UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

UN MODELO ESTADÍSTICO DE ECUACIONES SIMULTÁNEAS SOBRE LA INTERACCIÓN DE VARIABLES FONOLÓGICAS

Germán Coloma

Septiembre 2013
Nro. 519

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>

UN MODELO ESTADÍSTICO DE ECUACIONES SIMULTÁNEAS SOBRE LA INTERACCIÓN DE VARIABLES FONOLÓGICAS

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

Resumen

Este trabajo propone un método para analizar bases de datos lingüísticas utilizando regresiones con sistemas de ecuaciones simultáneas, y aplica dicho método a una base de datos de 100 idiomas y cuatro variables fonológicas (número de consonantes, número de vocales, distinción por acento y distinción por tono). El método resulta adecuado para replicar los coeficientes de correlación más significativos de la base de datos, y para resolver una contradicción acerca del signo de algunos de esos coeficientes. Con la ayuda de este método también encontramos cierta evidencia de un "fenómeno compensatorio" entre variables fonológicas, y también terminamos proponiendo dos modelos estadísticos generales de estructura fonológica de las lenguas. Según el primero de ellos, las consonantes, las vocales y el tono están negativamente correlacionados con la existencia de distinciones por acento. De acuerdo con el segundo modelo, el tono está negativamente correlacionado con el acento, que a su vez está negativamente correlacionado con el número de consonantes, que a su vez está negativamente correlacionado con el número de vocales, que a su vez está negativamente correlacionado con la existencia de distinción entre tonos.

Palabras clave: ecuaciones simultáneas, regresión, variables fonológicas, compensación, tipología lingüística.

1. Introducción

El objetivo de este trabajo es proponer un método para analizar bases de datos lingüísticas utilizando análisis de regresión estadístico. Este método se basa en correr sistemas de ecuaciones simultáneas, y su propósito es encontrar una serie de relaciones entre distintas variables que sean parte de un mismo fenómeno lingüístico.

Para ilustrar este método hemos construido una base de datos de 100 idiomas distintos con cuatro grandes características fonológicas (número de fonemas consonánticos, número de fonemas vocálicos, distinción por acento y distinción por tono). Después de calcular los principales estadísticos descriptivos de dicha base de datos (incluyendo los coeficientes de correlación entre las variables), le aplicamos el método de ecuaciones simultáneas, para ver si las relaciones halladas cuando cada variable se

[*] Las opiniones son personales y no representan necesariamente las de la Universidad del CEMA.

correlacionaba con otra permanecían si uno consideraba las posibles interacciones entre las cuatro variables. Después de eso, refinamos el método en busca de posibles "relaciones estructurales" entre las variables, usando un sistema recursivo en el cual cada variable influencia a otra y es a su vez influenciada por una tercera variable.

El resto del artículo se organiza del siguiente modo. En la sección 2 explicamos la lógica básica de la regresión de sistemas de ecuaciones simultáneas, y su posible uso para analizar bases de datos lingüísticos. En la sección 3 reseñamos parte de la literatura sobre análisis multilingüísticos de variables fonológicas, y resumimos los principales resultados que dicha literatura ha obtenido utilizando distintas metodologías. En la sección 4 describimos la base de datos construida por nosotros y calculamos sus principales estadísticos descriptivos. En la sección 5 aplicamos el método de regresión de ecuaciones simultáneas a la base de datos, y encontramos algunas relaciones que son consistentes con los estadísticos descriptivos de la sección 4. También construimos un sistema recursivo de ecuaciones, el cual nos sirve para resolver una contradicción que tiene que ver con el signo de algunas correlaciones halladas previamente, y que resultaban contraintuitivas. Finalmente, en la sección 6 elaboramos algunas conclusiones acerca de todo el trabajo.

2. El método de regresión con ecuaciones simultáneas

En el campo de la estadística, una regresión es un método por el cual el comportamiento de una variable se compara con el de una o más variables adicionales para detectar si existen relaciones entre ellas que ayuden a explicar un proceso natural que ligue las variables bajo análisis. Desarrollado originalmente para aplicaciones en las ciencias biológicas, el análisis de regresión ha sido utilizado de manera muy frecuente desde hace muchos años en relación a innumerables problemas de las ciencias físicas y sociales, así como también en distintas ramas de la lingüística tales como fonética, fonología, sociolingüística, etc [1].

La idea básica de una regresión es analizar varias variables que se supone que, de manera conjunta, explican determinado fenómeno, y ver si esa explicación puede ser medida a través de coeficientes que relacionan esas variables con otra que refleje la

[1] Para ilustraciones de estas aplicaciones, véase Johnson (2008), capítulo 3.

intensidad del fenómeno bajo estudio. Supongamos, por ejemplo, que queremos testear una teoría que postula que el número de fonemas consonánticos de un idioma está inversamente relacionado en el número de fonemas vocálicos. Supongamos que nuestra teoría predice además que el número de fonemas consonánticos también se podría relacionar con otras variables adicionales, como pueden ser su estructura en términos de acento y de tono. En un caso como ese, podríamos correr una regresión con la siguiente forma lineal:

Consonantes = c(1) +c(2)*Vocales +c(3)*Acento +c(4)*Tono (1) ;

donde *Consonantes* es la cantidad de fonemas consonánticos que corresponde a cada una de las lenguas que estamos analizando, *Vocales* es la cantidad de fonemas vocálicos, *Acento* es una variable categórica cuyo valor es igual a uno si el idioma usa al acento como un modo de distinguir entre palabras, y *Tono* es otra variable categórica cuyo valor es igual a uno si el idioma es tonal (e igual a cero en caso contrario). En un contexto como ese, *c(1), c(2), c(3)* y *c(4)* son los coeficientes a estimar en el análisis de regresión, y sus valores son las mejores aproximaciones lineales para las relaciones que existen entre la variable *Consonantes* y las tres variables explicativas postuladas.

El típico uso del análisis de regresión, ejemplificado mediante la ecuación (1), ocurre cuando uno postula ciertas relaciones que van en una "dirección particular", es decir, cuando se cree que varias variables (en este caso, *Vocales*, *Acento* y *Tono*) tienen algún poder explicativo acerca del comportamiento de otra variable (en este caso, *Consonantes*). En muchas situaciones, sin embargo, uno podría pensar que, así como las variables *Vocales, Acento* y *Tono* pueden tener algún poder para explicar parcialmente el número de fonemas consonánticos de un idioma, también puede haber alguna relación inversa que vaya desde *Consonantes* hacia variables como *Vocales*, *Acento* y *Tono*. En esas situaciones, la metodología ordinaria de correr regresiones tales como la que aparece en la ecuación (1) puede perder eficacia, ya que ahora no estamos seguros si la relación que encontremos implicará que *Vocales*, *Acento* y *Tono* son variables que explican el comportamiento de la variable *Consonantes*, o si en realidad es esta última variable la que tiene algo que ver con una posible explicación relacionada con la ocurrencia de fenómenos tales como el número de fonemas vocálicos o la distinción entre palabras según su acentuación o su tono en cierta muestra de idiomas.

