UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

LA EXISTENCIA DE CORRELACIÓN NEGATIVA ENTRE DISTINTOS ASPECTOS DE LA COMPLEJIDAD DE LOS IDIOMAS

Germán Coloma

Abril 2014
Nro. 536

www.cema.edu.ar/publicaciones/doc_trabajo.html
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 <jae@cema.edu.ar>

LA EXISTENCIA DE CORRELACIÓN NEGATIVA ENTRE DISTINTOS ASPECTOS DE LA COMPLEJIDAD DE LOS IDIOMAS

Germán Coloma (Universidad del CEMA, Buenos Aires, Argentina)[*]

Resumen

Este trabajo propone un procedimiento para evaluar la posible existencia de correlación negativa entre complejidad fonológica, morfológica y sintáctica de diferentes idiomas. El procedimiento se basa en el uso de coeficientes de correlación parcial, variables instrumentales y sistemas de ecuaciones simultáneas, y se lleva a cabo sobre datos tomados de la fábula "El viento norte y el sol", traducida a 40 idiomas distintos. La complejidad fonológica se mide por el número de fonemas por sílaba, la complejidad morfológica por el número de sílabas por palabra, y la complejidad sintáctica por el número de palabras por frase. Luego de controlar por fenómenos relacionados con factores geográficos y endogeneidad estadística, se llega a la conclusión de que la complejidad fonológica está negativamente correlacionada con la complejidad morfológica, y que esta, a su vez, está negativamente correlacionada con la complejidad sintáctica. Estos resultados pueden relacionarse con la llamada "ley de Menzerath".

Palabras clave: correlación parcial, variables instrumentales, ecuaciones simultáneas, complejidad lingüística, ley de Menzerath.

1. Introducción

El objetivo de este trabajo es proponer un procedimiento para evaluar la posible existencia de correlación negativa entre complejidad fonológica, morfológica y sintáctica de diferentes idiomas. El mismo se basa en el uso de un método de regresión de ecuaciones simultáneas, que es una técnica estadística ampliamente utilizada en otras ciencias sociales.

A fin de evaluar las relaciones entre las distintas medidas de complejidad en un contexto de análisis multilingüístico, utilizaremos un mismo texto disponible en muchas lenguas. Dicho texto es la fábula conocida como "El viento norte y el sol", que es un relato breve empleado por la Asociación Fonética Internacional (IPA) para ilustrar la fonética de los diferentes idiomas. Contando con dicho texto para un número considerable de lenguas de todo el mundo, hemos elegido una muestra de 40 casos para

[*] Agradezco los comentarios de Damián Blasi a una versión anterior de este trabajo. Agradezco también a Helen Eaton, Sameer Kahn y Justin Watkins por facilitarme el acceso a algunas de las fuentes utilizadas. Las opiniones son personales y no representan necesariamente las de la Universidad del CEMA.

los cuales hemos calculado tres índices básicos: fonemas por sílaba, sílabas por palabra y palabras por frase. Dichos índices pueden ser considerados como medidas de complejidad fonológica, morfológica y sintáctica para cada uno de los idiomas en cuestión.

La forma más simple de analizar la existencia de relaciones entre los indicadores mencionados es calcular los coeficientes de correlación simple (o de Pearson) entre ellos. Algo más complejo resulta calcular los denominados "coeficientes de correlación parcial", que tienen la ventaja de que utilizan información de las tres variables al mismo tiempo. Una forma alternativa para capturar las relaciones entre los índices de complejidad es correr regresiones en las que cada una de ellos depende de otras variables (por ejemplo, de las otras medidas de complejidad, y de ciertas variables categóricas relacionadas con la filiación de los idiomas, o con factores geográficos). Esas regresiones pueden correrse de manera simultánea, como un sistema de ecuaciones. Todos los resultados obtenidos tanto de los análisis de correlación como de los análisis de regresión son luego comparados entre sí, y relacionados también con resultados anteriores que aparecen en la literatura sobre tipología lingüística.

El resto de este artículo está organizado de la siguiente manera. En la sección 2 hacemos una reseña de la literatura sobre análisis multilingüísticos de complejidad, especialmente de aquella que encuentra resultados relacionados con posibles correlaciones entre variables fonológicas y gramaticales. En la sección 3 explicamos las principales características de la base de datos que hemos armado, basada en las transcripciones de "El viento norte y el sol" que aparecen en IPA (1999) y en otras fuentes relacionadas. En la sección 4 aplicamos técnicas de correlación y regresión a nuestra base de datos, y hallamos algunos resultados que pueden ser comparados con los de la literatura existente. Finalmente, en la sección 5 exponemos algunas conclusiones.

2. Reseña de la literatura

La literatura sobre análisis cuantitativo de la complejidad de los idiomas es relativamente amplia y diversa. Probablemente el enfoque más parecido al que utilizamos aquí es el que aparece en Shosted (2006), quien analiza la posible existencia de correlación negativa entre complejidad fonológica y morfológica en una muestra de 32 lenguas. En dicho trabajo, que no encuentra una correlación estadísticamente significativa

entre ambas variables, el autor mide la complejidad fonológica usando el número de tipos posibles de sílaba en cada idioma, en tanto que su medida de complejidad morfológica está dada por el número de categorías de la inflexión verbal.

Un enfoque diferente para medir la complejidad lingüística, que genera también una conclusión distinta en lo que se refiere a la existencia de correlación negativa, es el de Fenk-Oczlon y Fenk (2004). Estos autores calculan varios índices lingüísticos (fonemas por sílaba, sílabas por palabra, palabras por frase, etc.) usando como base traducciones a 33 idiomas de distintas oraciones simples escritas originalmente en alemán. Luego calculan varios coeficientes de correlación entre los índices propuestos, e incluyen algunas otras variables (por ejemplo, número de casos de los sustantivos, tendencia a la aparición de "posposiciones", orden habitual del objeto y el verbo) que se supone que tienen alguna relación con ciertas dimensiones de la complejidad lingüística. Llegan así a un conjunto de conclusiones que implican la presencia de correlación negativa entre varias de las variables analizadas, y la mayoría de ellas resultan ser estadísticamente significativas.

Los resultados de Fenk-Oczlon y Fenk también pueden interpretarse como un caso especial de la llamada "ley de Menzerath" (Menzerath, 1954), que predice una correlación negativa entre la medida de cada elemento lingüístico y la medida de los componentes de dicho elemento. En este caso, eso implica que el número de fonemas por sílaba debe estar negativamente correlacionado con el número de sílabas por palabra (ya que los fonemas son los componentes de cada sílaba) y que el número de sílabas por palabra debe estar negativamente correlacionado con el número de palabras por frase (ya que las sílabas son los componentes de cada palabra).

Otra rama de la literatura, que también genera resultados comparables a los de Fenk-Oczlon y Fenk, es la que relaciona la extensión de las palabras con medidas de complejidad fonológica. Dicha literatura comienza con el trabajo de Nettle (1995), que encuentra una fuerte correlación negativa entre el número de fonemas de un idioma y el número promedio de fonemas por palabra en dicho idioma, y continúa posteriormente con trabajos como el de Wichmann et al. (2011), que utiliza una medida alternativa de complejidad fonológica y una base de datos mucho más grande (al tiempo que controla por efectos relacionados con la geografía y con la filiación lingüística). Otra contribución

que puede mencionarse aquí es la de Oh et al. (2013), que calcula correlaciones entre sílabas por palabra y dos medidas de "densidad informativa" (que son los índices entre el número de sílabas y el número de palabras en el mismo texto traducido a distintos idiomas, tomando como base los valores obtenidos para el idioma vietnamita). Estos autores encuentran una correlación negativa altamente significativa entre densidad silábica (que puede verse como una medida de complejidad fonológica) y número promedio de sílabas por palabra.

Otros enfoques para el problema de la medición de la complejidad de los idiomas se circunscriben a subsistemas particulares del lenguaje, como pueden ser la fonología, la morfología o la sintaxis. Maddieson (2007), por ejemplo, mide la complejidad fonológica usando tres indicadores diferentes: número de fonemas (vocálicos y consonánticos), sistema de tonos, y complejidad del inventario de sílabas. Dicho autor intenta también encontrar correlaciones entre tales medidas, pero solo encuentra una correlación positiva significativa entre el número de consonantes y la estructura silábica, y una correlación negativa (mucho menos significativa) entre estructura silábica y complejidad del sistema de tonos. Moran y Blasi (2014), en cambio, hallan una correlación negativa relativamente grande entre el número de fonemas vocálicos y el número promedio de fonemas por palabra, al tiempo que proponen una medida alternativa de complejidad fonológica, basada en el mínimo conjunto de rasgos distintivos necesarios para describir los fonemas de cada idioma.

En lo que se refiere a la complejidad morfológica, Bane (2008) propone una medida que consiste en analizar el uso de afijos en las palabras que aparecen en un texto, y a fin de llevar a cabo su idea en un contexto multilingüístico utiliza distintas traducciones de la Biblia. También compara la medida propuesta con otras que provienen de contar palabras, hallando una fuerte correlación positiva entre su medida y las que surgen de esos otros criterios alternativos.

El recuento de palabras ha sido también propuesto como un modo de medir la complejidad sintáctica. Szmrecsányi (2004), por ejemplo, encuentra que el número de palabras por frase está altamente correlacionado con otras medidas tales como el número de nodos por frase y con el "índice de complejidad sintáctica". Dicho número de palabras, sin embargo, tiene la ventaja de que es más simple de computar, especialmente

cuando se trabaja con un número relativamente grande de idiomas diferentes.

