UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

UNA COMPARACIÓN INTERLINGÜÍSTICA DE DOS FORMULACIONES ALTERNATIVAS DE LA LEY DE MENZERATH

Germán Coloma

Mayo 2015
Nro. 564

ISBN 978-987-3940-00-2
Queda hecho el depósito que marca la Ley 11.723

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>

Coloma, Germán
Una : Comparación interlingüística de dos formulaciones alternativas de la ley de Menzerath . - 1a ed. - Ciudad Autónoma de Buenos Aires : Universidad del CEMA, 2015.
22 p. ; 22x15 cm.

ISBN 978-987-3940-00-2

1. Economia. 2. Ciencia Politica.
CDD 330

Fecha de catalogación: 08/06/2015

UNA COMPARACIÓN INTERLINGÜÍSTICA DE DOS FORMULACIONES ALTERNATIVAS DE LA LEY DE MENZERATH

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

Resumen

Este trabajo intenta comparar dos formulaciones alternativas de la ley de Menzerath para la relación entre dos indicadores lingüísticos (palabras por enunciado y fonemas por palabra) en una muestra de 50 idiomas. Las formulaciones alternativas son la función potencial (que es la especificación más tradicional de la ley de Menzerath), y una forma recientemente propuesta de tipo hiperbólico. Las estimaciones han sido modificadas para controlar por factores filogenéticos y geográficos, y por la presencia de posible endogeneidad entre las variables. Nada de eso, sin embargo, altera significativamente los resultados básicos, que muestran una leve preferencia por la función potencial por sobre la función hiperbólica.

Palabras clave: ley de Menzerath, indicadores lingüísticos, función potencial, función hiperbólica.

1. Introducción

La ley de Menzerath, o ley de Menzerath-Altmann, es una regularidad lingüística bien estudiada por la literatura, que establece que la medida de un elemento debería estar negativamente correlacionada con la medida de los componentes de dicho elemento. Propuesta originalmente por Menzerath (1954), la ley fue reformulada por Altmann (1980) como una relación potencial que puede escribirse del siguiente modo:

$$y = a \cdot x^{b} \qquad (1)\ ;$$

donde *y* es la medida de un elemento lingüístico, *x* es la medida promedio de los componentes de dicho elemento, y *a* y *b* son parámetros[1].

Existen numerosas aplicaciones de esta ley a diversas bases de datos lingüísticas. Las originales de Menzerath y de Altmann se referían a la comparación entre las medidas

* Agradezco los comentarios de Gabriel Altmann, Mariana Conte Grand, Stefan Gries, Reinhard Kohler y Fermín Moscoso a una versión anterior de este trabajo. Agradezco también a Federico Cápula, Helen Eaton, John Esling, Sameer Kahn, Kevin Schäfer y Justin Watkins por su ayuda para acceder a algunas de las fuentes utilizadas. Parte de la investigación para este trabajo fue llevada a cabo durante mi estadía como investigador visitante en la Universidad de California, Santa Bárbara (UCSB). Las opiniones son personales y no representan necesariamente las de la Universidad del CEMA ni las de la UCSB.

[1] En rigor, la formulación propuesta por Altmann incluía también un término de tipo exponencial (e^{cx}), el cual ha desaparecido en la mayoría de las aplicaciones posteriores de la ley de Menzerath.

de las palabras y las de las sílabas que las constituían, pero luego aparecieron estudios que relacionaron la medida de los enunciados con la de las palabras constituyentes (Teupenhayn y Altmann, 1984), la medida de los enunciados con la de las frases constituyentes (Kulacka, 2010), y la medida de las palabras con el número de palabras diferentes utilizadas (Eroglu, 2013), entre otras comparaciones[2]. Aunque la ley de Menzerath también ha sido mencionada como una posible explicación de la ocurrencia de este tipo de fenómenos en contextos interlingüísticos (Fenk-Oczlon y Fenk, 1999), la mayoría de los análisis efectuados utilizan textos escritos en un único idioma.

En un trabajo reciente (Milicka, 2014) se argumenta que la fórmula tradicional de la ley de Menzerath (basada en una función potencial) puede mejorarse si se emplea una alternativa hiperbólica, que puede escribirse de la siguiente manera:

$$y = a + \frac{b}{x} \qquad (2)\ .$$

Se supone que esta fórmula ajusta mejor ciertos datos y que tiene una explicación más intuitiva, relacionada con un posible efecto de compensación (*trade-off*) entre información pura y estructura informativa (Kohler, 1984)[3].

En las siguientes secciones de este trabajo procederemos a comparar las implicancias de las dos formulaciones de la ley de Menzerath en una muestra de 50 idiomas para los cuales hemos conseguido datos correspondientes a un mismo texto. En cada caso calcularemos el número de fonemas por palabra y el número de palabras por enunciado, y veremos qué versión de la ley ajusta mejor los datos. Nuestra comparación será posteriormente mejorada a través de dos tests de especificación diferentes, y de la inclusión de los posibles efectos de dos variables categóricas: ubicación geográfica y filiación genética de los idiomas. También incluiremos una corrección ligada con la posible endogeneidad del número de fonemas por palabra como variable explicativa del número de palabras por enunciado, utilizando variables instrumentales.

[2] Algunas aplicaciones de la ley de Menzerath han ido inclusive más allá, y han testeado la existencia de relaciones similares en áreas totalmente alejadas de la lingüística. Véase, por ejemplo, Boroda y Altmann (1991), aplicado a textos musicales, y Ferrer y Forns (2010), aplicado al estudio de los genomas.

[3] La especificación tradicional de la ley de Menzerath, sin embargo, también ha sido objeto de varias explicaciones teóricas. Eroglu (2014), por ejemplo, la ha interpretado como un caso particular de una "organización mecánica de tipo estadístico".

2. Descripción de los datos

El texto del cual derivaremos los resultados presentados en este trabajo es la fábula conocida como “El viento norte y el sol”, atribuida a Esopo, que es un cuento corto utilizado por la Asociación Fonética Internacional (IPA) como un “espécimen” o modelo para ilustrar la fonética de un importante número de lenguas. Dicho texto tiene la ventaja de que describe claramente los fonemas, las palabras y los enunciados, y de que es inmediatamente comparable entre idiomas. Por ejemplo, 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.

y su correspondiente transcripción fonémica es:

el 'biento 'noɾte i el 'sol por'fiaban sobɾe 'kual de 'eʌos 'eɾa el 'mas 'fueɾte | kuando aθeɾ'to a pa'saɾ un bia'xeɾo em'buelto en 'antʃa 'kapa || kombi'nieɾon en ke kien 'antes lo'gɾaɾa obli'gaɾ al bia'xeɾo a ki'taɾse la 'kapa se'ɾia konside'ɾado 'mas pode'ɾoso || el 'biento 'noɾte so'plo kon 'gran 'fuɾia | peɾo 'kuanto 'ma so'plaba 'mas se arebu'xaba en su 'kapa el bia'xeɾo || poɾ 'fin el 'biento 'noɾte abando'no la em'pɾesa || en'tonθes bɾi'ʌo el 'sol kon aɾ'doɾ | e imme'diata'mente se despo'xo de su 'kapa el bia'xeɾo | poɾ lo ke el 'biento 'noɾte 'ubo de rekono'θeɾ la supeɾioɾi'da del 'sol ||

