (2006) Sandawe. *Journal of the International Phonetic Association* 36: 235-242.

Edmonson, Jerold, John Esling y Li Shaoni (2021) Jianchuan Bai. *Journal of the International Phonetic Association* 51: 490-501.

Elias-Ulloa, José y Rolando Muñoz (2021) Upper-Chambira Urarina. *Journal of the International Phonetic Association* 51: 137-169.

Elliot, Raymond, Jerold Edmonson y Fausto Sandoval (2016) Chicahuaxtla Triqui. *Journal of the International Phonetic Association* 46: 351-365.

Engstrand, Olle (1999) Swedish. En IPA (1999), 140-142.

Evans, Nicholas y Julia Miller (2016) Nen. *Journal of the International Phonetic Association* 46: 331-349.

Garellek, Marc y Marija Tabain (2020) Tongan. *Journal of the International Phonetic Association* 50: 406-416.

Gil Burgoin, Carlos (2023) Northern Tepehuan. *Journal of the International Phonetic Association*, en prensa.

Gordon, Matthew, Pamela Munro y Peter Ladefoged (2001) Chickasaw. *Journal of the International Phonetic Association* 31: 287-290.

Hargus, Sharon y Virginia Beavert (2014) Northwest Sahaptin. *Journal of the International Phonetic Association* 44: 320-342.

Harry, Otelemate (2003) Kalabari-Ijo. *Journal of the International Phonetic Association* 33: 113-120.

Hayward, Katrina y Richard Hayward (1999) Amharic. En IPA (1999), 45-50.

Herrera Zendejas, Esther (2023) Mecapalapa Tepehua. *Journal of the International Phonetic Association*, en prensa.

Heston, Tyler y Stephanie Locke (2019) Fataluku. *Journal of the International Phonetic Association* 49: 419-425.

Howe, Penelope (2021) Central Malagasy. *Journal of the International Phonetic Association* 51: 103-136.

Hualde, José, Oihana Lujanbio y Juan Zubiri (2010) Goizueta Basque. *Journal of the International Phonetic Association* 40: 113-127.

Hussain, Qandeel, Michael Proctor, Mark Harvey y Katherine Demuth (2020) Punjabi (Lyallpuri variety). *Journal of the International Phonetic Association* 50: 282-297.

Ikekeonwu, Clara (1999) Igbo. En IPA (1999), 108-110.

Khan, Sameer (2010) Bengali (Bangladeshi Standard). *Journal of the International Phonetic Association* 40: 221-225.

Kanu, Sullay y Benjamin Tucker (2010) Temne. *Journal of the International Phonetic Association* 40: 247-253.

Keane, Elinor (2004) Tamil. *Journal of the International Phonetic Association* 34: 111-116.

Kirby, James (2011) Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association* 41: 381-392.

Kruspe, Nicole y John Hajek (2009) Mah Meri. *Journal of the International Phonetic Association* 39: 241-248.

Lee, Hyun Bok (1999) Korean. En IPA (1999), 120-123.

Lee, Wai-Sum y Eric Zee (2003) Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33: 109-112.

Lichtman, Karen, Shawn Chang, Jennifer Kramer, Claudia Crespo, Jill Hallett, Amanda Huensch y Alexandra Morales (2010) IPA Illustration of Q'anjob'al; *Studies in the Linguistic Sciences*. Urbana, Universidad de Illinois.

Lindsay, Kate (2022) Ende. *Journal of the International Phonetic Association* 52: 581-601.

Liu, Wen, Youjing Lin, Zhenghui Yang y Jiangping Kong (2020) Hmu (Xinzhai variety). *Journal of the International Phonetic Association* 50: 240-257.

Mai, Anna, Andrés Aguilar y Gabriela Cabellero (2019) J'aa Kumiai. *Journal of the International Phonetic Association* 49: 231-244.

Majidi, Mohammad y Elmar Ternes (1999) Persian (Farsi). En IPA (1999), 124-125.

Makasso, Emmanuel y Seunghun Lee (2015) Basaá. *Journal of the International Phonetic Association* 45: 71-79.