Teniendo en cuenta esa observación, en este trabajo usaremos el número de palabras por frase como nuestra medida básica de complejidad sintáctica. Del mismo modo, nuestras medidas de complejidad fonológica y morfológica serán respectivamente el número de fonemas por sílaba y el número de sílabas por palabra, y todos esos índices serán sometidos a un análisis estadístico para determinar la posible existencia de correlación negativa entre ellos. Otra característica importante de este artículo tiene que ver con su nivel de sofisticación estadística. Al revés de los estudios previos que se limitan al empleo de coeficientes de correlación simple, aquí utilizaremos también coeficientes de correlación parcial calculados de distintas maneras. Con dicho enfoque, controlaremos por factores relacionados con la interacción de varias variables adicionales y con la endogeneidad de los indicadores de complejidad propuestos, en el contexto de una estimación de ecuaciones simultáneas.

3. El viento norte y el sol

La fábula del viento norte y el sol, atribuida a Esopo, es un texto que ha sido usado durante muchas décadas por la Asociación Fonética Internacional como un "espécimen" o modelo para ilustrar los sonidos de los idiomas, y también los símbolos fonéticos necesarios para describir dichos sonidos[1]. Es por lo tanto un ejemplo único de un texto breve para el cual numerosos especialistas en la fonética de distintos idiomas han analizado meticulosamente los sonidos, los fonemas, las sílabas, las palabras y las frases de las lenguas y los dialectos bajo estudio.

La versión española de "El viento norte y el sol" es la siguiente:

El viento norte y el sol porfiaban sobre cuál de ellos era el más fuerte, cuando acertó a pasar un viajero envuelto en ancha capa. 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 arrebujaba en su capa el viajero; por fin el viento norte abandonó la empresa. Entonces brilló el sol con ardor, e inmediatamente se despojó de su capa el viajero; por lo que el viento norte hubo de reconocer la superioridad del sol.

Si contamos el número de frases, palabras, sílabas y fonemas en este texto,

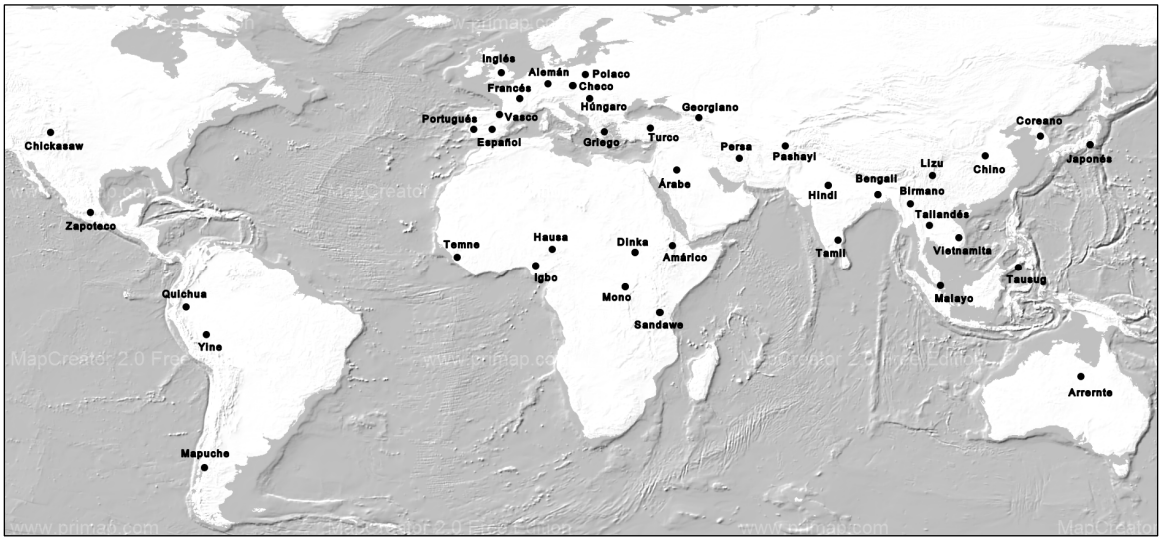
[1] Véase, por ejemplo, IPA (1949).

hallaremos que el mismo está compuesto por 9 frases o enunciados, 107 palabras, 193 sílabas y 425 fonemas. Esto nos permite calcular distintos índices surgidos como cocientes de las cifras en cuestión, entre los cuales los principales son el número de fonemas por sílaba (igual a 2,2021), el número de sílabas por palabra (igual a 1,8037) y el número de palabras por frase (igual a 11,89). Este mismo ejercicio puede llevarse a cabo para los textos disponibles en muchos otros idiomas. En la versión inglesa de "El viento norte y el sol", por ejemplo, el número de fonemas por sílaba es mayor (igual a 2,6783) pero el número de sílabas por palabra es menor (igual a 1,2655). En la versión traducida a la lengua árabe, en cambio, el número de sílabas por palabra (igual a 2,5529) es bastante mayor que el calculado para las versiones española e inglesa, en tanto que lo que es considerablemente menor es el número de palabras por frase (igual a 9,44).

La comparación sistemática de los cocientes entre fonemas y sílabas, sílabas y palabras, y palabras y frases para distintas traducciones de "El viento norte y el sol" puede ser por lo tanto una fuente muy útil para medir la complejidad relativa de los distintos idiomas, y para capturar la posible existencia de relaciones entre los indicadores calculados. A fin de llevar a cabo dicho análisis, seleccionamos una muestra de 40 lenguas para las cuales contamos con versiones del texto mencionado, que se encuentran en el *Handbook of the International Phonetic Association* (IPA, 1999) o en la serie "Illustrations of the IPA" (ilustraciones del alfabeto fonético internacional) publicada por el *Journal of the International Phonetic Association*. La muestra incluye 16 idiomas con un número muy grande de hablantes a nivel mundial (chino, inglés, español, hindi, árabe, portugués, bengalí, japonés, alemán, francés, tamil, vietnamita, coreano, malayo, persa y turco)[2], y 24 idiomas más (amárico, arrernte, birmano, checo, chickasaw, dinka, georgiano, griego, hausa, húngaro, igbo, lizu, mapuche, mono, pashayi, polaco, quichua, sandawe, tailandés, tausug, temne, vasco, yine y zapoteco) que son representantes de varias familias y áreas lingüísticas importantes. La ubicación geográfica aproximada de todos esos idiomas es la que aparece en el mapa del gráfico 1.

[2] Otros idiomas con un gran número de hablantes, tales como ruso, javanés, panyabí, maratí y telugu, no han sido incluidos por no disponerse para ellos de una ilustración del IPA.

Gráfico 1: Ubicación de los idiomas incluidos en la muestra



La descripción de nuestra base de datos en términos de fonemas por sílaba, sílabas por palabra, y palabras por frase aparece en el cuadro 1. En promedio, el número de fonemas por sílaba es 2,2957, pero dicho indicador va desde un mínimo de 1,7115 (correspondiente al idioma igbo, que es una lengua de la familia Níger-Congo que se habla en Nigeria) hasta un máximo de 2,8547 (correspondiente a la lengua vietnamita)[3]. El vietnamita es también el idioma que exhibe un menor valor para el cociente entre sílabas y palabras (igual a 1, ya que en la versión vietnamita de "El viento norte y el sol" todas las palabras son monosílabos), en tanto que el mayor valor para ese indicador (igual a 3,7460) corresponde al idioma yine (que es una lengua arahuaca, hablada en la selva amazónica peruana), en un contexto en el cual el promedio de sílabas por palabra es 2,1563. El promedio de palabras por frase, finalmente, es igual a 10,45, en tanto que el mínimo valor para dicho índice (que corresponde al chickasaw, lengua muskoguí hablada en Estados Unidos) es de 5,70, y el máximo (que corresponde al vietnamita) es de 16,71.

[3] Para definir el número de fonemas en cada versión de "El viento norte y el sol", tratamos de seguir los criterios usados por los autores que escribieron las correspondientes ilustraciones del IPA, pero también aplicamos algunos criterios unificadores. Por ejemplo, las vocales largas, cortas, orales y nasales fueron consideradas como fonemas diferentes cuando la duración o la nasalización eran rasgos distintivos en un idioma, pero los diptongos fueron considerados en todos los casos como combinaciones de dos fonemas. Las consonantes africadas y demás "articulaciones dobles", por su parte, también fueron consideradas como fonemas separados cuando eso resultaba aconsejable, pero las "consonantes geminadas" fueron siempre computadas como combinaciones de dos fonemas (idénticos) consecutivos.