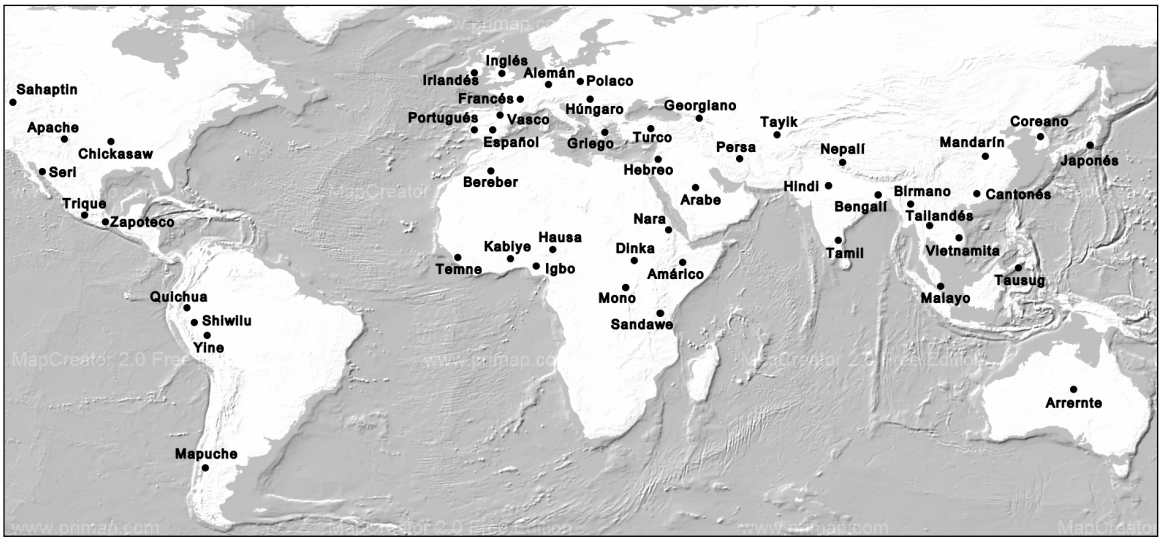
Si contamos el número de enunciados, palabras y fonemas en este texto, podemos hallar que el mismo consta de 9 enunciados[4], 107 palabras y 425 fonemas, y que por lo tanto tiene en promedio 3,97 fonemas por palabra y 11,89 palabras por enunciado. Estos mismos cálculos pueden llevarse a cabo para otros idiomas con relaciones diferentes entre los cocientes definidos. Por ejemplo, la versión turca de “El viento norte y el sol” tiene un número mayor de fonemas por palabra (igual a 6,53) que la versión española, pero un número menor de palabras por enunciado (igual a 7,33).

A fin de llevar a cabo nuestro análisis, hemos elegido una muestra de 50 idiomas

[4] El concepto de “enunciado” que usamos para este cálculo está basado en el número de pausas marcadas en el texto fonémico, y no en consideraciones sintácticas. Esto permite realizar más fácilmente las comparaciones cuando tomamos en cuenta las versiones del mismo texto en distintos idiomas.

para los cuales hemos hallado versiones del texto reproducido más arriba, ya sea en el *Handbook of the International Phonetic Association* (IPA, 1999) o en la serie de artículos titulada "Illustrations of the IPA" y publicada en el *Journal of the International Phonetic Association*. Esta muestra consta de diez idiomas de cada una de las cinco áreas en las que hemos dividido el mundo: América (sahaptin, apache, chickasaw, seri, trique, zapoteco, quichua, shiwilu, yine y mapuche), Europa (portugués, español, vasco, francés, irlandés, inglés, alemán, polaco, húngaro y griego), África (bereber, nara, dinka, amárico, sandawe, mono, hausa, igbo, kabiye y temne), Asia occidental (georgiano, turco, hebreo, árabe, persa, tayik, nepalí, hindi, bengalí y tamil) y Asia oriental (japonés, coreano, mandarín, cantonés, birmano, tailandés, vietnamita, malayo, tausug y arrernte)[5]. Incluye todos los idiomas disponibles cuyo número de hablantes supera los 80 millones[6], junto con otros ejemplos representativos de diversas familias lingüísticas y ubicaciones geográficas (ver mapa 1).

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



La base de datos completa está reproducida en el cuadro que aparece en el apéndice 1. En él puede verse que el número promedio de fonemas por palabra en toda la muestra es igual a 4,94, con un mínimo de 2,85 (que corresponde al idioma vietnamita) y

[5] En esta división, Australia (donde se habla la lengua arrernte) se considera parte de Asia oriental.
[6] Los únicos idiomas con más de 80 millones de hablantes que no están incluidos son el ruso, el panyabí y el javanés, para los cuales no hay disponible ninguna ilustración del IPA. Los mismos han sido reemplazados en la muestra por otras lenguas similares, que son el polaco, el nepalí y el malayo.

un máximo de 8,87 (que corresponde al yine, que es una lengua arahuaca hablada en Perú)[7]. El máximo valor para el cociente entre palabras y enunciados, en cambio, corresponde al idioma irlandés (y es igual a 18,43), en tanto que el mínimo valor para dicho cociente es igual a 5,70 (y corresponde al chickasaw, un idioma muskogueano hablado en Estados Unidos), en un contexto en el cual el número promedio de palabras por enunciado es igual a 10,37.

El número de fonemas por palabra y el número de palabras por enunciado tienen una correlación negativa relativamente elevada en esta muestra. Medida por el coeficiente de Pearson, dicha correlación es igual a -0,7158, y ese coeficiente es estadísticamente distinto de cero a cualquier nivel razonable de probabilidad[8].

3. Formulaciones alternativas de la ley de Menzerath

Para testear el desempeño relativo de las dos formulaciones alternativas de la ley de Menzerath que hemos mencionado en la sección 1, en esta sección correremos una serie de regresiones utilizando los datos descriptos en la sección 2. Dichas regresiones se basan las siguientes fórmulas:

$$Ln(Word/Clause) = c(1) + c(2)*Ln(Phon/Word) \quad (3)\ ;$$

$$Word/Clause = c(1) + c(2)*[1/(Phon/Word)] \quad (4)\ ;$$

las cuales son transformaciones lineales de las ecuaciones 1 y 2 para el caso en el que la variable independiente es el cociente entre fonemas y palabras (*Phon/Word*) y la variable dependiente es el cociente entre palabras y enunciados (*Word/Clause*).

Los principales resultados de dichas regresiones, utilizando mínimos cuadrados ordinarios (OLS), son los que aparecen en el cuadro 1. En él vemos que ambas

[7] A fin de calcular estos números, primero tuvimos que definir el número de palabras y fonemas en cada versión de "El viento norte y el sol". Para ello, seguimos básicamente los criterios utilizados por los autores que escribieron las correspondientes ilustraciones del IPA, pero también aplicamos algunos criterios unificadores. Por ejemplo, las vocales cortas, largas, orales y nasales son consideradas como fonemas diferentes cuando la duración o la nasalización son rasgos distintivos en un idioma en particular, pero los diptongos son siempre considerados como la combinación de dos fonemas. Las consonantes africadas y otras "articulaciones dobles" también han sido consideradas como fonemas separados cuando la fonología de un determinado idioma así lo indicaba, mientras que las "consonantes geminadas" fueron siempre analizadas como una combinación de dos fonemas idénticos consecutivos.