Una forma de encarar problemas como ese es chequear primero si existe alguna clase de relación entre las variables involucradas, y la forma más simple de hacer eso es calcular los llamados "coeficientes de correlación" entre las variables. Dichos coeficientes son por definición simétricos (o sea, la correlación entre *Consonantes* y *Vocales* es la misma que la correlación entre *Vocales* y *Consonantes*) y en su cálculo no se utiliza ninguna información que tenga que ver con la relación entre cada par de variables y una o más variables diferentes. Los coeficientes de correlación, sin embargo, pueden calcularse para cualquier par de variables respecto del cual tengamos información (así, en el ejemplo previo se pueden calcular seis coeficientes distintos, que corresponden a los pares *Consonantes/Vocales*, *Consonantes/Acento*, *Vocales/Acento*, *Vocales/Tono*, *Consonantes/Tono* y *Acento/Tono*).

El análisis de correlación, sin embargo, tiene una deficiencia importante en comparación con el análisis de regresión, ya que no permite controlar por factores relacionados con la interacción de varias variables. Una forma de resolver este problema es correr una regresión con más de una ecuación al mismo tiempo. Esta es precisamente la idea del método de regresión de ecuaciones simultáneas, que busca calcular un conjunto de coeficientes que correspondan a varias ecuaciones que se corren al mismo tiempo [2].

En este trabajo mostraremos varios resultados relacionados con las interacciones entre las cuatro variables fonológicas definidas en los párrafos anteriores. Una posibilidad para capturar las relaciones entre esas variables es correr un sistema de ecuaciones del siguiente tipo:

Consonantes = c(11) +c(2)*Vocales +c(3)*Acento +c(4)*Tono (2) ;
Vocales = c(21) +c(12)*Consonantes +c(5)*Acento +c(6)*Tono (3) ;
Acento = c(31) +c(13)*Consonantes +c(15)*Vocales +c(7)*Tono (4) ;
Tono = c(41) +c(14)*Consonantes +c(16)*Vocales +c(17)*Acento (5) ;

donde los coeficientes a estimar son *c(11)*, *c(21)*, *c(31)* y *c(41)* (que son las constantes de las cuatro ecuaciones), y *c(2)*, *c(3)*, *c(4)*, *c(5)*, *c(6)*, *c(7)*, *c(12)*, *c(13)*, *c(14)*, *c(15)*, *c(16)* y *c(17)* (que son las "pendientes" de cada variable respecto de cada una de las otras

[2] Para una explicación de la lógica detrás de este método, véase Kennedy (2008), capítulo 10.

variables incluidas en el análisis de regresión).

El uso de ecuaciones simultáneas permite la introducción de varios procedimientos que el análisis de regresión uniecuacional no puede utilizar. El más importante es el uso de coeficientes de correlación entre los residuos de las ecuaciones. Esto implica que, cuando estimamos una ecuación, estamos al mismo tiempo utilizando información de los resultados que obtenemos al estimar las otras ecuaciones, y dicha información puede ser útil para mejorar la precisión y la eficiencia estadística de los coeficientes a estimar [3].

Otra ventaja de usar ecuaciones simultáneas es la posibilidad de introducir relaciones entre los coeficientes de distintas ecuaciones. Supongamos, por ejemplo, que creemos que las relaciones entre las variables son tales que el cambio en una variable inducido por otra tiene que ser igual a la inversa del cambio en la segunda variable inducido por la primera (este podría ser el caso si estamos estimando "ecuaciones de equilibrio" en las que, por ejemplo, el efecto de *Consonantes* sobre *Vocales* es igual a la inversa del efecto de *Vocales* sobre *Consonantes*). En ese caso, nuestro sistema podría modificarse del siguiente modo:

Consonantes = c(11) +c(2)*Vocales +c(3)*Acento +c(4)*Tono (6) ;
Vocales = c(21) +(1/c(2))*Consonantes +c(5)*Acento +c(6)*Tono (7) ;
Acento = c(31) +(1/c(3))*Consonantes +(1/c(5))*Vocales +c(7)*Tono (8) ;
Tono = c(41) +(1/c(4))*Consonantes +(1/c(6))*Vocales +(1/c(7))*Acento (9) ;

lo cual equivale a correr el sistema descripto por las ecuaciones (2)/(5) imponiéndole las restricciones "*c(12) = 1/c(2)*", "*c(13) = 1/c(3)*", "*c(14) = 1/c(4)*", "*c(15) = 1/c(5)*", "*c(16) = 1/c(6)*" y "*c(17) = 1/c(7)*".

El uso de ecuaciones simultáneas también es bueno para incorporar una característica que es común en muchos problemas estadísticos, que es la "endogeneidad" de ciertas variables. Esto tiene que ver con que, si una variable (por ejemplo, *Consonantes*) depende del valor de otra variable (por ejemplo, *Vocales*) pero también ocurre que la segunda variable depende de la primera, entonces ninguna de ellas está

[3] El procedimiento que hace uso de estos coeficientes de correlación es conocido como "método de las regresiones aparentemente no relacionados" (SUR, por su sigla en inglés), y es el que utilizaremos en nuestras estimaciones de la sección 5. Para una explicación sobre la lógica detrás de este método, véase Greene (2011), capítulo 10.

determinada verdaderamente por la otra, sino que ambas están determinadas simultáneamente por un proceso gobernado por un entorno preestablecido. Para incorporar estos problemas de endogeneidad, tenemos que usar "variables instrumentales". Estas son variables que se supone que están relacionadas con las variables endógenas bajo análisis, pero que tienen la propiedad de estar determinadas exógenamente (es decir, fuera del problema estadístico que estamos analizando). En el caso de las cuatro variables fonológicas de los sistemas (2)/(5) y (6)/(9), podemos pensar que *Consonantes*, *Vocales*, *Acento* y *Tono* son todas variables endógenas, y podemos "instrumentarlas" utilizando variables relacionadas con la inclusión de los distintos idiomas de nuestra muestra en diferentes familias o áreas geográficas. Esas variables podrían ser variables categóricas que tomaran un valor igual a uno cuando cierta observación pertenece a un grupo particular (por ejemplo, cuando corresponde a un idioma sino-tibetano, o a un idioma sudamericano) y un valor igual a cero en caso contrario. Con la inclusión de estas variables, cada variable endógena queda reemplazada por una función de un conjunto de variables exógenas.