Cuadro 1: Principales indicadores de la base de datos de "El viento norte y el sol"

Idioma	Filiación	Fonemas/sílabas	Sílabas/palabras	Palabras/frases
Alemán	Germánica	2,5333	1,6667	10,80
Amárico	Semítica	2,5521	2,7553	11,75
Árabe	Semítica	2,2488	2,5529	9,44
Arrernte	Pama-Nyungan	2,2474	2,6575	6,08
Bengalí	Indoaria	2,3299	1,8942	10,40
Birmano	Sinotibetana	2,2901	3,1190	6,00
Checo	Eslava	2,4321	1,9286	8,40
Chickasaw	Muskoguí	2,5761	3,2281	5,70
Chino	Sinotibetana	2,6815	1,6020	9,80
Coreano	Coreánica	2,2952	2,7667	8,57
Dinka	Nilótica	1,9028	2,1022	13,70
Español	Latina	2,2021	1,8037	11,89
Francés	Latina	2,1572	1,4722	12,00
Georgiano	Caucásica	2,3616	2,5286	7,78
Griego	Helénica	2,1577	2,1346	11,56
Hausa	Chádica	2,2979	1,6988	13,83
Hindi	Indoaria	2,2452	1,6640	15,63
Húngaro	Urálica	2,2448	1,9200	10,00
Igbo	Níger-Congo	1,7115	1,9439	13,38
Inglés	Germánica	2,6783	1,2655	12,56
Japonés	Japónica	1,9559	2,5506	9,89
Lizu	Sinotibetana	2,0930	1,9545	13,75
Malayo	Austronesia	2,3014	2,6795	9,75
Mapuche	Araucana	2,3841	2,0133	8,33
Mono	Níger-Congo	1,8674	1,5739	11,50
Pashayi	Indoaria	2,2203	1,9888	12,71
Persa	Iraní	2,4897	2,1319	10,11
Polaco	Eslava	2,7089	1,7753	9,89
Portugués	Latina	2,0430	1,8980	12,25
Quichua	Quechua	2,1882	2,8830	8,55
Sandawe	Khoisan	2,1044	2,3038	8,78
Tailandés	Tai-Kadai	2,7746	1,3206	11,91
Tamil	Dravídica	2,1468	3,1899	8,78
Tausug	Austronesia	2,4034	2,0877	9,50
Temne	Níger-Congo	2,1546	1,6560	11,36
Turco	Túrquica	2,3552	2,7727	7,33
Vasco	Vascuence	2,1444	2,2530	11,86
Vietnamita	Austroasiática	2,8547	1,0000	16,71
Yine	Arahuaca	2,3686	3,7460	6,30
Zapoteco	Otomangueana	2,1234	1,7701	9,67
Promedio		**2,2957**	**2,1563**	**10,45**

La descripción de las variables del cuadro 1 puede complementarse con el cálculo de coeficientes de correlación. Ellos son los que aparecen en el cuadro 2, en el cual vemos que la mayor correlación en valor absoluto es la que corresponde a sílabas por palabra versus palabras por frase, la cual es igual a -0,7257. La correlación entre fonemas

por sílaba y sílabas por palabra, en cambio, es considerablemente menor (r = -0,1630), en tanto que la correlación entre fonemas por sílaba y palabras por frase (r = -0,0855) resulta ser menor aún.

Cuadro 2: Coeficientes de correlación de la base de datos utilizada

Variable	Fonemas/sílabas	Sílabas/palabras	Palabras/frases
Fonemas por sílaba	1,0000	-0,1630	-0,0855
Sílabas por palabra		1,0000	-0,7257
Palabras por frase			1,0000

Los valores absolutos de los coeficientes de correlación que aparecen en el cuadro 2 también pueden relacionarse con sus respectivos niveles de significación estadística. Usando estadísticos-t para los diferentes coeficientes, podemos ver que solo el que tiene un valor absoluto mayor (la correlación entre sílabas por palabra y palabras por frase) es significativamente distinto de cero a un nivel de probabilidad del 1% ($p \approx 0$), en tanto que el coeficiente de correlación entre fonemas por sílaba y sílabas por palabra no llega a ser estadísticamente significativo a un nivel de probabilidad del 10% ($p = 0,1575$). El coeficiente de correlación entre fonemas por sílaba y palabras por frase, por último, no es estadísticamente significativo a ningún nivel razonable de probabilidad ($p = 0,2999$).

Otra forma de apreciar las relaciones entre las distintas medidas de complejidad que hemos calculado para nuestra muestra de idiomas es representar cada observación como un punto de un diagrama en el cual computamos dos de esas medidas al mismo tiempo. Eso es lo que hemos hecho en los siguientes gráficos, que corresponden a complejidad fonológica versus complejidad morfológica (gráfico 2), complejidad morfológica versus complejidad sintáctica (gráfico 3) y complejidad fonológica versus complejidad sintáctica (gráfico 4).

Gráfico 2: Complejidad fonológica versus complejidad morfológica

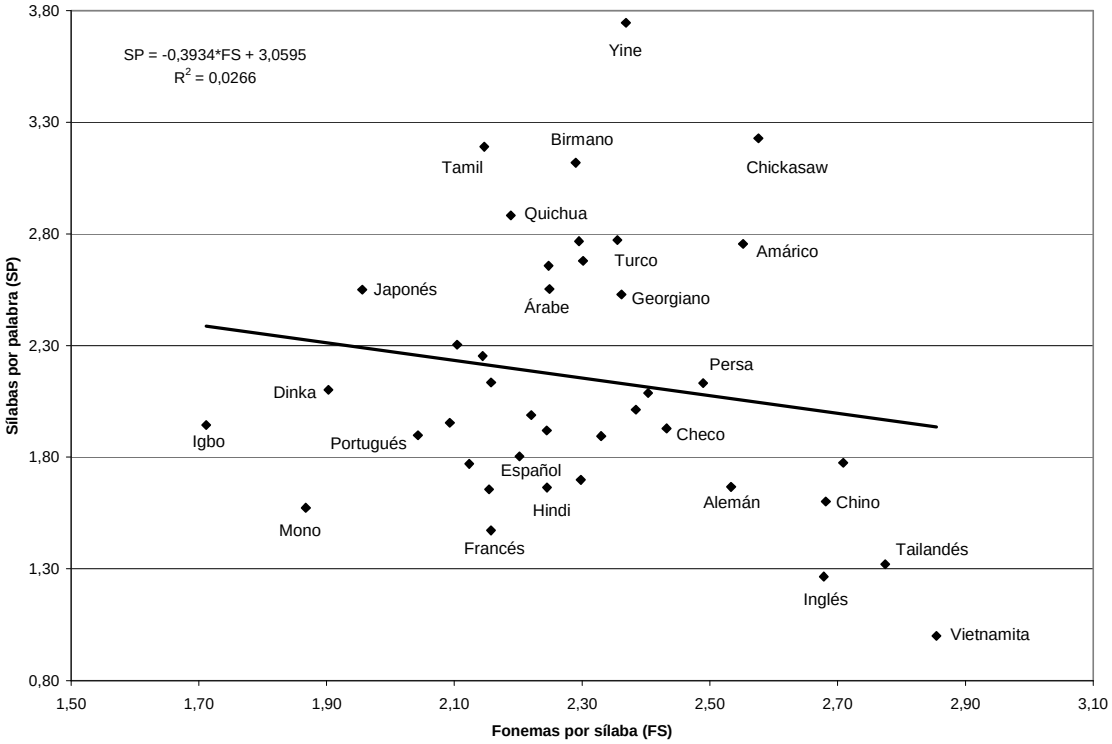


Gráfico 3: Complejidad morfológica versus complejidad sintáctica

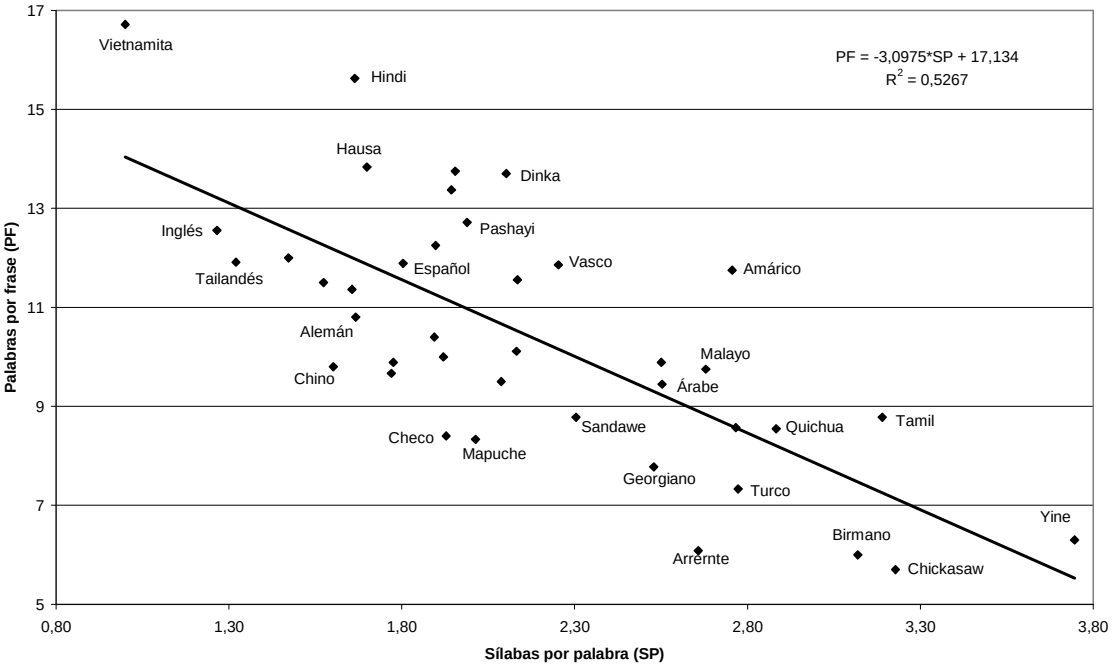
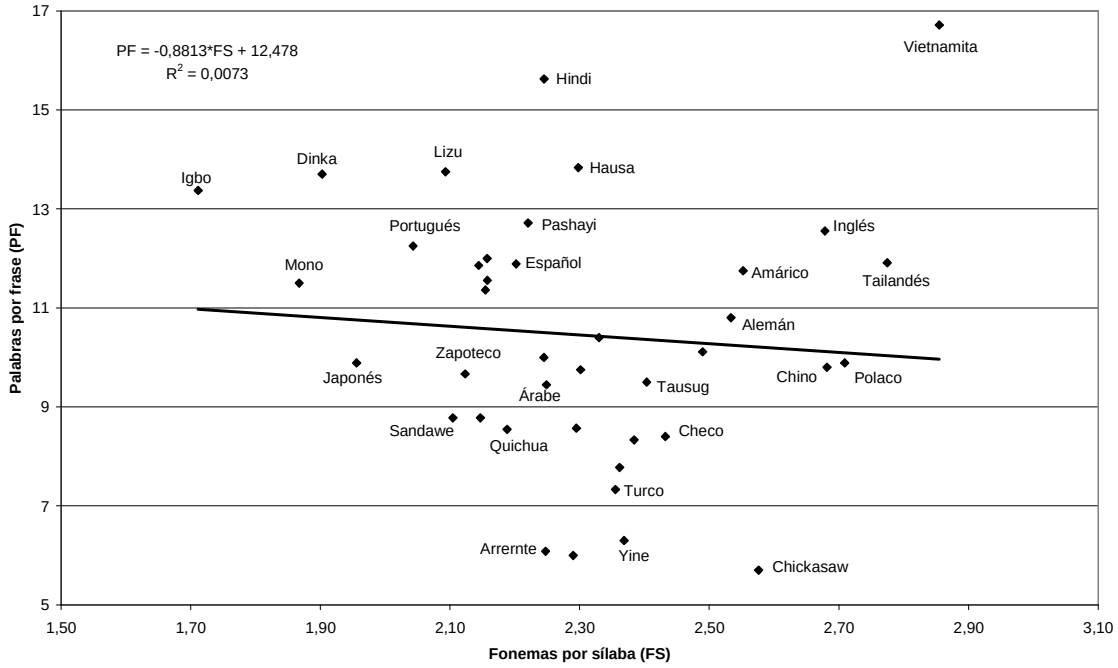