[8] En rigor, teniendo en cuenta que este coeficiente de correlación surge de una muestra de 50 observaciones (con 48 grados de libertad), su correspondiente estadístico-t es igual a -7,1025. Dicho estadístico genera un valor-p igual a 0,000000005.

especificaciones generan un buen ajuste de los datos, y que los coeficientes estimados son altamente significativos y tienen los signos esperados (ya que ambos implican una relación negativa entre *Word/Clause* y *Phon/Word*)[9]. Basándonos en los coeficientes de determinación (R^2) de estas regresiones, podemos ver también que el ajuste de la función potencial (R^2 = 0,5851) es levemente mejor que el que se obtiene con la función hiperbólica (R^2 = 0,5560)[10].

Cuadro 1: Resultados de las regresiones por mínimos cuadrados ordinarios

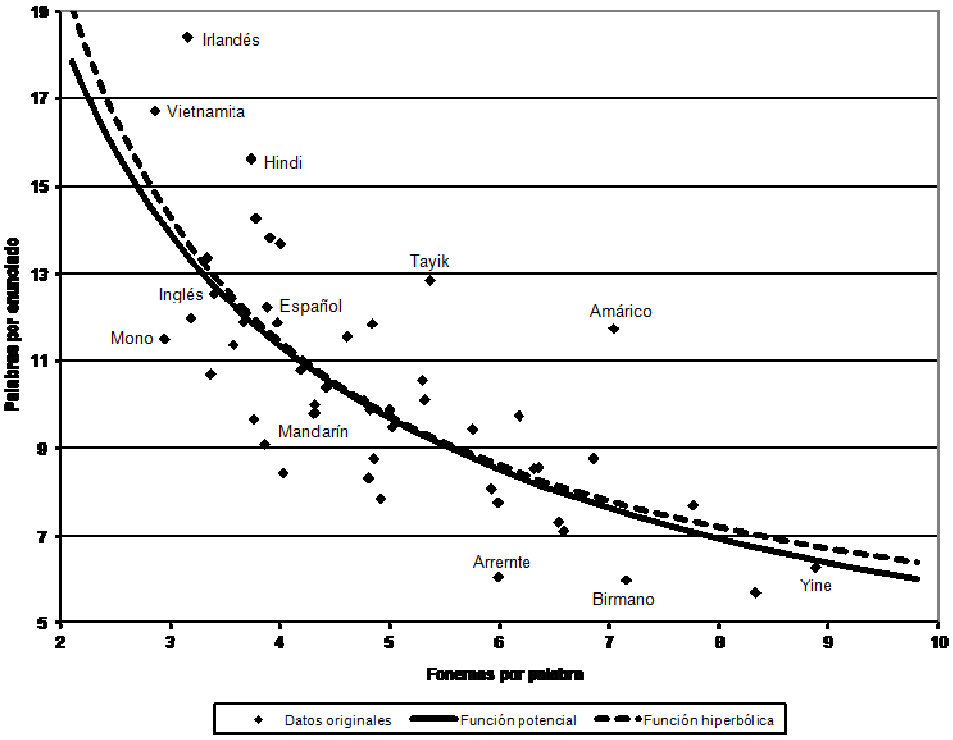
Concepto	Coeficiente	Error típico	Estadístico-t	Probabilidad
Función potencial				
Constante [c(1)]	3,4072	0,1359	25,0684	0,0000
Phon/Word [c(2)]	-0,7068	0,0859	-8,2272	0,0000
R-cuadrado	0,5851			
R^2 ajustado	0,5764			
Función hiperbólica				
Constante [c(1)]	2,9457	0,9912	2,9720	0,0046
Phon/Word [c(2)]	33,9531	4,3794	7,7530	0,0000
R-cuadrado	0,5560			
R^2 ajustado	0,5468			

Las funciones potencial e hiperbólica que surgen de las regresiones del cuadro 1 pueden representarse en un diagrama en el cual cada observación sea un punto en el espacio de fonemas por palabra versus palabras por enunciado. Dicha representación es la que aparece en el gráfico 1, en el cual vemos que la función hiperbólica predice un valor para el número de palabras por enunciado que es siempre mayor que el predicho por la función potencial. Esto genera un mejor ajuste para 24 idiomas (irlandés, vietnamita, hindi, español, amárico, etc.) pero un ajuste peor para los restantes 26 idiomas (inglés, mandarín, apache, birmano, arrernte, etc.).

[9] Estas dos regresiones, al igual que las demás cuyos resultados aparecen en el presente trabajo, fueron corridas utilizando el programa informático Eviews 3.1.

[10] Ambas formulaciones tienen también un ajuste mejor que el que se obtiene bajo una especificación de tipo lineal. Dicha especificación habría producido un coeficiente R^2 igual a 0,5124.

Gráfico 1: Líneas de regresión potencial e hiperbólica



4. Tests de especificación

Para tener una impresión más precisa de las ventajas relativas de cada una de las formulaciones de la ley de Menzerath analizadas en este trabajo, es posible realizar algunos tests estadísticos que sirven para comparar si los resultados de una regresión pueden explicar algunos fenómenos que la otra regresión no explica. Uno de dichos tests es el propuesto por Davidson y MacKinnon (1981), conocido también como "test J". El mismo consiste en correr ecuaciones como las siguientes:

$$Ln(Word/Clause) = c(1) + c(2)*Ln(Phon/Word) + c(3)*Ln(WC2fitted) \quad (5)\ ;$$

$$Word/Clause = c(1) + c(2)*[1/(Phon/Word)] + c(3)*WC1fitted \quad (6)\ ;$$

donde *WC1fitted* y *WC2fitted* son los valores de *Word/Clause* estimados por las regresiones correspondientes a las ecuaciones 3 y 4. La idea detrás de este test es analizar si el comportamiento de la variable dependiente explicado por un modelo puede ayudar a mejorar la estimación bajo el modelo alternativo, y el elemento básico para evaluar esto es la significación estadística de los coeficientes designados bajo el nombre de *c(3)* en las

ecuaciones 5 y 6.

En el cuadro 2 pueden verse los resultados de las regresiones de estas últimas ecuaciones, y en ambos casos se observa que el valor estimado para *c(3)* resulta muy poco significativo ("p = 0,9562" y "p = 0,9701"). Esto indica que los resultados generados por la especificación potencial no pueden ser mejorados de manera apreciable por los factores tenidos en cuenta en la especificación hiperbólica, en tanto que la inversa también es cierta (es decir, los resultados de la especificación hiperbólica tampoco pueden mejorarse usando los factores tenidos en cuenta por la especificación potencial). Más aún, si comparamos los coeficientes R^2 ajustados que aparecen en el cuadro 2 con los informados en el cuadro 1, vemos que en ambos casos dichos coeficientes han descendido, y esto es otra indicación de que los factores adicionales incluidos en las nuevas regresiones no ayudan a mejorar los resultados originales.