El procedimiento descripto en el párrafo anterior, conocido como "estimación en dos etapas", puede combinarse con el uso de coeficientes de correlación entre los residuos de las ecuaciones. Si hacemos eso, agregamos una etapa adicional a la estrategia de estimación, y por eso es que el procedimiento como un todo se conoce como "estimación en tres etapas" [4].

La endogeneidad, sin embargo, también puede estar ligada a procesos que relacionen variables de un modo más particular o "estructural". Podría ser posible que una variable (por ejemplo, el número de fonemas consonánticos) tuviera un papel en la determinación del valor de otra variable (por ejemplo, el número de fonemas vocálicos), pero que esa segunda variable fuera a su vez el principal factor explicativo de una tercera variable (por ejemplo, la probabilidad de que el tono sea un elemento distintivo en determinada lengua). El proceso podría inclusive continuar, ya que la tercera variable podría ser parte de la explicación del nivel de una cuarta variable (por ejemplo, la probabilidad de que el acento sea un elemento distintivo), y esa variable podría por su parte influir en el nivel del primer fenómeno analizado (que en nuestro ejemplo es el

[4] Véase Kennedy (2008), capítulo 10, o Greene (2011), capítulo 10.

número de fonemas consonánticos). Una situación como esa genera una especie de "sistema recursivo", en el cual las distintas variables endógenas se relacionan de cierta manera específica. El caso descripto más arriba, por ejemplo, puede representarse a través del siguiente conjunto de ecuaciones:

$$\text{Consonantes} = c(11) + c(2)*\text{Acento} \quad (10)\ ;$$

$$\text{Vocales} = c(21) + c(3)*\text{Consonantes} \quad (11)\ ;$$

$$\text{Acento} = c(31) + c(4)*\text{Tono} \quad (12)\ ;$$

$$\text{Tono} = c(41) + c(5)*\text{Vocales} \quad (13)\ .$$

Por supuesto, el sistema formado por las ecuaciones (10)/(13) es solo uno de los muchos ejemplos posibles de sistemas recursivos que pueden construirse utilizando las cuatro variables analizadas en nuestro ejemplo. Seis de esos sistemas tienen las mismas propiedades estructurales (o sea, cada variable como determinante de una segunda, y determinada a su vez por una tercera, sin repeticiones).

3. El análisis multilingüístico de variables fonológicas

El problema descripto en la sección anterior para ilustrar el posible uso de un método de regresión de ecuaciones simultáneas es un ejemplo de análisis multilingüístico de variables fonológicas. Que nosotros sepamos, la literatura sobre estos temas no ha utilizado dicho método. Resulta sin embargo útil reseñar algunas de las contribuciones de esa literatura, a fin de apreciar los principales resultados que podrían compararse con los que obtendremos en el presente trabajo.

El enfoque cuantitativo de la tipología fonológica se remonta por lo menos a un artículo de Kramsky (1959), que fue uno de los primeros en analizar una muestra de idiomas para encontrar relaciones entre el número de vocales, el número de consonantes y otras variables tales como el número de fonemas por palabra y la frecuencia relativa de los distintos tipos de sonido. Ese trabajo fue también el primero en hallar cierta correlación negativa entre distintos rasgos fonológicos (que en ese caso fueron la ocurrencia de consonantes oclusivas y la ocurrencia de consonantes líquidas).

La mayoría de las contribuciones de los análisis multilingüísticos de variables fonológicas se han concentrado desde ese momento en descubrir "principios fonológicos

universales" de los idiomas. En lo que respecta a los sistemas vocálicos, por ejemplo, existen muchos trabajos que respaldan la validez de la llamada "teoría de la dispersión", según la cual los sonidos vocálicos tienden a localizarse a la máxima distancia posible unos de otros [5]. En lo que respecta a los sistemas consonánticos, en cambio, la regularidad más importante parece ser el hecho de que los sistemas con pocas consonantes exhiben solo "sonidos básicos", y que las articulaciones más elaboradas y complejas solo aparecen cuando el número de fonemas consonánticos de un idioma se vuelve más grande [6].

El estudio estadístico de principios fonológicos universales también se ha enfocado en algunas relaciones entre las consonantes y las vocales. Marsico y otros (2004), por ejemplo, han explorado la existencia de "redundancia" en los sistemas fonológicos, encontrando que los inventarios fonéticos tienden a maximizar el uso de un conjunto relativamente pequeño de rasgos distintivos sin necesidad de hacer distinciones innecesarias entre los sonidos. Más aún, Coupé, Marsico y Pellegrino (2009) hallaron que no había correlación entre la complejidad de los sistemas consonánticos y la complejidad de los sistemas vocálicos de las distintas lenguas, pero sí encontraron una relación entre dichas medidas de complejidad fonológica y una serie de factores geográficos.

Los factores geográficos también aparecen de manera muy preponderante en la literatura sobre la relación entre complejidad fonológica (medida a través de un índice de fonemas vocálicos, fonemas consonánticos y estructura tonal) y la distancia a un punto en el cual supuestamente se produjo el origen del habla humana. Esta literatura, que comenzó con el trabajo de Atkinson (2011), ha producido una implicancia muy fuerte, relacionada con la posible aparición del lenguaje en el sudoeste de África, y generó también un intenso debate entre especialistas en tipología cuantitativa. Dicho debate sirvió para arrojar luz sobre posibles relaciones entre otras variables (tales como el número de fonemas y el tamaño de la población, la complejidad fonológica y la cantidad de fonemas por palabra, etc) [7].

La tipología cuantitativa también ha estudiado la relación entre el acento y el

[5] Para un buen resumen de esa teoría y un interesante análisis cuantitativo de las relaciones entre el número de vocales y él área abarcada por dichas vocales en el "espacio de formantes", véase Becker-Kristal (2010).
[6] Este resultado fue mencionado originalmente por Lindblom y Maddieson (1988).
[7] Para una serie de análisis acerca de estas relaciones, véase Donohue y Nichols (2011), Wichman, Rama y Holman (2011) y Jager y otros (2011).

tono. Hyman (2009), por ejemplo, encontró que los idiomas con un "sistema tonal" pueden separarse de manera estadísticamente significativa de los idiomas con un "sistema de distinción por el acento", y que el resto de los idiomas no incluidos en ninguna de las dos categorías puede verse como una mezcla de esos dos sistemas canónicos. Maddieson (2007), por otro lado, calculó los índices de correlación entre número de consonantes y número de vocales respecto de una medida de la complejidad del sistema tonal de las lenguas, pero no encontró relaciones significativas entre dichas variables.