Gráfico 4: Complejidad fonológica versus complejidad sintáctica



Tal como puede observarse, los mismos hechos estilizados detectados por los coeficientes de correlación del cuadro 2 aparecen también en los gráficos 2, 3 y 4. Nótese que las líneas de tendencia generadas por tales gráficos nos indican que, aunque todas las relaciones son negativas, la que corresponde a sílabas por palabra versus palabras por frase (gráfico 3) es mucho más fuerte que la que se da entre fonemas por sílaba y sílabas por palabra (gráfico 2), la cual es a su vez más fuerte que la se encuentra entre fonemas por sílaba y palabras por frase (gráfico 4). Dichas líneas de tendencia son en rigor el resultado de regresiones simples entre las variables representadas en los gráficos, las cuales exhiben además cierta "bondad de ajuste" (medida a través del coeficiente de determinación R^2). Dicho coeficiente es considerablemente más alto en la relación representada en el gráfico 3 ($R^2 = 0,5267$) que en las relaciones representadas en el gráfico 2 ($R^2 = 0,0266$) y en el gráfico 4 ($R^2 = 0,0073$)[4].

[4] Los valores de los coeficientes R^2 en estas regresiones simples están de hecho relacionados con los coeficientes de correlación anteriormente calculados. Efectivamente, cada uno de ellos es en realidad el cuadrado del coeficiente de correlación que corresponde a cada par de variables.

4. Análisis estadístico

4.1. Coeficientes de correlación parcial

Todos los estadísticos descriptivos calculados en la sección anterior son valores promedio de los indicadores lingüísticos de nuestra base de datos, o bien son coeficientes que representan relaciones entre esos indicadores, pero en ningún caso tienen en cuenta la posible interdependencia entre las variables analizadas. Un paso adicional en el análisis de nuestros datos consiste por lo tanto en considerar dicha interdependencia, y el modo más directo de hacerlo es calculando los denominados "coeficientes de correlación parcial".

Un coeficiente de correlación parcial es una medida de la dependencia lineal entre un par de variables, eliminando la influencia que puedan tener otras variables adicionales. Para calcular dicho coeficiente es necesario controlar por el posible efecto de otras variables sobre las dos variables que se quiere correlacionar, eliminándolo a través de algún procedimiento estadístico. Una posibilidad es partir de una matriz de correlación simple de todas las variables bajo análisis (en nuestro caso, de la que aparece en el cuadro 2) e invertirla. Una vez hecho eso, los coeficientes de correlación parcial pueden calcularse empleando la siguiente fórmula:

$$\mathrm{Pc}(x, y) = -\frac{p_{xy}}{\sqrt{p_{xx} \cdot p_{yy}}} \qquad (1)\ ;$$

donde *Pc(x,y)* es el coeficiente de correlación parcial entre las variables *x* e *y*, p_{xy} es el coeficiente que corresponde a dicho par de variables en la matriz de correlación inversa, y p_{xx} y p_{yy} son los coeficientes que corresponden a las variables *x* e *y* en dicha matriz.

Aplicando este procedimiento para todos los coeficientes de correlación parcial generados por las relaciones entre fonemas por sílaba, sílabas por palabra y palabras por frase, llegamos a una matriz como la que aparece en el cuadro 3. Tal como puede verse, todos los coeficientes calculados resultan negativos, y sus valores absolutos son mayores que los coeficientes de correlación que aparecen en el cuadro 2. Este incremento es particularmente evidente para los coeficientes de correlación entre fonemas por sílaba y sílabas por palabra (que sube de -0,1630 a -0,3283) y entre fonemas por sílaba y palabras por frase (que aumenta de -0,0855 a -0,3002). Ambos coeficientes se vuelven ahora

estadísticamente significativos a un nivel de probabilidad del 5%, ya que el primero de ellos toma un valor-p igual a 0,0193, en tanto que el segundo pasa a tener un valor-p igual a 0,0299.

Cuadro 3: Coeficientes de correlación parcial

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Frases
Fonemas por sílaba	1,0000	-0,3283	-0,3002
Sílabas por palabra		1,0000	-0,7525
Palabras por frase			1,0000

Los mismos coeficientes de correlación parcial, obtenidos a través del procedimiento de inversión matricial descripto, pueden hallarse utilizando un procedimiento de regresión estadística. El mismo consiste en correr tres regresiones separadas por mínimos cuadrados ordinarios, correspondientes a estas funciones:

$$FS = c(1) + c(2)*SP + c(3)*PF \quad (2)\ ;$$

$$SP = c(4) + c(5)*FS + c(6)*PF \quad (3)\ ;$$

$$PF = c(7) + c(8)*FS + c(9)*SP \quad (4)\ ;$$

donde *FS* (fonemas por sílaba), *SP* (sílabas por palabra) y *PF* (palabras por frase) son las medidas de complejidad lingüística, 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 en el análisis de regresión[5]. Los resultados de dicho análisis aparecen en el cuadro 4.

Los resultados que corresponden a cada uno de los parámetros estimados en nuestro análisis de regresión representan, respectivamente, un estimador puntual para cada coeficiente (primera columna) junto con su correspondiente error estándar (segunda columna). Con dichos valores, el procedimiento de regresión calcula a su vez un conjunto de estadísticos-t (tercera columna), que tienen una distribución que determina cierto valor de probabilidad de que el parámetro estimado sea en realidad igual a cero. Dichos valores de probabilidad son los que aparecen en la cuarta columna del cuadro 4.

[5] Estas regresiones, al igual que las demás que aparecen en el presente trabajo, fueron llevadas a cabo utilizando el programa informático EViews 3.5.

Cuadro 4: Resultados de las regresiones para calcular correlaciones parciales

Concepto	Coeficiente	Error std.	Estadístico-t	Probabilidad
Ec. de fonemas por sílaba				
Constante [c(1)]	3,157253	0,400474	7,883793	0,0000
Sílabas/Palabras [c(2)]	-0,196940	0,093152	-2,114188	0,0367
Palabras/Frases [c(3)]	-0,041789	0,021826	-1,914664	0,0581
R-cuadrado	0,114311			
Ec. de sílabas por palabra				
Constante [c(4)]	5,238005	0,672946	7,783692	0,0000
Fonemas/Sílabas [c(5)]	-0,547295	0,258868	-2,114188	0,0367
Palabras/Frases [c(6)]	-0,174585	0,025123	-6,949184	0,0000
R-cuadrado	0,577713			
Ec. de palabras por frase				
Constante [c(7)]	22,40026	2,936905	7,627166	0,0000
Fonemas/Sílabas [c(8)]	-2,157226	1,126687	-1,914664	0,0581
Sílabas/Palabras [c(9)]	-3,243070	0,466684	-6,949184	0,0000
R-cuadrado	0,569365			

Con los coeficientes de regresión estimados, las correlaciones parciales entre las distintas medidas de complejidad de los idiomas pueden calcularse utilizando las siguientes fórmulas:

$$\mathrm{Pc(FS,SP)} = \mathrm{c(2)} \cdot \sqrt{\frac{\mathrm{c(5)}}{\mathrm{c(2)}}} = -0,196940 \cdot \sqrt{\frac{-0,547295}{-0,196940}} = -0,3283 \qquad (5)\ ;$$

$$\mathrm{Pc(SP,PF)} = \mathrm{c(6)} \cdot \sqrt{\frac{\mathrm{c(9)}}{\mathrm{c(6)}}} = -0,174585 \cdot \sqrt{\frac{-3,243070}{-0,174585}} = -0,7525 \qquad (6)\ ;$$

$$\mathrm{Pc(FS,PF)} = \mathrm{c(3)} \cdot \sqrt{\frac{\mathrm{c(8)}}{\mathrm{c(3)}}} = -0,041789 \cdot \sqrt{\frac{-2,157226}{-0,041789}} = -0,3002 \qquad (7)\ .$$

Nótese que los valores obtenidos con este procedimiento son idénticos a los que aparecen en el cuadro 3.

4.2. Factores geográficos y regresiones con ecuaciones simultáneas

Los coeficientes de las regresiones llevadas a cabo en la sección anterior son el resultado de un análisis que supone que cada indicador de complejidad lingüística depende de los otros indicadores bajo estudio. En esta sección extenderemos dicho análisis para considerar la posibilidad de que tales indicadores dependan también de otras variables.