Cuadro 2: Resultados de las regresiones para los tests J

Concepto	Coeficiente	Error típico	Estadístico-t	Probabilidad
Función potencial				
Constante [c(1)]	3,9148	9,1936	0,4258	0,6722
Phon/Word [c(2)]	-0,8108	1,8861	-0,4299	0,6693
WC fitted [c(3)]	-0,1489	2,6970	-0,0552	0,9562
R-cuadrado	0,5851			
R^2 ajustado	0,5675			
Función hiperbólica				
Constante [c(1)]	2,6096	8,9764	0,2907	0,7725
Phon/Word [c(2)]	30,1439	101,1876	0,2979	0,7671
WC fitted [c(3)]	0,1143	3,0331	0,0377	0,9701
R-cuadrado	0,5560			
R^2 ajustado	0,5371			

Los test J corridos sobre nuestras dos formulaciones de la ley de Menzerath son ejemplos de "tests no anidados", que consideran a las distintas especificaciones como modelos alternativos a contrastar entre sí. En este caso, sin embargo, también puede pensarse en un "test anidado" que sirva para comparar las dos formulaciones, basado en un modelo más general que incluya las funciones potencial e hipebólica como casos particulares. El más simple de dichos modelos es el siguiente:

$$y = a + b \cdot x^c \qquad (7)\ ;$$

que en nuestro caso puede escribirse como:

Word/Clause = c(1) + c(2)(Phon/Word)^ c(3)* (8) .

A fin de estimar los parámetros de un modelo como este, resulta necesario correr una regresión no lineal como la que aparece expuesta en el cuadro 3. En dicho contexto, la función potencial es un caso particular para el cual se da que "*c(1) = 0*", en tanto que la función hiperbólica es otro caso particular para el cual se da que "*c(3) = -1*". La primera de dichas restricciones puede testearse observando el valor-p del coeficiente bajo análisis (p = 0,6770) y, dado eso, la hipótesis en cuestión no puede ser rechazada para ningún nivel razonable de probabilidad. Para testear la segunda restricción, en cambio, es necesario correr un test de "*c(3) = -1*" como puede ser el denominado "test de Wald". Dicho test nos da como resultado un estadístico ji cuadrado (χ^2) para el cual "p = 0,9817" (y esto tampoco puede rechazarse para ningún nivel razonable de probabilidad).

Cuadro 3: Resultados de una regresión no lineal general por OLS

Concepto	Coeficiente	Error típico	Estadístico-t	Probabilidad
Constante [c(1)]	2,7973	6,6733	0,4192	0,6770
Parámetro multiplicativo [c(2)]	33,6094	15,2603	2,2024	0,0326
Parámetro potencial [c(3)]	-0,9798	0,8789	-1,1148	0,2706
R-cuadrado	0,5560			
R^2 ajustado	0,5371			

Los resultados reproducidos en el cuadro 3 muestran también un coeficiente R^2 ajustado igual a 0,5371. Dicho coeficiente es menor que los coeficientes informados en el cuadro 1, y esto es otra indicación de que tanto la especificación potencial como la especificación hiperbólica son dos modelos que resultan adecuados para explicar los datos, y que un modelo general que los incluya a los dos no es eficiente para mejorar el poder explicativo de ninguna de estas dos formulaciones más simples.

5. Factores filogenéticos y geográficos

Una posible explicación para parte de la variación en el cociente entre palabras y enunciados que no puede ser explicada por las ecuaciones 3, 4 y 8 es la existencia de ciertos factores filogenéticos y geográficos que hagan que la relación funcional entre *Word/Clause* y *Phon/Word* no sea la misma para todos los idiomas. Para tener en cuenta algunos de esos factores, decidimos incluir dos variables categóricas adicionales,

referidas a las cinco regiones en las cuales está dividida nuestra muestra y a las cuatro principales familias lingüísticas representadas en ella. Esto es equivalente a incluir variables binarias para cuatro de las cinco regiones (*Africa*, *America*, *Westasia* y *Eastasia*) y para las cuatro familias principales (*Indoeuropean*, *Afroasiatic*, *Nigercongo* y *Sinotibetan*).

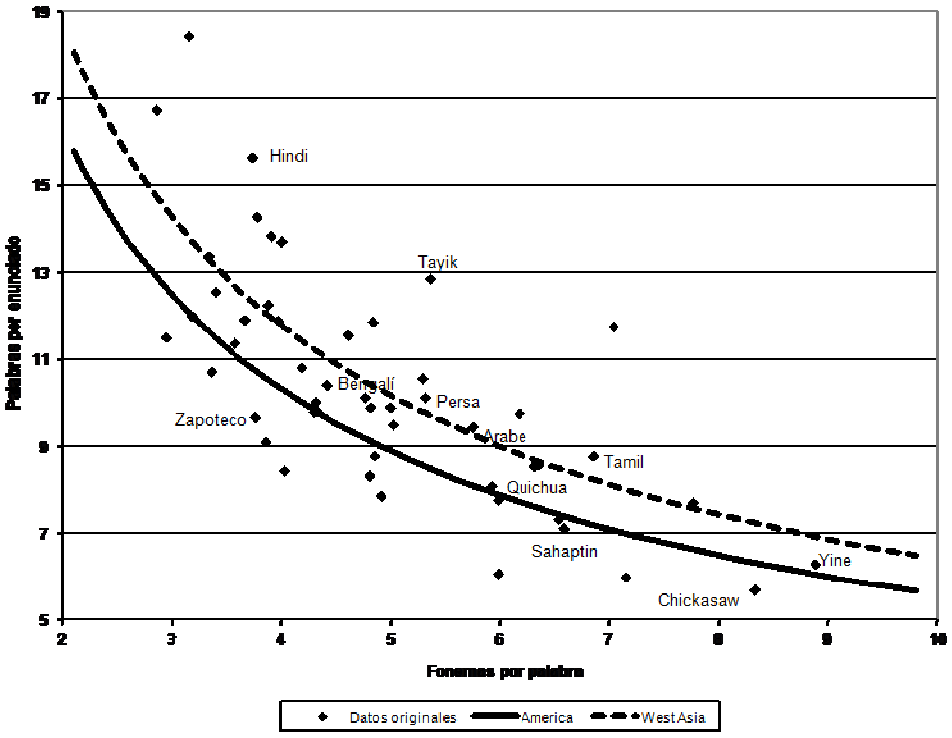
Cuadro 4: Resultados de las regresiones con factores filogenéticos y geográficos

Concepto	Coeficiente	Error típico	Estadístico-t	Probabilidad
Función potencial				
Constant [c(1)]	3,2712	0,1757	18,6133	0,0000
Africa [c(2)]	0,0709	0,1078	0,6578	0,5144
America [c(3)]	-0,0464	0,0982	-0,4719	0,6395
Westasia [c(4)]	0,0353	0,0799	0,4425	0,6605
Eastasia [c(5)]	0,0405	0,1025	0,3948	0,6951
Indoeuropean [c(6)]	0,1267	0,0850	1,4899	0,1441
Afroasiatic [c(7)]	0,0465	0,0945	0,4925	0,6251
Nigercongo [c(8)]	-0,0723	0,1158	-0,6246	0,5358
Sinotibetan [c(9)]	-0,1905	0,1102	-1,7286	0,0916
Phon/Word [c(10)]	-0,6454	0,0995	-6,4892	0,0000
R-cuadrado	0,6882			
R^2 ajustado	0,6181			
Función hiperbólica				
Constant [c(1)]	2,8399	1,4220	1,9971	0,0526
Africa [c(2)]	0,8472	1,1684	0,7251	0,4726
America [c(3)]	-0,3850	1,0598	-0,3633	0,7183
Westasia [c(4)]	0,4717	0,8684	0,5432	0,5900
Eastasia [c(5)]	0,5524	1,1114	0,4970	0,6219
Indoeuropean [c(6)]	1,3684	0,9219	1,4843	0,1456
Afroasiatic [c(7)]	0,3930	1,0242	0,3837	0,7032
Nigercongo [c(8)]	-1,2150	1,2660	-0,9597	0,3430
Sinotibetan [c(9)]	-1,8828	1,1947	-1,5760	0,1229
Phon/Word [c(10)]	32,2315	5,0899	6,3325	0,0000
R-cuadrado	0,6633			
R^2 ajustado	0,5875			