Marlett, Stephen, Xavier Moreno y Genaro Herrera (2005) Seri. *Journal of the International Phonetic Association* 35: 117-121.

Masaquiza, Fanny y Stephen Marlett (2008) Salasaca Quichua. *Journal of the International Phonetic Association* 38: 223-227.

McCollum, Adam y Si Chen (2021) Kazakh. *Journal of the International Phonetic Association* 51: 276-298.

McPherson, Laura (2020) Seenku. *Journal of the International Phonetic Association* 50: 220-239.

Ní Chasaide, Ailbhe (1999) Irish. En IPA (1999), 111-116.

Okada, Hideo (1999) Japanese. En IPA (1999), 117-119.

Olson, Kenneth (2004) Mono. *Journal of the International Phonetic Association* 34: 233-238.

Padayodi, Cécile (2008) Kabiye. *Journal of the International Phonetic Association* 38: 215-221.

Pearce, Mary (2011) Kera. *Journal of the International Phonetic Association* 41: 249-258.

Pickett, Velma, María Villalobos y Stephen Marlett (2010) Isthmus (Juchitán) Zapotec. *Journal of the International Phonetic Association* 40: 365-372.

Pompino-Marschall, Bernd, Elena Steriopolo y Marzena Zygis (2017) Ukrainian. *Journal of the International Phonetic Association* 47: 349-357.

Remijsen, Bert y Caguor Manyang (2009) Luanyjang Dinka. *Journal of the International Phonetic Association* 39: 123-124.

Repetti-Ludlow, Chiara, Haoru Zhang, Hugo Lucitante, Scott Ander Bois y Chelsea Sanker (2020) A'ingae (Cofán). *Journal of the International Phonetic Association* 50: 431-444.

Ridouane, Rachid (2014) Tashlhiyt Berber. *Journal of the International Phonetic Association* 44: 207-221.

Riehl, Anastasia y Dorothy Jauncey (2005) Tamambo. *Journal of the International Phonetic Association* 35: 255-259.

Rogers, Derek y Luciana d'Arcangeli (2004) Italian. *Journal of the International Phonetic Association* 34: 117-121.

Rojas-Berscia, Luis, Andrés Napurí y Lei Wang (2020) Shawi (Chayahuita). *Journal of the International Phonetic Association* 50: 417-430.

Rose, Françoise (2022) Mojeño Trinitario. *Journal of the International Phonetic Association* 52: 562-580.

Sadowsky, Scott, Héctor Painequeo, Gastón Salamanca y Heriberto Avelino (2013) Mapudungun. *Journal of the International Phonetic Association* 43: 87-96.

Schuh, Russell y Lawan Yalwa (1999) Hausa. En IPA (1999), 90-95.

Shosted, Ryan y Vakhtang Chikovani (2006) Standard Georgian. *Journal of the International Phonetic Association* 36: 255-264.

Soderberg, Craig, Seymour Ashley y Kenneth Olson (2012) Tausug (Suluk). *Journal of the International Phonetic Association* 42: 361-364.

Szende, Tamás (1999) Hungarian. En IPA (1999), 104-107.

Tabain, Marija y Andrew Butcher (2014) Pitjantjatjara. *Journal of the International Phonetic Association* 44: 189-200.

Tabain, Marija y Anthony Jukes (2016) Makasar. *Journal of the International Phonetic Association* 46: 99-111.

Tabain, Marija y Birgit Hellwig (2023) Qaqet. *Journal of the International Phonetic Association*, en prensa.

Tench, Paul (2007) Tera. *Journal of the International Phonetic Association* 37: 227-234.

Teo, Amos (2012) Sumi (Sema). *Journal of the International Phonetic Association* 42: 365-373.

Thelwall, Robin y Akram Sa'adeddin (1999) Arabic. En IPA (1999), 51-54.

Tingsabadh, Kalaya y Arthur Abramson (1999) Thai. En IPA (1999), 147-150.