Supongamos, por ejemplo, que la complejidad fonológica es un fenómeno que depende también de la filiación de los idiomas, o de factores de tipo geográfico. Una posibilidad para incorporar esto consiste en correr una regresión en la cual el número de fonemas por sílaba sea la variable dependiente, y las variables independientes sean el número de sílabas por palabra, el número de palabras por frase, y una serie de variables categóricas adicionales que tengan que ver con la ubicación geográfica de los idiomas bajo estudio. Esto puede escribirse del siguiente modo:

$$FS = c(1)*Europe + c(2)*Africa + c(3)*Westasia + c(4)*Eastasia + c(5)*America + c(6)*SP + c(7)*PF \quad (8)\ ;$$

donde *Europe*, *Africa*, *Westasia*, *Eastasia* y *America* son variables que toman un valor igual a uno cuando el idioma pertenece a cierto grupo, y cero si no pertenece[6].

A fin de definir los grupos en los cuales hemos reunido a los distintos idiomas, hemos seguido un criterio básicamente geográfico. Esto hace que los grupos representen "áreas lingüísticas" relativamente grandes, y que sirvan para depurar ciertos efectos relacionados con posibles factores adicionales que no tienen que ver con los indicadores de complejidad analizados[7].

Así como la ecuación 8 puede servirnos para explicar los distintos factores que influyen sobre la cantidad de fonemas por sílaba, también las sílabas por palabra y las palabras por frase pueden ser analizadas como variables dependientes de regresiones en las cuales las variables independientes son las mismas que hemos definido en los párrafos anteriores. Eso implica escribir ecuaciones como las siguientes:

$$SP = c(11)*Europe + c(12)*Africa + c(13)*Westasia + c(14)*Eastasia + c(15)*America + c(16)*FS + c(17)*PF \quad (9)\ ;$$

$$PF = c(21)*Europe + c(22)*Africa + c(23)*Westasia + c(24)*Eastasia + c(25)*America + c(26)*FS + c(27)*SP \quad (10)\ .$$

[6] Los idiomas de nuestra muestra han sido clasificados de la siguiente manera: los idiomas europeos (*Europe*) son inglés, alemán, checo, polaco, húngaro, griego, francés, vasco, español y portugués; los africanos (*Africa*) son amárico, dinka, hausa, temne, igbo, mono y sandawe; los idiomas de Asia Occidental (*Westasia*) son georgiano, turco, árabe, persa, pashayi, hindi, bengalí y tamil; los de Asia Oriental (*Eastasia*) son japonés, coreano, chino, birmano, lizu, tailandés, vietnamita, malayo, tausug y arrernte; y los idiomas amerindios (*America*) son chickasaw, zapoteco, quichua, yine y mapuche.

[7] Para un resumen sobre la teoría de las áreas lingüísticas, véase Campbell (2006).

Debido a que los determinantes de las ecuaciones 8, 9 y 10 son básicamente los mismos, este es un ejemplo en el cual el análisis de regresión puede mejorar si utilizamos un método de regresión de ecuaciones simultáneas. Dicho enfoque es relativamente común en otras ciencias sociales tales como la economía, ya que permite introducir varios procedimientos que no están disponibles en las regresiones de ecuaciones aisladas. El más importante tiene que ver con el empleo de coeficientes de correlación entre los errores estadísticos de las ecuaciones estimadas, a través de lo que se conoce como "regresiones aparentemente no relacionadas" (SUR, por su sigla en inglés). Esto implica que, cuando estimamos una ecuación, estamos usando al mismo tiempo información de los resultados que se obtienen al estimar las otras ecuaciones, y dicha información puede mejorar la precisión y la eficiencia estadística de los coeficientes calculados.

Las ecuaciones 8, 9 y 10 pueden correrse por lo tanto de manera simultánea, y ver si eso permite encontrar cierta significación estadística para los coeficientes *c(6), c(7), c(16), c(17), c(26)* y *c(27),* que son los que miden las relaciones entre los distintos indicadores de complejidad lingüística. Los resultados de dicho análisis aparecen en el cuadro 5, que en sus primeras dos columnas muestra lo obtenido utilizando regresiones por mínimos cuadrados ordinarios (OLS), y en sus últimas dos columnas muestra lo obtenido utilizando regresiones aparentemente no relacionadas (SUR). Nótese que la bondad del ajuste para estas regresiones, medidas por sus coeficientes R^2, es considerablemente más alta que la que habíamos obtenido en la sección 3 (cuando corrimos regresiones con una única variable). Esto se debe a que las regresiones del cuadro 5 incluyen otras variables (geográficas) que ayudan a explicar parte de los cambios en los distintos indicadores de complejidad de los idiomas, que no pueden explicarse usando solamente las variables originales.

Empleando el mismo procedimiento descripto en la sección 4.1, podemos usar los resultados de estas regresiones para calcular nuevos coeficientes de correlación parcial. Ellos son los que aparecen en el cuadro 6, en el cual vemos que la correlación más elevada es la que corresponde a la relación entre sílabas por palabra y palabras por frase. En segundo lugar se ubican los coeficientes de correlación que corresponden a fonemas por sílaba versus sílabas por palabra, y por último aparecen los que miden la relación entre fonemas por sílaba y palabras por frase. El lector notará también que los

coeficientes de correlación obtenidos cuando utilizamos el método SUR son en todos los casos mayores que los que hallamos cuando utilizamos OLS (y son también más grandes que los coeficientes que aparecen en los cuadros 2 y 3).

Cuadro 5: Resultados de las regresiones con ecuaciones simultáneas

Concepto	OLS		SUR	
	Coeficiente	Probabilidad	Coeficiente	Probabilidad
Ec. de fonemas por sílaba				
Europe [c(1)]	2,952568	0,0000	3,577936	0,0000
Africa [c(2)]	2,766303	0,0000	3,449805	0,0000
Westasia [c(3)]	2,997346	0,0000	3,672135	0,0000
Eastasia [c(4)]	3,054742	0,0000	3,702805	0,0000
America [c(5)]	3,030417	0,0000	3,678272	0,0000
Sílabas/Palabras [c(6)]	-0,183989	0,0615	-0,329566	0,0002
Palabras/Frases [c(7)]	-0,025995	0,2584	-0,058515	0,0047
R-cuadrado	0,255757		0,198499	
Ec. de sílabas por palabra				
Europe [c(11)]	4,816465	0,0000	6,399916	0,0000
Africa [c(12)]	5,025491	0,0000	6,555629	0,0000
Westasia [c(13)]	5,194430	0,0000	6,719133	0,0000
Eastasia [c(14)]	5,063630	0,0000	6,622065	0,0000
America [c(15)]	5,190098	0,0000	6,587653	0,0000
Fonemas/Sílabas [c(16)]	-0,531664	0,0615	-0,952332	0,0002
Palabras/Frases [c(17)]	-0,158804	0,0000	-0,213053	0,0000
R-cuadrado	0,631023		0,562827	
Ec. de palabras por frase				
Europe [c(21)]	20,04769	0,0000	26,16882	0,0000
Africa [c(22)]	21,20632	0,0000	27,08390	0,0000
Westasia [c(23)]	20,77653	0,0000	27,39568	0,0000
Eastasia [c(24)]	20,32043	0,0000	26,92848	0,0000
America [c(25)]	19,44177	0,0000	26,51820	0,0000
Fonemas/Sílabas [c(26)]	-1,449251	0,2584	-3,262352	0,0047
Sílabas/Palabras [c(27)]	-3,063921	0,0000	-4,110577	0,0000
R-cuadrado	0,609189		0,554796	

Cuadro 6: Coeficientes de correlación parcial de las ecuaciones simultáneas

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Frases
Regresión OLS			
Fonemas por sílaba	1,0000	-0,3128	-0,1941
Sílabas por palabra		1,0000	-0,6975
Palabras por frase			1,0000
Regresión SUR			
Fonemas por sílaba	1,0000	-0,5602	-0,4369
Sílabas por palabra		1,0000	-0,9358
Palabras por frase			1,0000

4.3. Variables instrumentales

Un refinamiento adicional que puede incluirse en el análisis de regresión de ecuaciones simultáneas descripto en la sección 4.2 es el empleo de variables instrumentales. Se trata de un procedimiento diseñado para problemas estadísticos en los cuales hay varias variables que se supone que se determinan al mismo tiempo en el sistema de ecuaciones propuesto (como son, en nuestro caso, las variables *FS*, *SP* y *PF*), y que por lo tanto son "endógenas" a dicho sistema. Esto quiere decir que ninguna de dichas variables está verdaderamente determinada por las otras, sino que todas están definidas mediante un proceso gobernado por un entorno preestablecido. Para tratar ese tipo de problemas es necesario utilizar variables instrumentales, es decir, variables que se supone que están relacionadas con las variables endógenas bajo análisis, pero que tienen la propiedad de estar determinadas de manera exógena (o sea, fuera del problema estadístico que estamos analizando).

En esta sección, por lo tanto, vamos a reestimar el sistema definido por las ecuaciones 8, 9 y 10, usando un conjunto de variables instrumentales para eliminar la posible endogeneidad de *FS*, *SP* y *PF*. Dicho conjunto está constituido por tres variables fonológicas, tres variables morfológicas y tres variables sintácticas, las cuales han sido elegidas debido a su importancia relativa y a su disponibilidad. Las variables fonológicas son el número de fonemas consonánticos (*Consonantes*), el número de fonemas vocálicos (*Vocales*), y el número de tonos distintos que posee cada idioma (*Tonos*)[8], y sus valores han sido calculados usando las mismas fuentes empleadas para las distintas versiones de "El viento norte y el sol" (es decir, las correspondientes ilustraciones del IPA).