En el cuadro 4 se muestran los resultados de estas nuevas regresiones, corridas bajo la formulación potencial y bajo la formulación hiperbólica. En él vemos que, aunque las variables binarias incluidas son en general poco significativas estadísticamente, las mismas ayudan a mejorar el ajuste de las estimaciones, que ahora pasan a tener coeficientes R^2 ajustados más altos. Así, para la función potencial, el R^2 ajustado pasa de 0,5764 a 0,6181, y para la función hiperbólica pasa de 0,5468 a 0,5875. Esta mejora, sin embargo, no altera el ránking de los coeficientes R^2, ya que el mismo sigue mostrando a la función potencial por encima de la función hiperbólica.

En el gráfico 2 podemos visualizar parte del efecto que tiene incluir variables geográficas para explicar la variación del cociente entre palabras y enunciados. Dicho gráfico exhibe las líneas de regresión potencial para las dos regiones cuyos coeficientes terminan estando más separados entre sí cuando tomamos en cuenta todos los factores relevantes (tanto geográficos como filogenéticos), y él podemos observar que los puntos que representan a los idiomas de Asia occidental (hindi, bengalí, árabe, tamil, etc.) tienden a estar más cerca de la línea superior, en tanto que los puntos que representan a las lenguas amerindias (zapoteco, apache, quichua, chickasaw, etc.) tienden a estar más cerca de la línea inferior.

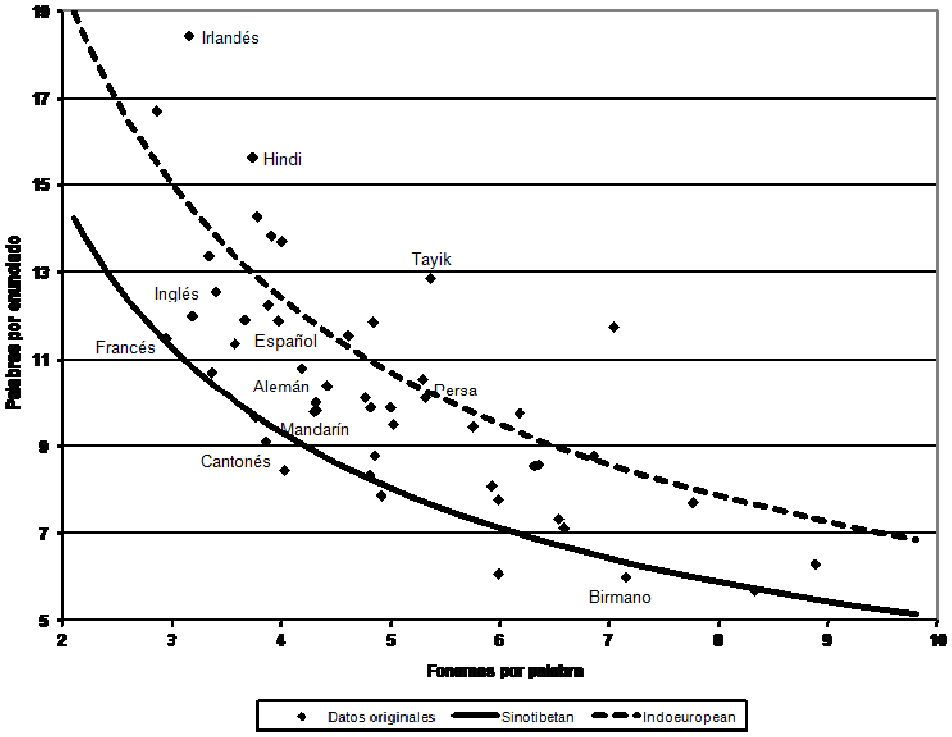
Gráfico 2: Líneas de regresión potencial para distintas regiones



Una observación semejante puede hacerse para los las líneas de regresión representadas en el gráfico 3, que corresponden a las dos familias lingüísticas (indoeuropea y sino-tibetana) que están más lejos una de la otra en el espacio de palabras por enunciado versus fonemas por palabra. En dicho gráfico vemos que, mientras los puntos que representan a las tres lenguas sino-tibetanas incluidas en nuestra muestra

(mandarín, cantonés y birmano) están muy cerca de la línea inferior, la mayoría de los idiomas indoeuropeos (por ejemplo, inglés, español, alemán, persa) están representados por puntos ubicados alrededor de la línea superior.

Gráfico 3: Líneas de regresión potencial para distintas familias



Las nuevas ecuaciones estimadas bajo una formulación potencial y bajo una formulación hiperbólica pueden ser también sometidas a tests J, para ver si los resultados de una regresión sirven para explicar fenómenos que la otra regresión no explica. En este caso, los coeficientes estimados adicionales resultan tener valores-p algo más bajos que los expuestos en la sección anterior, que son iguales a “p = 0,2809” para el coeficiente que mide el efecto de los fenómenos explicados por la función hiperbólica en la ecuación potencial, y a “p = 0,6170” para el coeficiente que mide el efecto de los fenómenos explicados por la función potencial en la ecuación hiperbólica. Dichos valores sin embargo, siguen sin ser estadísticamente significativos para ningún nivel razonable de probabilidad.

El efecto de los factores geográficos y filogenéticos que aparece implícito en

nuestras formulaciones alternativas de la ley de Menzerath parece también reducir la magnitud de la relación entre fonemas por palabra y palabras por enunciado. En efecto, si comparamos los coeficientes que aparecen en el cuadro 4 con los informados en el cuadro 1, vemos que el coeficiente negativo para *Phon/Word* en la función potencial baja de 0,7068 a 0,6454, lo cual implica una reducción del 8,7%. Del mismo modo, la inclusión de factores filogenéticos y geográficos implica una reducción en el coeficiente equivalente en la función hiperbólica de 33,9531 a 32,2315 (es decir, una disminución del 5,1%). No obstante esto, los nuevos coeficientes siguen siendo estadísticamente muy significativos, ya que sus valores de probabilidad son en ambos casos iguales a cero.

6. Variables instrumentales

Cuando uno lleva a cabo una regresión entre dos variables supone implícitamente que la variable incluida en el lado derecho de la ecuación (es decir, la variable independiente) es la que explica el comportamiento de la variable incluida en el lado izquierdo (es decir, la variable dependiente), y no al revés. Eso es una diferencia notable entre el análisis de regresión y el análisis de correlación, ya que la correlación es un concepto simétrico que no supone una relación causal particular entre una variable y la otra con la cual ella se relaciona.