Tuttle, Siri y Merton Sandoval (2002) Jicarilla Apache. *Journal of the International Phonetic Association* 32: 105-112.

Valenzuela, Pilar, Luis Márquez e Ian Maddieson (2001) Shipibo. *Journal of the International Phonetic Association* 31: 281-285.

Watkins, Justin (2001) Burmese. *Journal of the International Phonetic Association* 31: 291-295.

Wright, Johathan (2022) Khongso. *Journal of the International Phonetic Association* 52: 521-540.

Zimmer, Karl & Orhan Orgun (1999) Turkish. En IPA (1999), 154-156.

Apéndice 1: Ejemplos del texto de "El viento norte y el sol"

Español (no incluido en la muestra)

El viento norte y el sol discutían sobre cuál de ellos era el más fuerte, cuando pasó un viajero envuelto en una ancha capa. El viento y el sol convinieron en que quien antes lograra obligar al viajero a quitarse la capa sería considerado más poderoso. El viento norte sopló con gran furia, pero cuanto más soplaba, más se agarraba el viajero de su capa. Por fin el viento norte abandonó la empresa. Entonces brilló el sol con ardor, e inmediatamente el viajero se despojó de su capa, por lo que el viento norte tuvo que reconocer la superioridad del sol.

Árabe (Arabia, afroasiático)

Kaanat riihu al-shamaali tatazhaadalu wa al-shamsu fii ayyin minhumaa kaanat aqwaa min al-ukraa, wa id bi-musaafirin yatla‘u mutalaffi‘an bi-‘abaa‘atin sami ikatin. Fa ittafaqataa ‘alaa i‘tibaari al-saabiqi fii izhbaari al-musaafiri ‘alaa kal‘i ‘abaa'atihi al-aqwaa. ‘Asafat riihu al-shamaali bi-aqsaa maa istataa‘at min quuwatin. Wa laakin kullumaa izdaada al‘asfu, izdaada al-musaafiru tadatturan bi-‘abaa‘atihi, ilaa an usqita fii yadi al-riih fatakallat ‘an muhaawalatihaa. Ba‘da‘idin sata‘ati al-shamsu bi-dif'ihaa, famaakaana min al-musaafiri illaa an kala‘a ‘abaa‘atahu ‘alaa al-tauu. Wa hakadaa idtarrat riihu al-shamaali ilaa al-i‘tiraafi bi-anna al-shamsa kaanat hiya al-aqwaa.

Húngaro (Hungría, urálico)

Egyszer az északi szél és a nap vetélkedtek, hogy melyikük az erősebb, Épp arra jött egy vándor, vastag köpönyegbe burkolódzva. Az északi szél és a nap nyomban megegyeztek, hogy az lesz a gyóztes, aki hamarabb rábírja a vándort, hogy levegye a köpönyegét, Akkor az északi szél elkezdett süvölteni, ahogy csak bírt. De a vándor annál szorosabban vonta maga köré a köpenyt, minél erősebben fújt. Igy aztán az északi szél el is veszítette a versenyt. A nap meg elkezdte ontani túzó sugarait, mire a vándor egyszeriben kibújt a köpönyegéból. Az északi szél kénytelen volt megadni, hogy bizony a nap az erősebb.

Igbo (Nigeria, nigercongolés)

Ikùkù úgùrù nà Anwū nà-arúrítá úkà ónyé kā ībè yá íké mgbè há hùrù ótù ónyé íjè kà ó yì ùwé úgùrù yá nà-àbiá. Há kwèkòrìtàrà nà ónyé būrū úzò méé kà ónyé íjè áhù yípù ùwé yā kà á gà-éwè díkà ónyé kà íbè yá íké. Ikùkù úgùrù wéé màlíté féé, féé, òtù íké yā hà; mà kà ó nà-èfé kà ónyé íjè áhù nà-èjídésí ūwē yā īkē nà àhú yā. Yá fékàtá hápù. Mgbè áhù Anwū wéé chápùtá, chásíkē, méé kà ébé níílē kpòró ókū; ná-ātùfūghì ógè ónyé íjè áhù yìpùrù ùwé yā. Nké à mèrè ìkùkù úgùrù kwèrè nà Anwū kà yá íké.