La contribución de Maddieson también se enfocó en la posible correlación entre variables fonológicas básicas (consonantes, vocales, tono) y medidas de la complejidad de la estructura silábica de los idiomas, encontrando una correlación positiva entre el número de fonemas consonánticos y la complejidad silábica, y una correlación negativa entre la complejidad silábica y la complejidad tonal. Esta última correlación puede verse como un signo de la existencia de posibles "fenómenos compensatorios", a través de los cuales la complejidad en un subsistema del lenguaje debería estar compensada por la simplicidad en otro subsistema. Estos fenómenos compensatorios han sido también explorados en estudios multilingüísticos que utilizan otras variables además de las puramente fonológicas. Fenk-Oczlon y Fenk (2004), por ejemplo, han hallado relaciones negativas entre el número de fonemas por sílaba y el número de sílabas por palabra, y entre el número de sílabas por palabra y el número de palabras por enunciado. Shosted (2006), en cambio, no encontró ninguna relación clara entre la complejidad fonológica y morfológica de los idiomas que analizó, en el marco de un estudio en el cual la complejidad fonológica está medida a través del número de fonemas por sílaba y la complejidad morfológica está medida a través del grado en el cual los verbos aparecen marcados por distintos tipos de inflexión [8].

Aunque los coeficientes de correlación puedan ser útiles como herramientas descriptivas y como signos de posibles relaciones estructurales entre componentes de los sistemas lingüísticos, los mismos también podrían ser engañosos si las variables estuvieran interrelacionadas con otras características adicionales, tal como hemos mencionado en la sección anterior del presente trabajo. Dryer (2009), por ejemplo, ha

[8] Otro trabajo en el cual tampoco se encontró evidencia de correlación negativa entre estructuras lingüísticas es Silnitsky (2003). Dicho estudio, referido a los idiomas indoeuropeos, no halló ninguna correlación significativa entre las variables fonológicas y las variables gramaticales analizadas.

advertido acerca de posibles “conclusiones no fundamentadas” de las correlaciones tipológicas cuando existen factores geográficos que crean la ilusión de una relación entre dos características que en rigor no están vinculadas entre sí. En la misma línea de razonamiento, Bickel (2011) ha sostenido que una modelización exitosa de la distribución de las características de las lenguas necesita un enfoque multivariado, y que algunos fenómenos que podrían tener una “tendencia universal” (por ejemplo, el signo de un coeficiente de correlación entre dos variables) podría quedar estadísticamente distorsionado si los fenómenos en cuestión aparecieran mezclados con otro fenómeno que produce la impresión de que dicha tendencia no existe.

En nuestra opinión, el método que proponemos aquí para analizar la interacción entre distintas variables lingüísticas puede ser útil para resolver algunos de los problemas hallados por la tipología lingüística para llevar a cabo análisis multilingüísticos de variables fonológicas. En las próximas secciones veremos así que el método de regresión con ecuaciones simultáneas presenta una serie de posibilidades interesantes para descartar posibles “correlaciones espurias”, tal como quedará ilustrado a través de un ejercicio empírico que relaciona el número de consonantes, el número de vocales, la distinción entre sonidos acentuados y no acentuados, y la estructura tonal de los idiomas.

4. Descripción de los datos

Para ejemplificar la metodología propuesta en la solución del problema descripto en la sección 2, hemos construido una base de datos de 100 observaciones, que corresponden a diferentes idiomas. Dichos idiomas fueron elegidos en base a la disponibilidad de datos confiables, y por eso es que la muestra incluye a todos los idiomas con más de 40 millones de hablantes de primera lengua a nivel mundial. A fin de tomar en cuenta la importancia relativa de los distintos grupos idiomáticos, incluimos también ejemplos de las principales familias de idiomas que aparecen en fuentes tales como Lewis, Simons y Fennig (2013). Terminamos así con 34 idiomas indoeuropeos, 9 idiomas sino-tibetanos, 9 idiomas de la familia Níger-Congo, 7 idiomas afroasiáticos, 7 idiomas altaicos, 6 idiomas austronesios, 4 idiomas dravídicos, 3 idiomas de la familia Tai-Kadai, 2 idiomas austroasiáticos, 2 idiomas urálicos y 2 idiomas caucásicos. También incluimos representantes de varias familias menos importantes, que son sin embargo

significativas por razones históricas o por el número de idiomas que contienen. Esas familias son las que corresponden a los idiomas algonquianos, araucanos, atabascos, esquimo-aleutianos, jaqui, khoisan, mayas, nilo-saharianos, oto-mangueanos, pama-nyungan, papúes, quechuas, tupí, uto-aztecas y vascuences [9].

Cuadro 1: Valores promedio de las variables fonológicas por grupo de idiomas

Grupo	Idiomas	Consonantes	Vocales	Acento	Tono
Indoeuropeo	35	25.86	9.23	54%	0%
Latino	7	21.00	7.57	86%	0%
Germánico	7	21.43	16.00	57%	0%
Eslavo	7	24.57	6.57	71%	0%
Indoario	7	34.86	9.14	0%	0%
Otros	7	27.43	6.86	57%	0%
Níger-Congo	10	34.40	9.00	20%	80%
Atlántico	4	24.50	10.25	25%	75%
Bantú-khoisan	6	41.00	8.17	17%	83%
Sino-tibetano	9	22.22	7.22	0%	100%
Afroasiático	8	25.38	7.63	25%	38%
Altaico	7	21.86	9.43	14%	14%
Austronesio	8	16.13	5.75	25%	13%
Austro-Tai	5	23.40	13.00	0%	80%
Dravídico	4	30.50	11.00	0%	0%
Amerindio	10	20.00	7.20	10%	40%
Norteamericano	6	18.00	8.00	17%	67%
Sudamericano	4	23.00	6.00	0%	0%
Urálico-Caucásico	4	29.75	7.50	50%	0%
Total	100	24.92	8.62	29%	30%

En el cuadro 1 se muestran los valores promedio de las cuatro variables fonológicas que usamos en este estudio, agrupando a los idiomas en varias categorías. Algunas de ellas son propiamente familias idiomáticas (sino-tibetana, dravídica), mientras que otras son el resultado de agrupar distintas familias. Algunas familias importantes se dividen en subgrupos, tales como la familia indoeuropea que incluye a los idiomas latinos (portugués, gallego, español, catalán, francés, italiano y rumano), germánicos (inglés, holandés, alemán, danés, sueco, noruego e islandés), eslavos (ruso,

[9] En la descripción de los idiomas de nuestra muestra tratamos siempre de usar la fuente disponible más confiable. La mayoría de las descripciones proviene así de trabajos que aparecen en IPA (1999), Gary y Rubino (2001), Brown (2006), Comrie (2009), y de artículos publicados por el *Journal of the International Phonetic Association* en su sección denominada "Illustrations of the IPA".

ucraniano, búlgaro, serbio-croata, esloveno, checoslovaco y polaco), indoarios (hindi-urdu, bengalí, punjabi, marathi, gujarati, sindhi y nepalí) y otros idiomas (persa, pashto, kurdo, armenio, griego e irlandés, más la lengua vasca, que si bien no es indoeuropea tiene una larga tradición de contacto con idiomas de esa familia).