Las tres variables morfológicas que utilizaremos como instrumentos en nuestras regresiones, en cambio, han sido extraídas de la versión *online* del Atlas Mundial de Estructuras Lingüísticas (WALS)[9]. Ellas son el número de géneros posibles para los sustantivos que posee cada idioma (*Géneros*), el número de casos distintos que pueden presentar tales sustantivos (*Casos*) y el número de categorías de inflexión de los verbos

[8] Este último número es igual a uno para los idiomas en los cuales el tono no es un rasgo distintivo (por ejemplo, en los idiomas indoeuropeos), e igual al número de tonos distintivos (dos o más) para el caso de las lenguas tonales (por ejemplo, para los idiomas chino, japonés, igbo, vietnamita, zapoteco, etc.).
[9] Véase Dryer y Haspelmath (2013).

(*Inflexiones*).

Por último, las variables sintácticas que usaremos también tienen como fuente el WALS, y son el orden de objeto y verbo que es más común en cada idioma (*OV*), el orden de adjetivo y sustantivo que predomina en dichos idiomas (*AS*), y una tercera variable que mide el tipo de alineamiento morfosintáctico (*Acusativo*). Las tres están definidas como variables categóricas que toman un valor igual a uno en ciertos casos (cuando el orden es primero el objeto y luego el verbo, cuando el orden es primero el adjetivo y luego el sustantivo, y cuando el alineamiento morfosintáctico es de tipo nominativo-acusativo) y cero en los restantes.

El procedimiento básico para incluir variables instrumentales en una regresión de ecuaciones simultáneas recibe el nombre de "mínimos cuadrados en dos etapas" (2SLS, por su sigla en inglés). Consta de una primera etapa en la cual se corren las variables endógenas (en nuestro caso, *FS*, *SP* y *PF*) contra las variables instrumentales, usando mínimos cuadrados ordinarios. Una vez hecho esto, existe una segunda etapa en la cual los valores estimados en las regresiones de la primera etapa se incorporan a una estimación de ecuaciones simultáneas (en nuestro caso, a la estimación del sistema formado por las ecuaciones 8, 9 y 10), en lugar de los valores originales de las variables endógenas. El método admite también el uso de regresiones aparentemente no relacionadas, lo cual es equivalente a introducir una tercera etapa en la estimación. Por ello, esa variación del procedimiento se conoce con el nombre de "mínimos cuadrados en tres etapas" (3SLS)[10].

Los resultados de nuestra estimación por mínimos cuadrados en dos y en tres etapas son los que aparecen en cuadro 7. En él puede verse que los coeficientes R^2 que miden la bondad de ajuste son inferiores a los del cuadro 5, pero eso es algo que ocurre habitualmente cuando uno reemplaza variables endógenas por variables instrumentales exógenas (ya que dicho reemplazo implica sacrificar cierto grado de precisión a cambio de una mejora en la consistencia y en la eficiencia asintótica de la estimación).

De las cifras del cuadro 7 suge también que los valores correspondientes a los coeficientes de todas las variables geográficas toman valores con signo positivo, tanto en las regresiones en las que utilizamos 2SLS como en las que utilizamos 3SLS. En lo que

[10] Para una explicación más completa de estos métodos, véase Kennedy (2008), capítulo 10.

respecta a los coeficientes relacionados con las medidas de complejidad, en cambio, puede verse que *c(6)*, *c(16)*, *c(17)* y *c(27)* toman en ambos casos un valor negativo, pero que *c(7)* y *c(26)* toman un valor positivo (que no es estadísticamente significativo a ningún nivel razonable de probabilidad).

Cuadro 7: Resultados de las regresiones utilizando variables instrumentales

Concepto	2SLS		3SLS	
	Coeficiente	**Probabilidad**	**Coeficiente**	**Probabilidad**
Ec. de fonemas por sílaba				
Europe [c(1)]	2,593449	0,0024	2,615247	0,0000
Africa [c(2)]	2,378100	0,0105	2,403289	0,0000
Westasia [c(3)]	2,676996	0,0033	2,721860	0,0000
Eastasia [c(4)]	2,734213	0,0019	2,773210	0,0000
America [c(5)]	2,806677	0,0013	2,876394	0,0000
Sílabas/Palabras [c(6)]	-0,201131	0,2849	-0,238226	0,0147
Palabras/Frases [c(7)]	0,009094	0,8468	0,013178	0,6294
R-cuadrado	0,139173		0,063178	
Ec. de sílabas por palabra				
Europe [c(11)]	6,748064	0,0001	7,194962	0,0000
Africa [c(12)]	6,709945	0,0000	7,129098	0,0000
Westasia [c(13)]	7,115285	0,0001	7,549851	0,0000
Eastasia [c(14)]	7,069582	0,0001	7,516732	0,0000
America [c(15)]	7,190363	0,0001	7,605395	0,0000
Fonemas/Sílabas [c(16)]	-1,459411	0,1168	-1,607044	0,0088
Palabras/Frases [c(17)]	-0,138101	0,0382	-0,147353	0,0041
R-cuadrado	0,505005		0,466544	
Ec. de palabras por frase				
Europe [c(21)]	13,25793	0,2868	12,90198	0,1751
Africa [c(22)]	15,13207	0,1917	14,85197	0,0939
Westasia [c(23)]	14,07458	0,2756	13,77836	0,1624
Eastasia [c(24)]	13,35659	0,3084	13,02303	0,1937
America [c(25)]	12,65648	0,3476	12,39238	0,2285
Fonemas/Sílabas [c(26)]	1,463208	0,7441	1,693494	0,6225
Sílabas/Palabras [c(27)]	-3,062144	0,0177	-3,161858	0,0021
R-cuadrado	0,543246		0,529134	

Usando las cifras que aparecen en el cuadro 7, junto con fórmulas similares a las descriptas en las ecuaciones 5, 6 y 7, resulta posible hallar estimadores para los coeficientes de correlación parcial entre nuestros indicadores de complejidad de los idiomas. Dichos coeficientes son los que aparecen en el cuadro 8, y nos muestran una vez más el resultado de que la correlación más alta es la que se genera entre sílabas por palabra y palabras por frase (“r = -0,6503” y “r = -0,6826”). Un nuevo resultado que aparece ahora, y que difiere del que surge utilizando métodos menos sofisticados, es que

los coeficientes de correlación entre *FS* y *SP* (“r = -0,5418” y “r = -0,6187”) son sustancialmente más elevados que los coeficientes de correlación entre *FS* y *PS* (“r = 0,1154” and “r = 0,1494”). Estos últimos coeficientes, además, resultan ser positivos en vez de negativos, y su significación estadística es muy baja (“p = 0,2392” y “p = 0,1788”). La correlación entre fonemas por sílaba y sílabas por palabra, en cambio, se vuelve estadísticamente significativa a un nivel de probabilidad del 1% (“p = 0,0002” y “p ≈ 0”), lo mismo que la correlación entre sílabas por palabra y palabras por frase.

Cuadro 8: Coeficientes de correlación parcial usando variables instrumentales

Variable	Fonemas/Sílabas	Sílabas/Palabras	Palabras/Frases
Regresión 2SLS			
Fonemas por sílaba	1,0000	-0,5418	0,1154
Sílabas por palabra		1,0000	-0,6503
Palabras por frase			1,0000
Regresión 3SLS			
Fonemas por sílaba	1,0000	-0,6187	0,1494
Sílabas por palabra		1,0000	-0,6826
Palabras por frase			1,0000

Basándonos en los resultados de los cuadros 7 y 8, entonces, procedimos a realizar una reestimación de todo el sistema de ecuaciones, imponiendo las restricciones de que los coeficientes ligados con la correlación entre fonemas por sílaba y palabras por frase (es decir, *c(7)* y *c(26)*) fueran iguales a cero. Dichas estimaciones produjeron nuevos resultados, que a su vez nos permitieron calcular nuevos coeficientes de correlación parcial entre *FS* y *SP* (iguales a “r = -0,5820” y “r = -0,6682”), y entre *SP* y *PF* (iguales a “r = -0,6854” y “r = -0,7422”). Todos esos coeficientes de correlación resultaron ser negativos y estadísticamente significativos a un nivel de probabilidad del 1%, tanto cuando usamos el procedimiento de mínimos cuadrados en dos etapas como cuando empleamos mínimos cuadrados en tres etapas.

5. Consideraciones finales

Luego de llevar a cabo los ejercicios numéricos descriptos en las secciones 3 y 4, podemos ahora concluir con algunas consideraciones que resumen los resultados obtenidos. La principal conclusión a la que hemos arribado es que, usando la base de datos armada por nosotros con 40 versiones diferentes de “El viento norte y el sol”, la

complejidad morfológica (medida a través del número promedio de sílabas por palabra para cada idioma) está negativamente correlacionada con la complejidad sintáctica (medida a través del número promedio de palabras por frase), y que dicha relación es lo suficientemente fuerte como para aparecer en todos los tipos de análisis aplicados a nuestros datos.

Cuando introducimos la idea de que los distintos tipos de complejidad pueden estar interrelacionados entre sí, encontramos también una correlación negativa relativamente fuerte entre complejidad fonológica (medida a través del número de fonemas por sílaba) y complejidad morfológica. Esta conclusión se refuerza con el empleo de ecuaciones simultáneas y de variables instrumentales, que corrigen el problema de endogeneidad generado por un sistema en el que los distintos indicadores de complejidad lingüística son al mismo tiempo variables independientes y dependientes.

Estas dos conclusiones pueden relacionarse con un resultado tradicional de la lingüística cuantitativa, que es la ley de Menzerath. De acuerdo con dicha ley, la medida de cada elemento lingüístico debería estar negativamente correlacionada con la medida de los componentes de dicho elemento. Como las sílabas son los componentes de las palabras, y las palabras son los componentes de las frases, entonces las correlaciones halladas pueden verse como un ejemplo de dicho resultado más general.