Italiano (Italia, indoeuropeo)

Il vento del nord ed il sole stavano discutendo su chi, tra i due, fosse il più forte quando arrivò un viaggiatore avvolto in un mantello. I due decisero che il primo di loro che fosse riuscito a far togliere il mantello al viaggiatore sarebbe stato il più forte tra i due. Quindi il vento del nord soffiò più forte che mai, ma più lui soffiava più il viaggiatore si avvolgeva nel suo mantello; fin a quando il vento rinunciò. Allora il sole lo riscaldò con i suoi raggi e, immediatamente, il viaggiatore si tolse il mantello. Fu così che il vento del nord ammise che il sole era il più forte tra i due.

Japonés (Japón, japónico)

Arutoki Kitakaze to Taiyou ga chikara-kurabe wo shimashita. Tabibito no gaitou wo nugaseta hou ga kachi to yuu koto ni kimete, mazu Kitakaze kara hajimemashita. Kitakaze wa, “Nani, hitomakuri ni shite miseyou”, to, hageshiku fukitatemashita. Suru to tabibito wa, Kitakaze ga fukeba fuku hodo gaitou wo shikkari to karada ni kuttsukemashita. Kondo wa Taiyou no ban ni narimashita. Taiyou wa kumo no aida kara yasashii kao dashite, atatakana hikari wo okurimashita. Tabibito wa dandan yoi kokoromochi ni natte, shimai ni wa gaitou wo nugimashita. Soko de Kitakaze no make ni narimashita.

Malayo (Malasia, austronesio)

Ketika Angin Utara dan Matahari sedang bertengkar mengenai siapa yang lebih kuat, datang seorang pengembara yang memakai jubah. Keduanya bersetuju bahawa siapa yang berjaya menyebabkan pengembara tersebut menanggalkan jubahnya akan dianggap lebih kuat. Lalu Angin Utara pun meniup sekuatnya, namun semakin kuat angin bertiup semakin rapat pula pengembara tersebut memeluk jubahnya, sehingga akhirnya Angin Utara pun mengalah. Kemudian Matahari memancarkan sinarnya dan dengan segera pengembara tersebut menanggalkan jubahnya. Akhirnya Angin Utara terpaksa mengaku bahawa Matahari lebih kuat daripadanya.

Mandarín (China, sinotibetano)

Yǒu yì huí běifēng gēn tàiyáng zhèngzàinar zhēnglùn shéide běnshi dà, shuōzhe shuōzhe láile yíge zǒudàorde, shēnshang chuānzhe yíjiàn hòu páozi. Tāmen liǎ jiù shāngliang hǎo le shuō, “shéi néng xiān jiào zhège zǒudàorde bǎ tāde páozi tuōle xiàlai a, jiù suàn shéide běnshi dà”. Hǎo, běifēng jiù shǐqi dà jìn lá jǐnguā jǐnguā, kěshì tā guāde yuè lìhai, nèige rén bǎ páozi guǒde yuè jǐn; daà mòliǎor běifēng méile fázi, zhǐhǎo jiù suànle. Yìhuǐr tàiyáng jiù chūlái rèrērde yí shài, nèi zǒudàorde mǎshàng jiù bǎ páozi tuōle xiàlai. Suǒyǐ běifēng bù néng bù chéngrèn dàodǐ háishi tàiyáng bǐ tā běnshi dà.

Mapuche (Chile, araucano)

Piku kürüf engu antü n’otukawmekelu tuchi ñi doy newenngen, ruparumi kiñe wentru makuñtulelu. Fey mew feypiwingu: tuchi nentuñmafile ñi makuñ feytichi wentru fey doy newenngerkey pinngey. Fey mew fütra newentu tripay ti piku kürüf; welu tiyechi wentru fey doy ümpülhuwi ñi makuñ mew. Femlu ti wentru fey ti piku kürüf rupay. Fey mew ti antü wülüfüy ñi alof tripan. Arelu ti wentru, nentuy ñi makuñ. Fey mew ti piku kürüf kimi antü ñi doy newenngen.