En el caso bajo estudio en el presente trabajo, la lógica de la ley de Menzerath parece indicar que es la naturaleza de los constituyentes del lenguaje (es decir, el número de fonemas por palabra) la que determina la estructura de la categoría de nivel más alto (es decir, del número de palabras por enunciado). Sin embargo, esta causalidad no resulta completamente clara, y puede ser objetada por la idea de que tanto el cociente entre palabras y enunciados como el cociente entre fonemas y palabras podrían ser variables cuyo valor estuviera determinado simultáneamente por un proceso externo.

Para corregir este tipo de problemas de endogeneidad es posible utilizar “variables instrumentales”, es decir, variables que se supone que están relacionadas con la variable independiente pero que tienen la propiedad de estar determinadas exógenamente (o sea, fuera del problema estadístico que estamos analizando). Para este caso particular, hemos elegido seis variables numéricas, cuyos valores provienen de las gramáticas de los diferentes idiomas (y no de los textos usados para computar el número

de palabras por enunciado y el número de fonemas por palabra). Esas variables son el número de consonantes en el inventario de fonemas de cada idioma (*Consonants*), el número de vocales en dicho inventario (*Vowels*), y el número de tonos distintivos que posee cada idioma (*Tones*)[11], junto con el número de géneros distintos que tienen los sustantivos (*Genders*), el número de casos de dichos sustantivos (*Cases*), y la cantidad de inflexiones que presentan los verbos (*Inflections*)[12]. Los valores para las primeras tres variables instrumentales fueron extraídos de las mismas fuentes utilizadas para obtener las distintas versiones de "El viento norte y el sol" (o sea, de las correspondientes ilustraciones del IPA). Para imputar los valores de las otras tres variables, en cambio, usamos la versión electrónica del atlas mundial de estructuras lingüísticas (WALS)[13].

En un caso como este, el procedimiento más simple para incluir las variables instrumentales en la estimación de los coeficientes de las ecuaciones es el que se conoce como "mínimos cuadrados en dos etapas" (2SLS). El mismo consiste en una primera etapa en la cual se corre una regresión entre la variable independiente endógena (en nuestro caso, *Phon/Word*) y todas las variables instrumentales, utilizando mínimos cuadrados ordinarios. Luego viene una segunda etapa en la cual los valores estimados en la primera etapa se incluyen en la estimación de la ecuación que uno quiere verdaderamente llevar a cabo (en nuestro caso, en cada una de las especificaciones de la ley de Menzerath), en lugar de los valores originales de la variable independiente[14].

Para llevar a cabo la primera etapa de este procedimiento, utilizamos una especificación potencial linealizada en la cual relacionamos el logaritmo natural de *Phon/Word* con una constante, con cuatro variables geográficas binarias, con cuatro variables filogenéticas binarias, y con los logaritmos naturales de las seis variables instrumentales adicionales[15]. Luego utilizamos los valores estimados para *Phon/Word* surgidos de dicha regresión para reemplazar los valores originales de dicha variable, en

[11] Este número es igual a uno para los idiomas en los cuales el tono no es distintivo, e igual al número de tonos para el caso de las "lenguas tonales".
[12] El conjunto de datos utilizado para este ejercicio está reproducido en el cuadro que aparece en el apéndice 2.
[13] Véase Dryer y Haspelmath (2013).
[14] Este procedimiento fue propuesto originalmente por Basmann (1957). Para una explicación más completa, véase Davidson y MacKinnon (2003), capítulo 8.
[15] Para esta etapa, intentamos también emplear una especificación lineal, pero finalmente optamos por la especificación logarítmica porque exhibía un mejor ajuste de los datos.

sendas regresiones para las que se emplearon las mismas formulaciones de las ecuaciones 3 y 4.

Los resultados de estas regresiones por mínimos cuadrados en dos etapas aparecen en el cuadro 5. En él vemos que los coeficientes R^2 son en todos los casos menores que los informados en el cuadro 1, pero esto tiene básicamente que ver con que la estimación que utiliza variables instrumentales es siempre menos eficiente que la que emplea las variables originales, si bien es en general más consistente que aquella (es decir, estima valores para los parámetros que están supuestamente más cerca de los que uno obtendría si pudiera conocer el conjunto completo de datos que genera el proceso bajo estudio). En este caso, los resultados obtenidos están en línea con los hallados en las secciones anteriores, en el sentido de que los coeficientes estimados son significativos e implican una relación negativa entre fonemas por palabra y palabras por enunciado. Una vez más, la función potencial exhibe una leve ventaja respecto de la función hiperbólica, tanto en la comparación entre coeficientes R^2 estándar como en la comparación entre coeficientes R^2 ajustados.

Cuadro 5: Resultados de las regresiones por mínimos cuadrados en dos etapas

Concepto	Coeficiente	Error típico	Estadístico-t	Probabilidad
Función potencial				
Constante [c(1)]	3,4907	0,1783	19,5820	0,0000
Phon/Word [c(2)]	-0,7604	0,1132	-6,7192	0,0000
R-cuadrado	0,4847			
R^2 ajustado	0,4740			
Función hiperbólica				
Constante [c(1)]	2,6718	1,2175	2,1944	0,0331
Phon/Word [c(2)]	35,5499	5,4721	6,4966	0,0000
R-cuadrado	0,4679			
R^2 ajustado	0,4568			

El mismo procedimiento de estimación por mínimos cuadrados en dos etapas puede usarse también para las versiones más complejas de nuestro modelo (es decir, para las que incluyen como variables adicionales a los factores filogenéticos y geográficos). Al hacer eso, nos encontramos con nuevos valores para los coeficientes correspondientes a la variable *Phon/Word* en nuestras dos formulaciones alternativas de la ley de Menzerath, los cuales no resultan ser muy distintos de los reportados en el cuadro 4. El ránking de los coeficientes R^2 tampoco se modifica en este caso, ya que la especificación

potencial sigue teniendo un mejor ajuste que la especificación hiperbólica.

El procedimiento de mínimos cuadrados en dos etapas puede también utilizarse para correr una regresión de la formulación general de nuestro modelo, que anide en ella tanto a la función potencial como a la hiperbólica. Dicha regresión fue llevada a cabo empleando la misma serie de valores estimados para *Phon/Word* que usamos para las regresiones del cuadro 5, y sus resultados son los que aparecen en el cuadro 6. Dichos resultados son bastante distintos de los informados en el cuadro 3 (es decir, de los que se obtienen de hacer la misma regresión usando mínimos cuadrados ordinarios), pero los tests de hipótesis que se pueden hacer con ellos acerca de la razonabilidad de nuestros dos modelos competitivos siguen dando esencialmente lo mismo. Esto implica que la restricción asociada con el modelo potencial (*c(1) = 0*) y la asociada con el modelo hiperbólico (*c(3) = -1*) toman valores de probabilidad que no son significativamente distintos de cero a ningún nivel razonable. En este caso, además, el valor-p para la primera de estas restricciones (p = 0,9560) es más alto que el valor-p correspondiente a la segunda restricción (p = 0,7952), y esto señala una vez más que la función potencial parece tener un poder explicativo algo mayor que la función hiperbólica.