La relación entre fonemas por sílaba y palabras por frase, en cambio, no es estadísticamente significativa en la mayor parte de nuestros análisis. Aunque el coeficiente de correlación simple entre dichas variables es negativo, su valor absoluto es pequeño, y eso hace que no resulte estadísticamente significativo a ningún nivel razonable de probabilidad. Si bien vemos que, cuando calculamos un coeficiente de correlación parcial entre las los dos indicadores, el valor absoluto de dicho coeficiente se incrementa, vemos también que, cuando introducimos variables instrumentales, el mismo se vuelve positivo y estadísticamente insignificante[11].

Todas estas conclusiones se basan en las cifras que hemos resumido en el cuadro 9. En él vemos que los coeficientes de correlación calculados para la relación entre

[11] Este resultado también puede relacionarse con la ley de Menzerath. Dado que las sílabas no son componentes directos de las frases (ya que su relación es indirecta, a través de las palabras formadas por tales sílabas), resulta entonces razonable que no encontremos ninguna correlación significativa entre fonemas por sílaba y palabras por frase.

sílabas por palabra y palabras por frase son siempre negativos y mayores que 0,65 en valor absoluto. También son estadísticamente significativos a cualquier nivel razonable de probabilidad, ya que sus valores-p son siempre iguales a cero. Cuando calculamos coeficientes de correlación parcial, la relación entre fonemas por sílaba y sílabas por palabra también se vuelve estadísticamente significativa a un nivel de probabilidad del 5%, y su valor absoluto (negativo) pasa a estar en un rango entre 0,31 y 0,67. Esto no ocurre con los coeficientes de correlación para la relación entre fonemas por sílaba y palabras por frase, cuyo valor absoluto se vuelve positivo y menor que 0,15 cuando usamos procedimientos basados en el empleo de variables instrumentales (y resulta insignificante desde el punto de vista estadístico).

Cuadro 9: Resumen de los coeficientes de correlación estimados

Concepto	**FS vs. SP**	**SP vs. PF**	**FS vs. PF**
Correlación simple			
Coeficiente	-0,1630	-0,7257	-0,0855
Probabilidad	(0,1575)	(0,0000)	(0,2999)
Correlación parcial			
Coeficiente	-0,3283	-0,7525	-0,3002
Probabilidad	(0,0193)	(0,0000)	(0,0299)
OLS con variables geográficas			
Coeficiente	-0,3128	-0,6975	-0,1941
Probabilidad	(0,0247)	(0,0000)	(0,1151)
SUR con variables geográficas			
Coeficiente	-0,5602	-0,9358	-0,4369
Probabilidad	(0,0001)	(0,0000)	(0,0024)
2SLS con variables geográficas			
Coeficiente	-0,5418	-0,6503	0,1154
Probabilidad	(0,0002)	(0,0000)	(0,2392)
3SLS con variables geográficas			
Coeficiente	-0,6187	-0,6826	0,1494
Probabilidad	(0,0000)	(0,0000)	(0,1788)
2SLS con restricciones			
Coeficiente	-0,5820	-0,6854	0,0000
Probabilidad	(0,0000)	(0,0000)	
3SLS con restricciones			
Coeficiente	-0,6682	-0,7422	0,0000
Probabilidad	(0,0000)	(0,0000)	

Las conclusiones obtenidas a través de nuestros análisis estadísticos pueden compararse con los resultados de trabajos anteriores. Los más compatibles son sin duda los que obtuvieron Fenk-Oczlon y Fenk (2004), quienes también hallaron correlaciones negativas y significativas entre complejidad fonológica y morfológica, y entre

complejidad morfológica y sintáctica, en un contexto en el cual sus resultados pueden verse como un corolario de la ley de Menzerath[12]. La causa más probable de esta coincidencia es el hecho de que dichos autores miden la complejidad lingüística usando los mismos indicadores que nosotros, si bien su base de datos es completamente distinta de la nuestra y muchos de los idiomas utilizados por ellos son diferentes.

La correlación entre las distintas medidas de complejidad, en cambio, parece ser insignificante en otros estudios tales como el de Shosted (2006). En dicho trabajo el autor mide la complejidad de los idiomas usando medidas teóricas de ciertos aspectos fonológicos y morfológicos (tales como el número de tipos posibles de sílabas, y el número de categorías de inflexión de los verbos), en vez de usar indicadores estadísticos reales basados en el análisis de textos. Shosted, además, calcula la correlación entre las variables utilizando únicamente coeficientes de Pearson, y no incluye por lo tanto en su análisis a los coeficientes de correlación parcial ni a los que surgen de procedimientos de regresión con ecuaciones simultáneas, ni utiliza tampoco variables instrumentales para corregir posibles problemas de endogeneidad. Todo esto puede ser la causa por la cual sus resultados discrepan considerablemente de los nuestros.

Cuando los estudios multilingüísticos de complejidad incluyen alguno de los factores mencionados en el párrafo anterior, entonces sí es factible que aparezcan coeficientes de correlación con signo negativo, que además resulten estadísticamente significativos. Tal es el caso de los resultados de Winchmann et al. (2011), que muestran la existencia de correlación negativa entre el número de fonemas por palabra y un indicador basado en el número de fonemas de cada idioma. Lo mismo ocurre en el trabajo de Moran y Blasi (2014), que encuentra una correlación negativa entre el número de fonemas vocálicos y el número de fonemas por palabra.

En un trabajo anterior nuestro (Coloma, 2013) también nos apareció un fenómeno de correlación negativa dentro del subsistema fonológico de los idiomas analizados. En efecto, usando técnicas de regresión con ecuaciones simultáneas, en dicho trabajo encontramos correlaciones negativas significativas entre una variable que medía el uso del acento como elemento distintivo y otras tres variables adicionales (número de consonantes, número de vocales y distinción por tonos), en el contexto de una muestra de

[12] Véase también Fenk-Oczlon y Fenk (2008).

100 idiomas pertenecientes a diferentes familias lingüísticas y a distintas ubicaciones geográficas.

Referencias bibliográficas

Bane, Max (2008). Quantifying and Measuring Morphological Complexity. En *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 69-76. Somerville: Cascadilla Proceedings Project.

Campbell, Lyle (2006). Areal Linguistics: A Closer Scrutiny. En Y. Matras, A. McMahon & N. Vincent (eds.), *Linguistic Areas*, 1-31. Hampshire: Palgrave Macmillan.

Coloma, Germán (2013). Un modelo estadístico de ecuaciones simultáneas sobre la interacción de variables fonológicas, Documento de Trabajo Nro 519. Buenos Aires: Universidad del CEMA.

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

Fenk-Oczlon, Gertraud & August Fenk (2004). Systemic Typologies and Crosslinguistic Regularities. En V. Solovyev & V. Polyakov (eds.), *Text Processing and Cognitive Technologies*, 229-234. Moscú: MISA.

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

IPA (1949). *Principles of the International Phonetic Association*. Londres: University College.

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

Kennedy, Peter (2008). *A Guide to Econometrics*, 6ta edición. Nueva York: Wiley (hay versión en castellano: *Introducción a la econometría*; México: Fondo de Cultura Económica).

Maddieson, Ian (2007). Issues of Phonological Complexity: Statistical Analysis of the Relationship Between Syllable Structures, Segment Inventories and Tone Contrasts. En M. Solé, P. Beddor & M. Ohala (eds.), *Experimental Approaches to Phonology*, 93-103. New York: Oxford University Press.

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Moran, Steven & Damián Blasi (2014). Cross-Linguistic Comparison of Complexity Measures in Phonological Systems. En F. Newmayer & L. Preston (eds.), *Measuring Grammatical Complexity*, en prensa. Nueva York: Oxford University Press.

Nettle, Daniel (1995). Segmental Inventory Size, Word Length and Communicative Efficiency. *Linguistics* 33: 359-367.

Oh, Yoon-Mi, François Pellegrino, Egidio Marsico & Christophe Coupé (2013). A Quantitative and Typological Approach to Correlating Linguistic Complexity. En *Proceedings of the 5th Conference on Quantitative Investigations in Theoretical Linguistics*, 71-75. Lovaina: University of Leuven.

Shosted, Ryan (2006). Correlating Complexity: A Typological Approach. *Linguistic Typology* 10: 1-40.
Szmrecsányi, Benedikt (2004). On Operationalizing Syntactic Complexity. En *Annals of the 7th International Conference of Statistical Text Analysis (JADT)*, 1031-1038.
Wichmann, Soren, Taraka Rama & Eric Holman (2011). Phonological Diversity, Word Length and Population Sizes Across Languages: The ASJP Evidence. *Linguistic Typology* 15: 157-177.