También hemos dividido en dos a la familia Níger-Congo, distinguiendo por un lado un grupo atlántico africano (wolof, ewé, igbo y yoruba) y un grupo bantú (lingala, swahili, shona, xhosa y zulú, al que le hemos adicionado un idioma khoisan llamado "khoekhoe"). Los idiomas afroasiáticos, en cambio, aparecen en una única categoría, que incluye tres lenguas semíticas (árabe, amárico y hebreo), dos lenguas cushíticas (oromo y somalí), una lengua chádica (hausa), una lengua del grupo bereber (shilha) y el idioma nilo-sahariano llamado "dinka". Lo mismo ocurre con las lenguas sino-tibetanas incluidas en nuestra muestra, que son siete variedades del idioma chino (mandarín, cantonés, taiwanés, changsha, gan, hakka y wu), más dos lenguas tibeto-birmanas (tibetano y birmano). Los idiomas altaicos incluidos en nuestra muestra, por su parte, son tres idiomas del grupo turco (turco, azerí y uzbeco) y cuatro idiomas orientales (mongol, manchuriano, coreano y japonés).

Las seis lenguas austronesias que hemos seleccionado (malayo-indonesio, javanés, tagalog, malgache, maorí y hawaiano) han sido agrupadas con la lengua papú llamada "skou" y con el idioma australiano (pama-nyungan) llamado "nyangumarta", básicamente por razones de tipo geográfico. Lo mismo ocurre con los idiomas Tai-Kadai (tailandés, lao y kam) y los idiomas austroasiáticos (vietnamita y camboyano), que constituyen una única categoría "Austro-Tai", y con los idiomas caucásicos (georgiano y cabardiano) que han sido agrupados junto con los idiomas urálicos (finlandés y húngaro). En contraposición, las cuatro lenguas dravídicas incluidas en la muestra (tamil, telugu, kannada y malayalam) constituyen una única categoría relativamente homogénea.

Los diez idiomas aborígenes americanos que hemos incluido, finalmente, han sido agrupados en dos categorías, correspondientes a América del Norte y a América del Sur. En el primero de dichos grupos incluimos a los idiomas de las familias algonquiana (cheyenne), atabasca (navajo), esquimo-aleutiana (inuit), maya (yucateco), oto-mangueana (zapoteco) y uto-azteca (nahuatl), en tanto que en el segundo grupo aparecen los idiomas que pertenecen a las familias quechua (cusqueño), jaqui (aymara), tupí

(guaraní) y araucana (mapuche).

Tal como puede observarse en el cuadro 1, los 100 idiomas incluidos en la muestra tienen un promedio de cerca de 25 fonemas consonánticos y 8,6 fonemas vocálicos. Veintinueve por ciento de ellos usan al acento como un rasgo fonológico distintivo, y 30% de ellos son lenguas tonales. La distribución de estas características entre los grupos, sin embargo, es muy desigual. Mientras los idiomas bantú-khoisan tienen un promedio de 41 fonemas consonánticos, el grupo austronesio tiene un promedio de 16 fonemas consonánticos y menos de 6 fonemas vocálicos. Este último número contrasta con los 16 fonemas vocálicos que tienen en promedio las lenguas germánicas. En lo que se refiere al acento, este rasgo resulta distintivo en el 86% de los idiomas latinos incluidos en nuestra muestra, pero no está presente para nada en las lenguas indoarias, sino-tibetanas, Austro-Tai, dravídicas y norteamericanas. La proporción de lenguas tonales en nuestra muestra, en cambio, alcanza el 100% en la familia sino-tibetana, pero es igual a cero en toda la familia indoeuropea (y también en los grupos correspondientes a los idiomas dravídicos, caucásicos, urálicos y sudamericanos).

Cuadro 2: Coeficientes de correlación entre las variables

Variable	Consonantes	Vocales	Acento	Tono
Consonantes	1.0000			
Vocales	0.0159	1.0000		
Acento	-0.1198	-0.1270	1.0000	
Tono	0.0308	0.0178	-0.2741	1.0000

Las relaciones entre las cifras de nuestra base de datos pueden ser descriptas también en términos de los coeficientes de correlación entre las variables, que son los números que aparecen en el cuadro 2. En él vemos que las cifras más altas en valor absoluto corresponden a los coeficientes que relacionan a *Acento* con las otras tres variables (que son todas negativas y mayores que 0,1), mientras que los otros tres coeficientes (que corresponden a las correlaciones entre *Consonantes* y *Vocales*, *Consonantes* y *Tono*, y *Vocales* y *Tono*) son positivos y menores que 0,04 [10]. Esto parece

[10] De hecho, el único coeficiente estadísticamente significativo del cuadro es el que corresponde a la correlación entre *Acento* y *Tono* (p = 0.0029), en tanto que los que corresponden a los pares *Consonantes/Acento* (p = 0.1176) y *Vocales/Acento* (p = 0.1040) están cerca de ser significativos a un nivel de probabilidad del 10%. Los tres coeficientes con signo positivo, en cambio, no resultan para nada

indicar que dichas correlaciones son menos importantes que las otras tres. Las tres correlaciones menos significativas muestran también un signo contraintuitivo, ya que las tres variables analizadas (*Consonantes*, *Vocales* y *Tono*) deberían supuestamente tener coeficientes de correlación negativos entre ellas [11].

Una posible explicación de esta contradicción acerca del signo de los coeficientes de correlación tiene que ver con la idea de que las cuatro variables fonológicas analizadas en este trabajo están interrelacionadas. Así, cuando tratamos de capturar la relación parcial entre dos esas cuatro variables (por ejemplo, entre *Consonantes* y *Vocales*) usando un coeficiente de correlación, es posible que dicho coeficiente de correlación esté sesgado por la existencia de una relación indirecta entre cada una de esas variables y una tercera (por ejemplo, *Acento*). Esa relación indirecta puede ser lo suficientemente fuerte como para crear la falsa impresión de que la relación directa entre las variables originales tiene cierto signo cuando en realidad tiene un signo diferente. Una manera práctica de enfrentar este problema es correr un sistema de ecuaciones simultáneas, en el cual las relaciones entre las diferentes variables se determinen al mismo tiempo. Eso debería producir los "signos correctos" a la hora de calcular los coeficientes de correlación parcial, y controlar por las interferencias generadas por los efectos indirectos entre varias variables que están interactuando de manera simultánea.