Cuadro 6: Resultados de una regresión no lineal general por 2SLS

Concepto	Coeficiente	Error típico	Estadístico-t	Probabilidad
Constante [c(1)]	-1,2608	22,7220	-0,0555	0,9560
Parámetro multiplicativo [c(2)]	32,3550	5,0320	6,4300	0,0000
Parámetro potencial [c(3)]	-0,6647	1,2845	-0,5175	0,6072
R-cuadrado	0,4684			
R^2 ajustado	0,4458			

Este resultado general también aparece cuando llevamos a cabo tests J para las estimaciones de las distintas ecuaciones de la ley de Menzerath usando 2SLS. Por ejemplo, cuando incluimos a los factores geográficos y filogenéticos como variables dependientes adicionales, no nos aparece ninguna mejora apreciable en el ajuste de las regresiones. Lo que sí ocurre en este caso es que el valor-p para el coeficiente correspondiente a las variables *WC1fitted* y *WC2fitted* nos da un poco menor (es decir, un poco más significativo) cuando incluimos los valores ajustados de la especificación potencial en la ecuación hiperbólica (p = 0,2044) que cuando incluimos los valores ajustados de la especificación hiperbólica en la ecuación potencial (p = 0,6067).

7. Consideraciones finales

La principal conclusión que puede extaerse de los distintos análisis llevados a cabo en este estudio es que, para nuestra base de datos construida para medir la relación entre fonemas por palabra y palabras por enunciado en un contexto interlingüístico, no hay evidencia de que la especificación hiperbólica de la ley de Menzerath (Milicka, 2014) ajuste mejor los datos que la especificación potencial tradicional (Altmann, 1980).

Ambas formas de expresar la ley de Menzerath, sin embargo, parecen ser relativamente buenas para explicar los datos bajo estudio, ya que la alta correlación negativa que existe entre fonemas por palabra y palabras por enunciado queda bien explicada tanto por una función potencial como por una función hiperbólica. En todos los casos, los coeficientes obtenidos para el cociente entre fonemas y palabras como una variable explicativa del cociente entre palabras y enunciados tienen el signo esperado, y son también significativamente distintas de cero a un nivel de probabilidad del 1%.

La función potencial, sin embargo, genera siempre mayores coeficientes de determinación que la función hiperbólica. Esta ventaja en el ajuste de los datos aparece tanto cuando usamos la especificación más simple de nuestro modelo (es decir, cuando corremos regresiones por mínimos cuadrados ordinarios en las que *Word/Clause* es solo función de *Phon/Word*), como cuando incluimos factores filogenéticos y geográficos, y cuando utilizamos variables instrumentales (consonantes, vocales, tonos, géneros, casos, inflexiones verbales) para corregir posibles problemas de endogeneidad.

La mayoría de los tests realizados para evaluar los méritos relativos de las funciones potenciales e hiperbólicas muestran además que la variación en el cociente entre palabras y enunciados que queda sin explicar utilizando cierta especificación permanece también inexplicada cuando empleamos la especificación alternativa. Dichos tests nos muestran también que, cuando anidamos los dos modelos dentro de una formulación más general, los parámetros adicionales se vuelven estadísticamente insignificantes. Ese resultado se ve con mayor claridad cuando uno testea separadamente cada modelo como una alternativa restringida del modelo más general. De todos modos, la función potencial sigue siendo siempre una alternativa algo mejor que la función hiperbólica, y esa impresión se refuerza cuando corremos los tests de especificación en

un contexto de estimación por mínimos cuadrados en dos etapas (es decir, cuando corregimos la posible endogeneidad del cociente entre fonemas y palabras como una variable explicativa del cociente entre palabras y enunciados).

Referencias bibliográficas

Altmann, Gabriel (1980). Prolegomena to Menzerath's Law. *Glottometrika* 2: 1-10.

Basmann, Robert (1957). A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation. *Econometrica* 25: 77–83.

Boroda, Moisei & Gabriel Altmann (1991). Menzerath's Law in Musical Texts. *Musikometrica* 3: 1-13.

Coloma, Germán (2014). La existencia de correlación negativa entre distintos aspectos de la complejidad de los idiomas, Documento de Trabajo Nro 536, Universidad del CEMA.

Cramer, Irene (2005). The Parameters of the Menzerath-Altmann Law. *Journal of Quantitative Linguistics* 12: 41-52.

Davidson, Russell & James MacKinnon (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. *Econometrica* 49: 781-793.

Davidson, Russell & James MacKinnon (2003). *Econometric Theory and Methods*. Nueva York: Oxford University Press.

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

Eroglu, Sertac (2013). Menzerath-Altmann Law for Distinct Word Distribution Analysis in a Large Text. *Physica A* 392: 2775-2780.

Eroglu, Sertac (2014). Menzerath-Altmann Law: Statistical Mechanical Interpretation as Applied to a Linguistic Organization. *Journal of Statistical Physics* 157: 392-405.

Fenk-Oczlon, Gertraud & August Fenk (1999). Cognition, Quantitative Linguistics and Systemic Typology. *Linguistic Typology* 3: 151-177.

Ferrer, Ramón & Nuria Forns (2010). The Self-Organization of Genomes. *Complexity* 15: 34-36.

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

Kohler, Reinhard (1984). Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika* 6: 177-183.

Kulacka, Agnieszka (2010). The Coefficients in the Formula for the Menzerath-Altmann Law. *Journal of Quantitative Linguistics* 17: 257-268.

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

Milicka, Jiri (2014). Menzerath's Law: The Whole is Greater than the Sum of its Parts. *Journal of Quantitative Linguistics* 21: 85-99.

Teupenhayn, Regina & Gabriel Altmann (1984). Clause Length and Menzerath's Law. *Glottometrika* 6: 127-138.

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.
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.
Dawd, Abushush & Richard Hayward (2002). Nara. *Journal of the International Phonetic Association* 32: 249-255.
DiCanio, Christian (2010). Itunyoso Trique. *Journal of the International Phonetic Association* 40: 227-238.
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.
Hargus, Sharon & Virginia Beavert (2014). Northwest Sahaptin. *Journal of the International Phonetic Association* 44: 320-342.
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.
Ido, Shinji (2014). Bukharan Tajik. *Journal of the International Phonetic Association* 44: 87-102.
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.
Khatiwada, Rajesh (2009). Nepali. *Journal of the International Phonetic Association* 39: 373-380.
Kirby, James (2011). Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association* 41: 381-392.
Kohler, Klaus (1999). German. En IPA (1999), 86-89.
Laufer, Asher (1999). Hebrew. En IPA (1999), 96-99.
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.
Marlett, Stephen, Xavier Moreno & Genaro Herrera (2005). Seri. *Journal of the International Phonetic Association* 35: 117-121.
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.
Ní Chasaide, Ailbhe (1999). Irish. En IPA (1999), 111-116.
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.
Padayodi, Cécile (2008). Kabiye. *Journal of the International Phonetic Association* 38: 215-221.
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.
Ridouane, Rachid (2014). Tashlhiyt Berber. *Journal of the International Phonetic Association* 44: 207-221.
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 (1999). 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.
Tuttle, Siri & Merton Sandoval (2002). Jicarilla Apache. *Journal of the International Phonetic Association* 32: 105-112.
Urquía, Rittma & Stephen Marlett (2008). Yine. *Journal of the International Phonetic Association* 38: 365-369.
Valenzuela, Pilar & Carlos Gussenhoven (2013). Shiwilu (Jebero). *Journal of the International Phonetic Association* 43: 97-106.
Watkins, Justin (2001). Burmese. *Journal of the International Phonetic Association* 31: 291-295.
Zee, Eric (1999). Chinese (Hong Kong Cantonese). En IPA (1999), 58-60.
Zimmer, Karl & Orhan Orgun (1999). Turkish. En IPA (1999), 154-156.