Turco (Turquía, túrquico)

Poyrazla günes, birbirlerinden daha kuvvetli olduklarirn ileri sürerek iddialashiyorlardi. Derken, kahn bir palto giymish bir yolcu gördüler. Bu yolcuya paltosunu çikarttirabilenin daha kuvvetli oldugunu kabul etmeye karar verdiler. Poyraz, var gücüyle esmeye bashladi. Ancak, yolcu paltosuna gitgide daha siki sariniyordu. Sonunda poyraz ugrashmaktan vazgeçti. Bu sefer günesh açti; ortalik isininca yolcu paltosunu hemen çikardi. Böylece poyraz, güneshin kendisinden daha kuvvetli oldugunu kabul etmeye mecbur kaldi.

Zapoteco (México, otomangueano)

Ti dxi cacaa yu bi yooxho ne gubidxa. Tu tobi de laaca jma nadipa’. Raqué nuu ca bedandá ti nguiiu renda ti lari ngola. Para uni’ stiidxa ca’, guuya ca’ tu laa gaxha’ lari ladi nguiiu que. Bizulú nda’ bi yooxho bindubi ne stale stipa. Peru laga jma nadipa cundubi la?, jma rusi birenda dxiichi nguiiu que xhaba. Ni bi’ni ti bi la?, biaana dxi. Óraque bizulú gubidxa bizaani ne irá xtuxhu. Óraqueca gulee nguiiu que xhaba, ne zacá biiya’ bi jma nadipa gubidxa que laa.

Apéndice 2: Cocientes lingüísticos

Idioma	Familia	Región	Fon/Sil	Sil/Pal	Pal/Enunc
Abjasio	Caucásica NO	Asia occidental	2,1469	3,1053	7,13
Aingae	Cofán	América	2,0260	2,8657	8,38
Albanés	Indoeuropea	Europa	2,1613	1,6316	16,29
Amárico	Afroasiática	África	2,5521	2,7553	11,75
Ambel	Austronesia	Australasia	2,1162	1,9127	9,69
Apache (Oriental)	Atabascana	América	2,1287	2,3051	7,87
Árabe	Afroasiática	Asia occidental	2,2488	2,5529	9,44
Arrernte	Pamañungana	Australasia	2,2474	2,6575	6,08
Bai	Sinotibetana	Asia oriental	2,0546	1,5913	10,45
Basaá	Nigercongolesa	África	2,2752	1,4057	17,67
Bengalí	Indoeuropea	Asia occidental	2,3299	1,8942	10,40
Bereber (Marruecos)	Afroasiática	África	2,8898	1,8438	9,14
Birmano	Sinotibetana	Asia oriental	2,2901	3,1190	6,00
Chickasaw	Muskogueana	América	2,5761	3,2281	5,70
Coreano	Coreánica	Asia oriental	2,2952	2,7667	8,57
Dinka	Sudánica oriental	África	1,9028	2,1022	13,70
Ende	Pahoturi (Papú)	Australasia	2,1458	2,0140	10,21
Estonio	Urálica	Europa	2,6057	2,0349	9,56
Fataluku	Timoresa (Papú)	Australasia	2,1053	2,1269	9,57
Georgiano	Caucásica Sur	Asia occidental	2,3616	2,5286	7,78
Griego	Indoeuropea	Europa	2,1577	2,1346	11,56
Hausa	Afroasiática	África	2,2979	1,6988	13,83
Hmong (Hmu)	Hmong-Mien	Asia oriental	2,1617	1,0637	15,70
Huehuetla (Tepehua)	Totonaca	América	2,4413	2,8871	8,86
Húngaro	Urálica	Europa	2,2448	1,9200	10,00
Igbo	Nigercongolesa	África	1,7115	1,9439	13,38
Ijo (Kalabari)	Nigercongolesa	África	1,8882	1,8914	14,58
Irlandés	Indoeuropea	Europa	2,2809	1,3798	18,43
Italiano	Indoeuropea	Europa	2,3085	1,7478	12,78
Japonés	Japónica	East Asia	1,9559	2,5506	9,89
Kabiye	Nigercongolesa	África	1,9593	2,4286	10,11
Kazajo	Túrquica	Asia occidental	2,5785	2,4778	11,25
Kera	Afroasiática	África	2,3935	1,5650	9,32
Khongso	Sinotibetana	Asia oriental	2,5170	1,3968	21,00
Kumiai	Yumana	América	2,9024	1,3443	7,63
Kunama	Kunamana	África	2,1337	3,0656	12,20
Lizu	Sinotibetana	East Asia	2,0930	1,9545	13,75
Mah Meri	Austroasiática	East Asia	2,5385	1,6364	10,21
Makasar	Austronesia	Australasia	2,2873	2,4123	14,25
Malayo	Austronesia	Australasia	2,3014	2,6795	9,75
Malgache	Austronesia	África	2,0435	2,1905	12,60
Mambay	Nigercongolesa	África	2,4023	1,3034	10,68
Mandarín	Sinotibetana	Asia oriental	2,6815	1,6020	9,80
Mapuche	Araucana	América	2,3841	2,0133	8,33
Mojeño	Arahuaca	América	2,2065	2,3084	13,38
Mono	Nigercongolesa	África	1,8674	1,5739	11,50
Nara	Sudánica oriental	África	2,3417	1,8426	9,82
Nen	Yam (Papú)	Australasia	2,3021	2,3267	12,63
Paiute	Utoazteca	América	2,1604	2,9444	4,50
Panyabí	Indoeuropea	Asia occidental	2,2644	1,5963	13,63
Persa	Indoeuropea	Asia occidental	2,4897	2,1319	10,11