Fuentes de los datos utilizados

Arvaniti, Amalia (1999). Standard Modern Greek. *Journal of the International Phonetic Association* 29: 167-172.
Breen, Gavan & Veronica Dobson (2005). Central Arrernte. *Journal of the International Phonetic Association* 35: 249-254.
Chirkova, Katia & Yiya Chen (2013). Lizu. *Journal of the International Phonetic Association* 43: 75-86.
Clynes, Adrian & David Deterding (2011). Standard Malay (Brunei). *Journal of the International Phonetic Association* 41: 259-268.
Cruz-Ferreira, Madalena (1999). Portuguese (European). En IPA (1999), 126-130.
Dankovicova, Jana (1999). Czech. En IPA (1999), 70-73.
Eaton, Helen (2006). Sandawe. *Journal of the International Phonetic Association* 36: 235-242.
Fougeron, Cécile & Caroline Smith (1999). French. En IPA (1999), 78-81.
Gordon, Matthew, Pamela Munro & Peter Ladefoged (2001). Chickasaw. *Journal of the International Phonetic Association* 31: 287-290.
Hayward, Katrina & Richard Hayward (1999). Amharic. En IPA (1999), 45-50.
Hualde, José, Oihana Lujanbio & Juan Zubiri (2010). Goizueta Basque. *Journal of the International Phonetic Association* 40: 113-127.
Ikekeonwu, Clara (1999). Igbo. En IPA (1999), 108-110.
Jassem, Wiktor (2003). Polish. *Journal of the International Phonetic Association* 33: 103-107.
Kahn, Sameer (2010). Bengali (Bangladeshi Standard). *Journal of the International Phonetic Association* 40: 221-225.
Kanu, Sullay & 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.
Kohler, Klaus (1999). German. En IPA (1999), 86-89.
Lamuwal, Abd-El-Malek & Adam Baker (2013). Southeastern Pashayi. *Journal of the International Phonetic Association* 43: 243-246.
Lee, Hyun Bok (1999). Korean. En IPA (1999), 120-123.
Lee, Wai-Sum & Eric Zee (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33: 109-112.
Majidi, Mohammad & Elmar Ternes (1999). Persian (Farsi). En IPA (1999), 124-125.

Martínez, Eugenio, Ana Fernández & Josefina Carrera (2003). Castilian Spanish. *Journal of the International Phonetic Association* 33: 255-260.
Masaquiza, Fanny & Stephen Marlett (2008). Salasaca Quichua. *Journal of the International Phonetic Association* 38: 223-227.
Ohala, Manjari (1999). Hindi. En IPA (1999), 100-103.
Okada, Hideo (1999). Japanese. En IPA (1999), 117-119.
Olson, Kenneth (2004). Mono. *Journal of the International Phonetic Association* 34: 233-238.
Pickett, Velma, María Villalobos & Stephen Marlett (2010). Isthmus (Juchitán) Zapotec. *Journal of the International Phonetic Association* 40: 365-372.
Remijsen, Bert & Caguor Manyang (2009). Luanyjang Dinka. *Journal of the International Phonetic Association* 39: 123-124.
Roach, Peter (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association* 34: 239-245.
Sadowsky, Scott, Héctor Painequeo, Gastón Salamanca & Heriberto Avelino (2013). Mapudungun. *Journal of the International Phonetic Association* 43: 87-96.
Schuh, Russell & Lawan Yalwa (1999). Hausa. En IPA (1999), 90-95.
Shosted, Ryan & Vakhtang Chikovani (2006). Standard Georgian. *Journal of the International Phonetic Association* 36: 255-264.
Soderberg, Craig, Seymour Ashley & Kenneth Olson (2012). Tausug (Suluk). *Journal of the International Phonetic Association* 42: 361-364.
Szende, Tamás. Hungarian. En IPA (1999), 104-107.
Thelwall, Robin & Akram Sa'adeddin (1999). Arabic. En IPA (1999), 51-54.
Tingsabadh, Kalaya & Arthur Abramson (1999). Thai. En IPA (1999), 147-150.
Urquía, Rittma & Stephen Marlett (2008). Yine. *Journal of the International Phonetic Association* 38: 365-369.
Watkins, Justin (2001). Burmese. *Journal of the International Phonetic Association* 31: 291-295.
Zimmer, Karl & Orhan Orgun (1999). Turkish. En IPA (1999), 154-156.

Apéndice 1: Datos extraídos de los textos de "El viento norte y el sol"

Los siguientes datos corresponden a las cifras usadas para calcular los indicadores para los 40 idiomas de nuestra base de datos. Se mencionan también las fuentes de dicha información.

Idioma	Frases	Palabras	Sílabas	Fonemas	Fuente
Alemán	10	108	180	456	Kohler (1999)
Amárico	8	94	259	661	Hayward & Hayward (1999)
Árabe	9	85	217	488	Thelwall & Sa'adeddin (1999)
Arrernte	12	73	194	436	Breen & Dobson (2005)
Bengalí	10	104	197	459	Kahn (2010)
Birmano	7	42	131	300	Watkins (2001)
Checo	10	84	162	394	Dankovicova (1999)
Chickasaw	10	57	184	474	Gordon et al. (2001)
Chino	10	98	157	421	Lee & Zee (2003)
Coreano	7	60	166	381	Lee (1999)
Dinka	10	137	288	548	Remijsen & Manyang (2009)
Español	9	107	193	425	Martínez et al. (2003)
Francés	9	108	159	343	Fougeron & Smith (1999)
Georgiano	9	70	177	418	Shosted & Chikovani (2006)
Griego	9	104	222	479	Arvaniti (1999)
Hausa	12	166	282	648	Schuh & Yalwa (1999)
Hindi	8	125	208	467	Ohala (1999)
Húngaro	10	100	192	431	Szende (1999)
Igbo	8	107	208	356	Ikekeonwu (1999)
Inglés	9	113	143	383	Roach (2004)
Japonés	9	89	227	444	Okada (1999)
Lizu	8	110	215	450	Chirkova & Chen (2013)
Malayo	8	78	209	481	Clynes & Deterding (2011)
Mapuche	9	75	151	360	Sadowsky et al. (2013)
Mono	10	115	181	338	Olson (2004)
Pashayi	7	89	177	393	Lamuwal & Baker (2013)
Persa	9	91	194	483	Majidi & Ternes (1999)
Polaco	9	89	158	428	Jassem (2003)
Portugués	8	98	186	380	Cruz-Ferreira (1999)
Quichua	11	94	271	593	Masaquiza & Marlett (2008)
Sandawe	9	79	182	383	Eaton (2006)
Tailandés	11	131	173	480	Tingsabadh & Abramson (1999)
Tamil	9	79	252	541	Keane (2004)
Tausug	12	114	238	572	Soderberg et al. (2012)
Temne	11	125	207	446	Kanu & Tucker (2010)
Turco	9	66	183	431	Zimmer & Orgun (1999)
Vasco	7	83	187	401	Hualde et al. (2010)
Vietnamita	7	117	117	334	Kirby (2011)
Yine	10	63	236	559	Urquía & Marlett (2008)
Zapoteco	9	87	154	327	Pickett et al. (2010)

Apéndice 2: Datos de las variables instrumentales

Los siguientes datos corresponden a las cifras usadas como variables instrumentales en nuestros análisis. Se mencionan también los grupos a los cuales pertenecen los distintos idiomas, de acuerdo con el criterio geográfico utilizado.

Idioma	Grupo	Consonantes	Vocales	Tonos	Casos
Alemán	Europe	20	15	1	4
Amárico	Africa	27	7	1	2
Árabe	Westasia	29	6	1	1
Arrernte	Eastasia	27	4	1	8
Bengalí	Westasia	29	7	1	6
Birmano	Eastasia	34	9	4	8
Checo	Europe	25	9	1	5
Chickasaw	America	16	9	1	2
Chino	Eastasia	19	6	4	1
Coreano	Eastasia	19	18	1	6
Dinka	Africa	20	7	4	1
Español	Europe	19	5	1	1
Francés	Europe	20	13	1	1
Georgiano	Westasia	28	5	1	6
Griego	Europe	18	5	1	3
Hausa	Africa	28	10	2	1
Hindi	Westasia	34	11	1	2
Húngaro	Europe	25	14	1	10
Igbo	Africa	26	8	3	1
Inglés	Europe	24	11	1	2
Japonés	Eastasia	16	5	2	8
Lizu	Eastasia	39	8	2	1
Malayo	Eastasia	18	6	1	1
Mapuche	America	22	6	1	2
Mono	Africa	32	8	3	1
Pashayi	Westasia	27	11	1	4
Persa	Westasia	23	6	1	2
Polaco	Europe	31	6	1	6
Portugués	Europe	19	13	1	1
Quichua	America	23	3	1	8
Sandawe	Africa	44	15	2	1
Tailandés	Eastasia	21	9	5	1
Tamil	Westasia	15	10	1	6
Tausug	Eastasia	17	3	1	1
Temne	Africa	19	9	2	1
Turco	Westasia	22	8	1	6
Vasco	Europe	23	5	1	10
Vietnamita	Eastasia	22	11	8	1
Yine	America	16	5	1	2
Zapoteco	America	20	5	3	1

Idioma	Géneros	Inflexiones	OV	AS	Acusativo
Alemán	3	2	0	1	1
Amárico	2	6	1	1	1
Árabe	2	6	0	0	1
Arrernte	1	4	1	0	0
Bengalí	2	2	1	1	1
Birmano	1	2	1	0	0
Checo	3	4	0	1	1
Chickasaw	1	6	1	0	0
Chino	1	1	0	1	0
Coreano	1	6	1	1	0
Dinka	1	6	1	0	1
Español	2	4	0	0	1
Francés	2	4	0	0	1
Georgiano	1	8	1	1	1
Griego	3	4	0	1	1
Hausa	2	6	0	1	0
Hindi	2	2	1	1	1
Húngaro	1	4	0	1	1
Igbo	1	6	0	0	0
Inglés	1	2	0	1	1
Japonés	1	4	1	1	0
Lizu	1	3	1	0	0
Malayo	1	4	0	0	1
Mapuche	1	8	0	1	0
Mono	5	6	1	0	0
Pashayi	2	2	1	1	1
Persa	1	4	1	0	1
Polaco	3	4	0	1	1
Portugués	2	4	0	0	1
Quichua	1	8	1	1	1
Sandawe	5	8	1	0	1
Tailandés	1	2	0	0	0
Tamil	3	2	1	1	1
Tausug	1	4	0	0	0
Temne	5	2	0	0	1
Turco	1	6	1	1	1
Vasco	1	4	1	0	0
Vietnamita	1	1	0	0	0
Yine	4	6	1	0	0
Zapoteco	1	8	0	0	1