5. Regresiones con ecuaciones simultáneas

Siguiendo la metodología descripta en la sección 2, procedimos a estimar una serie de coeficientes que corresponden a las regresiones especificadas en el sistema de ecuaciones (2)/(5). A efectos de hacer eso, usamos primero el método de las regresiones aparentemente no relacionadas (SUR), y luego controlamos por endogeneidad usando una estimación en tres etapas basada en mínimos cuadrados lineales (el llamado "método de mínimos cuadrados en tres etapas") [12]. Como varios de los coeficientes resultaron ser estadísticamente insignificantes (ya que sus valores de probabilidad eran todos mayores

significativos desde el punto de vista estadístico, ya que sus valores de probabilidad son iguales a 0,4377, 0,3803 y 0,4301.

[11] Ese, por lo menos, sería el signo esperado si creyéramos en la existencia de fenómenos compensatorios, por los cuales la complejidad en una dimensión estuviera compensada por cierta simplicidad en otra dimensión.

[12] Para una explicación de este método, véase Greene (2011), capítulo 10.

que 0,1), eliminamos de las regresiones a las variables que los generaban, e incluimos solo aquellas variables cuyos coeficientes fueran estadísticamente significativos. Finalmente impusimos varias restricciones relacionadas con la simetría de los efectos en cada una de las ecuaciones, y terminamos con un sistema que se parece al descripto en las ecuaciones (6)/(9).

Cuadro 3: Resultados de las regresiones de ecuaciones simultáneas

Concepto	Regresión A		Regresión B		Regresión C		Regresión D	
	Coefic	Prob	Coefic	Prob	Coefic	Prob	Coefic	Prob
Ec 1 (Consonantes)								
Constante	26.603	0.0000	26.139	0.0000	26.399	0.0000	26.739	0.0000
Vocales	-0.0250	0.9059						
Acento	-4.4045	0.0282	-4.2025	0.0282	-5.1010	0.0000	-6.2727	0.0000
Tono	-0.6363	0.7476						
Ec 2 (Vocales)								
Constante	9.6174	0.0000	9.2329	0.0000	9.5925	0.0000	9.5266	0.0000
Consonantes	-0.0056	0.9059						
Acento	-2.3267	0.0142	-2.1136	0.0198	-3.3534	0.0000	-3.1262	0.0000
Tono	-0.6101	0.5148						
Ec 3 (Acento)								
Constante	0.9223	0.0000	0.8966	0.0000	8.0311	0.0000	7.2005	0.0000
Consonantes	-0.0107	0.0282	-0.0103	0.0276	-0.1960	0.0000	-0.1594	0.0000
Vocales	-0.0253	0.0142	-0.0238	0.0155	-0.2982	0.0000	-0.3199	0.0000
Tono	-0.4903	0.0000	-0.4827	0.0000	-0.9510	0.0000	-0.6013	0.0000
Ec 4 (Tono)								
Constante	0.5499	0.0012	0.4457	0.0000	0.6050	0.0000	0.7823	0.0000
Consonantes	-0.0016	0.7476						
Vocales	-0.0070	0.5148						
Acento	-0.5149	0.0000	-0.5024	0.0000	-1.0516	0.0000	-1.6630	0.0000

Todos estos resultados aparecen en el cuadro 1, que muestra los coeficientes estimados y los valores de probabilidad para cuatro sistemas diferentes [13]. El primero de ellos (Regresión A) incluye todas las variables como explicativas en todas las ecuaciones, y está estimado mediante el método SUR. La regresión B es similar a la regresión A, pero excluye a las variables cuyos coeficientes tienen valores de probabilidad mayores que 0,1 (o sea, *Vocales* y *Tono*, en la ecuación 1, *Consonantes* y *Tono*, en la ecuación 2, y *Consonantes* y *Vocales*, en la ecuación 4). La regresión C es similar a la regresión B, pero impone las restricciones de que los coeficientes de la variable *Acento* en las

[13] All the regressions whose results are reported in this paper were performed using the software package EViews 3.5.

ecuaciones 1, 2 y 4 tienen que ser igual a la inversa de los coeficientes de las variables *Consonantes*, *Vocales* y *Tono* en la ecuación 3. La regresión D, por último, es similar a la regresión C, pero fue corrida utilizando un método de mínimos cuadrados en tres etapas en el cual las cuatro variables dependientes (*Consonantes*, *Vocales*, *Acento* y *Tono*) se consideran endógenas, y las variables instrumentales utilizadas son dieciséis variables categóricas que representan los distintos grupos idiomáticos (que son los mismos que aparecen para describir los datos en el cuadro 1).

Los resultados obtenidos son satisfactorios en el sentido de que resultan consistentes con los coeficientes de correlación reportados en el cuadro 2. Esto es así porque los seis coeficientes que son estadísticamente significativos en la regresión A corresponden a los tres coeficientes de correlación con los mayores valores absolutos en el cuadro 2 (*Consonantes/Acento*, *Vocales/Acento* y *Acento/Tono*). Los signos de estos coeficientes, además, son siempre negativos (tal como ocurre con los coeficientes de correlación reportados en el cuadro 2), y el mayor de ellos (*Acento/Tono*) está relacionado con los coeficientes de regresión más significativos. Más aún, los valores obtenidos en la regresión A para los coeficientes que miden la correlación entre *Consonantes* y *Vocales*, *Consonantes* y *Tono*, y *Vocales* y *Tono* también son negativos, si bien todos ellos resultan ser estadísticamente insignificantes.

El enfoque alternativo de estimación por ecuaciones simultáneas propuesto en la sección 2 (al que denominamos “sistema recursivo”) también puede ser aplicado utilizando la información disponible en nuestra base de datos. Para chequear cuál de las posibles alternativas era la mejor, procedimos a correr los seis sistemas univariados mencionados al final de la sección 2. Los coeficientes obtenidos aparecen en el cuadro 4. Todos ellos fueron calculados utilizando mínimos cuadrados en tres etapas, las cuatro variables dependientes fueron consideradas endógenas, y las variables instrumentales empleadas fueron las mismas dieciséis variables categóricas usadas en la regresión D (reportada en el cuadro 3). En este caso, todos los coeficientes estimados para todas las variables en todas las regresiones resultaron ser estadísticamente significativos a un nivel de probabilidad del 1%.