Idioma	Familia	Región	Fon/Sil	Sil/Pal	Pal/Enunc
Pitjantjatjara	Pamañungana	Australasia	2,1792	2,9444	7,20
Qanjobal	Maya	América	2,3750	2,4615	13,00
Qaqet	Baining (Papú)	Australasia	2,2967	2,1329	9,29
Quichua (Ecuador)	Quechua	América	2,1882	2,8830	8,55
Sahaptin	Penutí	América	2,4351	2,7018	7,13
Sandawe	Joisana	África	2,1044	2,3038	8,78
Seenku	Nigercongolesa	África	2,1078	1,3360	13,89
Seri	Comcaac	América	2,4504	1,5414	14,27
Setswana	Nigercongolesa	África	1,9188	1,6281	15,13
Shawi	Cahuapana	América	2,1312	3,4000	9,29
Shipibo	Panoana	América	1,7905	2,6079	11,95
Sueco	Indoeuropea	Europa	2,5917	1,5794	11,89
Sumi	Sinotibetana	Asia occidental	1,8448	2,6667	12,43
Tailandés	Tai-Kadai	Asia oriental	2,7746	1,3206	11,91
Tamambo	Austronesia	Australasia	2,1200	1,8072	11,86
Tamil	Dravídica	Asia occidental	2,1468	3,1899	8,78
Tausug	Austronesia	Australasia	2,4034	2,0877	9,50
Telugu	Dravídica	Asia occidental	2,1154	3,6000	8,13
Temne	Nigercongolesa	África	2,1546	1,6560	11,36
Tepehuano	Utoazteca	América	2,1326	2,2625	8,89
Tera	Afroasiática	África	2,2390	1,6016	14,22
Tongano	Austronesia	Australasia	1,8566	1,9308	21,67
Triqui (Chicahuaxtla)	Otomangueana	América	2,2814	1,5182	18,33
Turco	Túrquica	Asia occidental	2,3552	2,7727	7,33
Ucraniano	Indoeuropea	Europa	2,5000	2,1667	12,00
Urarina	Urarínica	América	1,9349	2,9912	6,65
Vasco	Vascónica	Europa	2,1444	2,2530	11,86
Vietnamita	Austroasiática	Asia oriental	2,8547	1,0000	16,71
Zapoteco (Istmo)	Otomangueana	América	2,1234	1,7701	9,67
Promedio			**2,2491**	**2,1541**	**11,15**