Apéndice 1: Base de datos de “El viento norte y el sol”

Idioma	Ubicación	Familia	Fonemas	Palabras	Enunciados	Phon/Word	Word/Clause
Alemán	Europe	Indoeuropea	452	108	10	4.19	10.80
Amárico	Africa	Afroasiática	661	94	8	7.03	11.75
Apache	America	Na-Dené	579	118	15	4.91	7.87
Árabe	West Asia	Afroasiática	488	85	9	5.74	9.44
Arrernte	East Asia	Pama-Nyungan	436	73	12	5.97	6.08
Bengalí	West Asia	Indoeuropea	459	104	10	4.41	10.40
Bereber	Africa	Afroasiática	306	76	9	4.03	8.44
Birmano	East Asia	Sino-Tibetana	300	42	7	7.14	6.00
Cantonés	East Asia	Sino-Tibetana	351	91	10	3.86	9.10
Chickasaw	America	Muskogeana	474	57	10	8.32	5.70
Coreano	East Asia	Coreánica	381	60	7	6.35	8.57
Dinka	Africa	Nilo-Sahara	548	137	10	4.00	13.70
Español	Europe	Indoeuropea	425	107	9	3.97	11.89
Francés	Europe	Indoeuropea	343	108	9	3.18	12.00
Georgiano	West Asia	Caucásica	418	70	9	5.97	7.78
Griego	Europe	Indoeuropea	479	104	9	4.61	11.56
Hausa	Africa	Afroasiática	648	166	12	3.90	13.83
Hebreo	West Asia	Afroasiática	526	89	11	5.91	8.09
Hindi	West Asia	Indoeuropea	467	125	8	3.74	15.63
Húngaro	Europe	Urálica	431	100	10	4.31	10.00
Igbo	Africa	Níger-Congo	356	107	8	3.33	13.38
Inglés	Europe	Indoeuropea	383	113	9	3.39	12.56
Irlandés	Europe	Indoeuropea	406	129	7	3.15	18.43
Japonés	East Asia	Japónica	444	89	9	4.99	9.89
Kabiye	Africa	Níger-Congo	433	91	9	4.76	10.11
Malayo	East Asia	Austronesia	481	78	8	6.17	9.75
Mandarín	East Asia	Sino-Tibetana	421	98	10	4.30	9.80
Mapuche	America	Araucana	360	75	9	4.80	8.33
Mono	Africa	Níger-Congo	338	115	10	2.94	11.50
Nara	Africa	Nilo-Sahara	466	108	11	4.31	9.82
Nepalí	West Asia	Indoeuropea	502	95	9	5.28	10.56
Persa	West Asia	Indoeuropea	483	91	9	5.31	10.11
Polaco	Europe	Indoeuropea	428	89	9	4.81	9.89
Portugués	Europe	Indoeuropea	380	98	8	3.88	12.25
Quichua	America	Quechua	593	94	11	6.31	8.55
Sahaptin	America	Penutiana	375	57	8	6.58	7.13
Sandawe	Africa	Khoisan	383	79	9	4.85	8.78
Seri	America	Hokan	593	157	11	3.78	14.27
Shiwilu	America	Kawapana	837	108	14	7.75	7.71
Tailandés	East Asia	Tai-Kadai	480	131	11	3.66	11.91
Tamil	West Asia	Dravídica	541	79	9	6.85	8.78
Tausug	East Asia	Austronesia	572	114	12	5.02	9.50
Tayik	West Asia	Indoeuropea	482	90	7	5.36	12.86
Temne	Africa	Níger-Congo	446	125	11	3.57	11.36
Trique	America	Otomangueana	359	107	10	3.36	10.70
Turco	West Asia	Túrquica	431	66	9	6.53	7.33
Vasco	Europe	Vascónica	401	83	7	4.83	11.86
Vietnamita	East Asia	Austroasiática	334	117	7	2.85	16.71
Yine	America	Arahuaca	559	63	10	8.87	6.30
Zapoteco	America	Otomangueana	327	87	9	3.76	9.67
Promedio			**455.32**	**96.94**	**9.48**	**4.94**	**10.37**

Apéndice 2: Variables instrumentales

Idioma	Ubicación	Consonantes	Vocales	Tonos	Casos	Géneros	Inflexiones
Alemán	Europe	23	15	1	4	3	2
Amárico	Africa	27	7	1	2	2	6
Apache	America	33	8	3	1	1	5
Árabe	West Asia	29	6	1	1	2	6
Arrernte	East Asia	27	4	1	8	1	4
Bengalí	West Asia	29	7	1	6	2	2
Bereber	Africa	34	3	1	2	2	6
Birmano	East Asia	34	9	4	8	1	2
Cantonés	East Asia	19	11	6	1	1	1
Chickasaw	America	16	9	1	2	1	6
Coreano	East Asia	19	18	1	6	1	6
Dinka	Africa	20	7	4	1	1	6
Español	Europe	19	5	1	1	2	4
Francés	Europe	20	13	1	1	2	4
Georgiano	West Asia	28	5	1	6	1	8
Griego	Europe	18	5	1	3	3	4
Hausa	Africa	28	10	2	1	2	6
Hebreo	West Asia	25	5	1	1	2	4
Hindi	West Asia	34	11	1	2	2	2
Húngaro	Europe	25	14	1	10	1	4
Igbo	Africa	26	8	3	1	1	6
Inglés	Europe	24	11	1	2	1	2
Irlandés	Europe	35	11	1	2	2	2
Japonés	East Asia	16	5	2	8	1	4
Kabiye	Africa	21	9	2	1	1	2
Malayo	East Asia	18	6	1	1	1	4
Mandarín	East Asia	19	6	4	1	1	1
Mapuche	America	22	6	1	2	1	8
Mono	Africa	32	8	3	1	5	6
Nara	Africa	25	10	2	5	2	4
Nepalí	West Asia	27	11	1	2	1	4
Persa	West Asia	23	6	1	2	1	4
Polaco	Europe	31	6	1	6	3	4
Portugués	Europe	19	13	1	1	2	4
Quichua	America	23	3	1	8	1	8
Sahaptin	America	32	7	1	4	1	10
Sandawe	Africa	44	15	2	1	5	8
Seri	America	18	8	1	1	1	5
Shiwilu	America	17	4	1	6	1	6
Tailandés	East Asia	21	9	5	1	1	2
Tamil	West Asia	15	10	1	6	3	2
Tausug	East Asia	17	3	1	1	1	4
Tayik	West Asia	22	6	1	2	1	4
Temne	Africa	19	9	2	1	5	2
Trique	America	29	8	9	1	1	6
Turco	West Asia	22	8	1	6	1	6
Vasco	Europe	23	5	1	10	1	4
Vietnamita	East Asia	22	11	8	1	1	1
Yine	America	16	5	1	2	4	6
Zapoteco	America	20	5	3	1	1	8
Promedio		**24.10**	**8.08**	**1.94**	**3.08**	**1.70**	**4.50**