Teniendo en cuenta los signos de los coeficientes en estos seis sistemas de ecuaciones, los mejores resultados parecen ser los que corresponden a la regresión 4, que

fueron obtenidos con el sistema usado como ejemplo en la sección 2 (ecuaciones (10)/(13)). Esos resultados sugieren que *Acento* tiene un efecto negativo sobre el número de consonantes (o sea, que la existencia de distinción entre sonidos acentuados y no acentuados vuelve innecesario el uso de un gran número de fonemas consonánticos), el cual tiene a su vez un efecto negativo sobre el número de vocales (o sea, que un idioma con un gran número de fonemas consonánticos necesita menos fonemas vocálicos que un idioma con menos fonemas consonánticos, controlando por los demás factores). Un número grande de fonemas vocálicos, sin embargo, parece reducir la probabilidad de que un idioma desarrolle una distinción fonológica entre distintos tonos, pero la existencia de dicha distinción fonológica reduce a su vez la probabilidad de que el acento sea un rasgo distintivo en el idioma en cuestión.

Cuadro 4: Coeficientes de las regresiones de ecuaciones recursivas

Concepto	Regr 1	Regr 2	Regr 3	Regr 4	Regr 5	Regr 6
Ec 1 (Consonantes)						
Constante	40.7951	5.2225	34.0545	32.0154	19.5118	29.8536
Vocales	-1.8417	2.2851				
Acento			-31.4984	-24.4669		
Tono					18.0275	-16.4454
Ec 2 (Vocales)						
Constante	5.8736	10.7897	10.8758	18.4667	4.1987	-0.8407
Consonantes				-0.3951		0.3796
Acento	9.4705				15.2459	
Tono		-7.2324	-7.5193			
Ec 3 (Acento)						
Constante	1.1720	0.1159	-0.4680	0.7273	1.1328	-0.0308
Consonantes		0.0070			-0.0338	
Vocales			0.0879			0.0372
Tono	-2.9398			-1.4575		
Ec 4 (Tono)						
Constante	-0.1860	2.8117	-0.8965	0.9118	1.2273	1.5483
Consonantes	0.0195		0.0480			
Vocales				-0.0710	-0.1076	
Acento		-8.6612				-4.3044

Esta forma de analizar las relaciones entre *Consonantes*, *Vocales*, *Acento* y *Tono* es también útil para resolver la contradicción empírica que encontramos al final de la sección 3, donde obtuvimos coeficientes de correlación positivos para las relaciones *Consonantes/Vocales*, *Consonantes/Tono* y *Vocales/Tono*. Con el sistema recursivo

correspondiente a la regresión 4, lo que hallamos son relaciones negativas (y no positivas) entre *Consonantes* y *Vocales*, y entre *Vocales* y *Tono*. La relación supuestamente positiva entre *Tono* y *Consonantes*, por su parte, la podemos explicar por el efecto indirecto que juega la variable *Acento*, la cual está influida negativamente por la distinción por tonos y es a su vez un determinante negativamente relacionado con el número de fonemas consonánticos. Todo esto puede por lo tanto verse como una evidencia en favor de la existencia de fenómenos compensatorios entre las variables fonológicas, por los cuales la complejidad en una dimensión está relacionada con la simplicidad en otra.

Los resultados de todos los otros sistemas recursivos propuestos cuyos resultados aparecen en el cuadro 4 exhiben en cambio una o más incoherencias, y pueden por lo tanto ser descartados. La regresión 1, por ejemplo, sugiere que el número de fonemas vocálicos se incrementa cuando hay distinción por acento, en tanto que la regresión 2 parece indicar que tener más consonantes induce la aparición de distinciones por acento. La regresión 3, por el contrario, parece indicar que cuantos más fonemas vocálicos tenga un idioma, más probable es que exhiba distinciones basadas en el acento y en el tono, mientras que la regresión 5 nos dice que la distinción por el acento tiende a incrementar el número de vocales, y que la distinción por tonos tiende a incrementar el número de consonantes. La regresión 6, por último, sugiere que un mayor número de vocales incrementa la probabilidad que el acento sea un rasgo fonológico distintivo, y que un mayor número de consonantes induce el uso de un mayor número de fonemas vocálicos.

6. Conclusiones

El método propuesto en este artículo, que es relativamente común en otras ciencias sociales tales como la economía y la política, es, según nuestra opinión, una fuente muy prometedora para detectar posibles relaciones en sistemas lingüísticos en los cuales haya varias variables que juegan un papel al mismo tiempo. En particular, dicho método puede ser muy útil cuando enfrentamos situaciones en las cuales no sabemos bien qué fenómenos están influenciando el comportamiento de las variables involucradas, y especialmente cuando sospechamos que dicho comportamiento es recursivo (es decir, que una variable está influenciando a otra, la cual está a su vez influenciando a una tercera,

que a su vez parece tener influencia sobre la primera variable).

Aplicando el método de regresión con ecuaciones simultáneas, podemos dar un primer paso en la detección de la estructura lingüística que está detrás de las variables involucradas en determinado proceso. En este trabajo, por ejemplo, tratamos de avanzar hacia un modelo estadístico de interacción de variables fonológicas que relacionara el número de fonemas consonánticos, el número de fonemas vocálicos y la distinción entre sonidos por acento y por tono, para una muestra representativa de los idiomas del mundo. Estimando varios sistemas de regresión con ecuaciones simultáneas, terminamos con algunas conclusiones acerca de las variables analizadas, y esas conclusiones fueron capaces de llevarnos más allá de lo que puede obtenerse observando solamente los valores promedios de las variables o los coeficientes de correlación entre ellas.

Como resultado de nuestros análisis de regresión, aparecieron dos modelos relativamente generales: un "modelo de ecuaciones de equilibrio" y un "modelo recursivo". El primero de ellos enfatiza la importancia de la distinción por acento como un determinante de las otras tres variables, y señala que los idiomas en los que el acento es distintivo tienden a tener pocos fonemas vocálicos, pocos fonemas consonánticos y ninguna distinción entre tonos [14]. En el lado opuesto del espectro, el modelo predice también que las lenguas tonales, que tienen relativamente muchos fonemas vocálicos y consonánticos, no desarrollarán una distinción fonológica entre sonidos acentuados y no acentuados [15].

En el modelo recursivo, en cambio, *Tono* está negativamente correlacionado con *Acento*, que a su vez está negativamente correlacionado con *Consonantes*, que a su vez está negativamente correlacionado con *Vocales*, que a su vez está negativamente correlacionado con *Tono*. Esto implica que si un idioma no es tonal, tenderá a tener una distinción fonológica entre sonidos acentuados y no acentuados, pocos fonemas consonánticos y relativamente muchos fonemas vocálicos [16]. Por el contrario, los idiomas en los cuales existen distinciones por tono tenderán a no desarrollar una distinción por

[14] Este modelo es bueno para explicar la estructura fonológica de la mayoría de los idiomas latinos, en los cuales el acento es generalmente distintivo. Dichos idiomas tienden a tener menos fonemas vocálicos y consonánticos que el promedio, y no han desarrollado la distinción por tonos.