Apéndice 3: Variables tipológicas exógenas

Idioma	Consonantes	Vocales	Tonos	Acento	Sílabas Complejas	Casos
Abjasio	59	2	1	1	1	2
Aingae	27	10	1	1	0	6
Albanés	29	7	1	0	1	4
Amárico	27	7	1	1	0	2
Ambel	14	5	2	0	0	1
Apache	33	8	3	0	0	1
Árabe	29	6	1	0	1	1
Arrernte	27	4	1	1	0	8
Bai	21	15	5	0	0	1
Basaá	30	14	4	0	0	1
Bengalí	29	7	1	0	1	6
Bereber	34	3	1	0	1	2
Birmano	34	9	4	0	0	8
Chickasaw	16	9	1	1	0	2
Coreano	19	18	1	0	0	6
Dinka	20	7	4	0	0	1
Ende	19	7	1	0	0	6
Estonio	17	18	1	0	1	10

Idioma	Consonantes	Vocales	Tonos	Acento	Sílabas Complejas	Casos
Fataluku	15	5	1	0	0	1
Georgiano	28	5	1	1	1	6
Griego	18	5	1	1	1	3
Hausa	28	10	2	0	0	1
Hmong	32	8	8	0	0	1
Huehuetla	21	6	1	0	1	1
Húngaro	25	14	1	0	1	10
Igbo	26	8	3	0	0	1
Ijo	20	18	2	0	0	1
Irlandés	35	11	1	0	1	2
Italiano	21	7	1	1	1	1
Japonés	16	5	2	1	0	8
Kabiye	21	9	2	0	0	1
Kazajo	20	11	1	0	0	6
Kera	24	6	3	0	1	1
Khongso	26	10	5	0	0	1
Kumiai	17	10	1	0	1	6
Kunama	22	10	3	0	0	6
Lizu	39	8	2	0	0	1
Mah Meri	30	19	2	0	0	3
Makasar	19	5	1	0	0	1
Malayo	18	6	1	0	1	1
Malgache	29	4	1	0	0	1
Mambay	25	10	2	0	0	1
Mandarín	19	6	4	0	0	1
Mapuche	22	6	1	0	0	2
Mojeño	29	12	1	0	1	1
Mono	32	8	3	0	0	1
Nara	25	10	2	0	0	5
Nen	18	8	1	1	1	3
Paiute	17	11	1	0	0	5
Panyabí	27	17	3	0	1	2
Persa	23	6	1	1	1	2
Pitjantjatjara	17	6	1	0	0	10
Qanjobal	25	5	1	0	0	1
Qaqet	16	4	1	0	0	1
Quichua	23	3	1	0	0	8
Sahaptin	32	7	1	1	1	4
Sandawe	44	15	2	0	0	1
Seenku	20	12	4	0	0	1
Seri	18	8	1	0	1	1
Setswana	28	7	2	0	0	1
Shawi	12	4	1	1	0	6
Shipibo	15	8	1	1	0	6
Sueco	16	17	1	1	1	2
Sumi	29	6	3	0	0	6
Tailandés	21	9	5	0	0	1
Tamambo	16	5	1	0	0	1
Tamil	15	10	1	0	0	6
Tausug	17	3	1	0	0	1
Telugu	35	12	1	0	0	8
Temne	19	9	2	0	1	1
Tepehuano	12	5	2	0	0	1

Idioma	Consonantes	Vocales	Tonos	Acento	Sílabas Complejas	Casos
Tera	35	11	3	0	0	1
Tongano	12	5	1	1	0	1
Triqui	22	8	5	0	0	1
Turco	22	8	1	0	0	6
Ucraniano	32	6	1	1	1	7
Urarina	13	13	2	0	0	1
Vasco	23	5	1	1	1	10
Vietnamita	22	11	8	0	0	1
Zapoteco	20	5	3	0	0	1
Promedio	**23,89**	**8,50**	**1,96**	**23%**	**31%**	**3,14**

Apéndice 4: Variables tipológicas exógenas (continuación)

Idioma	Géneros	Inflexiones	Objeto-Verbo	Adjet-Sustant	Acusativo	Aislante
Abjasio	3	10	1	0	0	0
Aingae	1	6	1	1	1	0
Albanés	3	7	0	0	1	0
Amárico	2	6	1	1	1	0
Ambel	1	6	0	0	1	1
Apache	1	5	1	0	1	0
Árabe	2	6	0	0	1	0
Arrernte	1	4	1	0	0	0
Bai	1	2	0	1	0	1
Basaá	5	6	0	0	0	0
Bengalí	2	2	1	1	1	0
Bereber	2	6	0	0	1	0
Birmano	1	2	1	0	0	1
Chickasaw	1	6	1	0	0	0
Coreano	1	6	1	1	0	0
Dinka	1	6	1	0	1	1
Ende	1	6	1	1	0	0
Estonio	1	2	0	1	1	0
Fataluku	1	4	1	0	0	1
Georgiano	1	8	1	1	1	0
Griego	3	4	0	1	1	0
Hausa	2	6	0	1	0	1
Hmong	1	2	0	0	0	1
Huehuetla	1	4	0	1	1	0
Húngaro	1	4	0	1	1	0
Igbo	1	6	0	0	0	1
Ijo	1	6	1	1	1	1
Irlandés	2	2	0	0	1	0
Italiano	2	4	0	0	1	0
Japonés	1	4	1	1	0	0
Kabiye	1	2	0	0	0	1
Kazajo	1	6	1	1	1	0
Kera	2	6	0	0	1	1
Khongso	1	1	1	1	0	1
Kumiai	1	6	1	0	1	0
Kunama	2	4	1	0	1	0
Lizu	1	3	1	0	0	1
Mah Meri	1	1	0	0	0	1

Idioma	Géneros	Inflexiones	Objeto-Verbo	Adjet-Sustant	Acusativo	Aislante
Makasar	1	5	0	0	0	1
Malayo	1	4	0	0	1	1
Malgache	1	4	0	0	1	0
Mambay	1	1	0	0	0	0
Mandarín	1	1	0	1	0	1
Mapuche	1	8	0	1	0	0
Mojeño	1	6	0	0	0	0
Mono	5	6	1	0	0	0
Nara	2	4	1	0	1	1
Nen	1	10	1	0	0	0
Paiute	1	4	1	1	1	0
Panyabí	2	3	1	1	1	0
Persa	1	4	1	0	1	0
Pitjantjatjara	1	4	1	0	1	0
Qanjobal	1	4	0	1	0	0
Qaqet	8	3	0	0	0	0
Quichua	1	8	1	1	1	0
Sahaptin	1	10	0	1	0	0
Sandawe	5	8	1	0	1	1
Seenku	1	2	1	0	0	1
Seri	1	5	1	1	0	0
Setswana	5	4	0	0	1	0
Shawi	1	6	1	1	0	0
Shipibo	1	6	1	1	0	0
Sueco	3	2	0	1	1	0
Sumi	1	4	1	0	1	0
Tailandés	1	2	0	0	0	1
Tamambo	1	6	0	0	1	1
Tamil	3	2	1	1	1	0
Tausug	1	4	0	0	0	1
Telugu	3	2	1	1	1	0
Temne	5	2	0	0	1	0
Tepehuano	1	4	0	1	1	0
Tera	1	2	0	0	0	1
Tongano	1	6	0	0	0	1
Triqui	1	6	0	0	1	0
Turco	1	6	1	1	1	0
Ucraniano	3	4	0	1	1	0
Urarina	1	8	1	0	0	0
Vasco	1	4	1	0	0	0
Vietnamita	1	1	0	0	0	1
Zapoteco	1	8	0	0	1	0
Promedio	**1,65**	**4,63**	**49%**	**40%**	**53%**	**33%**