[15] Ese sería el caso general de los idiomas de la familia Níger-Congo, en los cuales la distinción por acento no es común pero sí lo es la distinción por tono, junto con un número relativamente grande de vocales y consonantes.

[16] Ese sería el caso general para los idiomas germánicos.

acento, y en general tendrán muchos fonemas consonánticos y relativamente pocos fonemas vocálicos [17].

Todas estas conclusiones pueden ser relacionadas con resultados similares hallados en la literatura reseñada en la sección 3. Las predicciones del modelo recursivo, por ejemplo, son consistentes con lo encontrado por Hyman (2009) acerca de los sistemas de distinción por acento y por tono, en tanto que las predicciones del modelo de ecuaciones de equilibrio pueden usarse para explicar por qué Maddieson (2007) no encontró ninguna correlación apreciable entre *Consonantes*, *Vocales* y *Tono* (cuya relación, de acuerdo con dicho modelo, es indirecta y ocurre a través de la interacción con la variable *Acento*).

Los resultados obtenidos a través del uso de nuestras regresiones con ecuaciones simultáneas pueden verse como un punto a favor de la búsqueda de generalizaciones estadísticas y no categóricas dentro del campo de la tipología lingüística[18]. También pueden interpretarse como tendencias del lenguaje, en el sentido de que pueden representar "estados estacionarios" o "puntos estables" hacia los cuales deberían converger las estructuras fonológicas. De acuerdo con esta perspectiva, por ejemplo, un idioma sin distinción por tonos que tiene relativamente pocas consonantes y vocales sería propenso a desarrollar una distinción entre sonidos acentuados y no acentuados, en tanto que un idioma tonal (sin distinción por acento) tendría presumiblemente que exhibir una tendencia a incrementar su número de fonemas. Los idiomas que no siguen estas reglas, por lo tanto, podrían estar en una situación "inestable", y la predicción del modelo sería que algunas de sus características fonológicas actuales cambiarán probablemente en el futuro, hasta alcanzar alguna de las configuraciones de equilibrio halladas a través del análisis estadístico.

Los métodos de regresión con ecuaciones simultáneas, además, podrían ser usados también para analizar distintos conjuntos de variables en comparaciones multilingüísticas. Una posible vía para futuras investigaciones podría ser aumentar el conjunto de características fonológicas incluidas en el análisis, introduciendo variables

[17] Este último conjunto de rasgos no aparece como una característica general de ningún grupo de idiomas de nuestra muestra, pero se aplica a varios casos particulares tales como los de los idiomas birmano, ewé, igbo, kam, khoekhoe, shona, tibetano y wu.
[18] Para una discusión acerca de estas concepciones, véase Bickel (2013).

adicionales que ayuden a explicar el grado de complejidad de las consonantes y las vocales (tales como el uso de consonantes no pulmónicas y de articulaciones dobles, la distinción por sonoridad y por aspiración, el uso de vocales anteriores labializadas, la distinción por duración y por nasalización de los sonidos, etc). Con esos agregados, el conjunto de ecuaciones simultáneas puede ampliarse, y pueden llegar a descubrirse nuevas relaciones estructurales. Otra posible extensión de este trabajo es usar ecuaciones simultáneas para hacer una regresión de un sistema de relaciones que también incluya otras variables lingüísticas relacionadas con características morfológicas, sintácticas o léxicas. En ese caso, podríamos esperar que este método fuera útil para arrojar algo más de luz sobre la posible existencia de fenómenos compensatorios más sofisticados entre los diferentes componentes de los sistemas lingüísticos.

Apéndice: Base de datos de variables fonológicas

El siguiente cuadro muestra todos los datos utilizados en los sistemas de regresión de ecuaciones simultáneas corridos en el presente trabajo.

Idioma	Grupo	Consonantes	Vocales	Acento	Tono
Alemán	GE	20	15	1	0
Amárico	AA	25	5	1	0
Árabe	AA	29	6	0	0
Armenio	OI	30	6	0	0
Aymara	SA	27	3	0	0
Azerí	AL	24	9	0	0
Bengalí	IA	32	8	0	0
Birmano	ST	31	8	0	1
Búlgaro	ES	22	6	1	0
Cabardiano	CU	53	3	1	0
Camboyano	AT	21	21	0	0
Cantonés	ST	19	11	0	1
Catalán	LA	23	7	1	0
Changsha	ST	19	6	0	1
Checoslovaco	ES	25	9	0	0
Cheyenne	NA	10	3	0	1
Coreano	AL	19	18	0	0
Danés	GE	15	20	1	0
Dinka	AA	20	7	0	1
Esloveno	ES	21	8	1	0
Español	LA	19	5	1	0
Ewé	AC	28	7	0	1
Finlandés	CU	13	8	0	0

Francés	LA	20	13	0	0
Gallego	LA	21	7	1	0
Gan	ST	18	7	0	1
Georgiano	CU	28	5	1	0
Griego	OI	18	5	0	0
Guaraní	SA	18	12	0	0
Gujarati	IA	31	8	0	0
Hakka	ST	17	7	0	1
Hausa	AA	28	10	0	1
Hawaiano	AU	8	10	0	0
Hebreo	AA	22	5	1	0
Hindi-Urdu	IA	34	11	0	0
Holandés	GE	18	14	0	0
Húngaro	CU	25	14	0	0
Igbo	AC	26	8	0	1
Inglés	GE	24	11	0	0
Inuit	NA	14	6	0	0
Irlandés	OI	35	11	0	0
Islandés	GE	32	16	0	0
Italiano	LA	23	7	1	0
Japonés	AL	16	5	1	1
Javanés	AU	20	6	0	0
Kam	AT	25	6	0	1
Kannada	DR	34	11	0	0
Khoekhoe	BK	31	8	0	1
Kurdo	OI	31	8	1	0
Lao	AT	28	18	0	1
Lingala	BK	24	7	1	1
Malayalam	DR	39	12	0	0
Malayo-Indonesio	AU	18	6	0	0
Malgache	AU	27	4	0	0
Manchuriano	AL	20	6	0	0
Mandarín	ST	19	6	0	1
Maorí	AU	10	5	0	0
Mapuche	SA	22	6	0	0
Marathi	IA	44	6	0	0
Maya Yucateco	NA	21	10	0	1
Mongol	AL	26	14	0	0
Nahuatl	NA	15	8	0	0
Navajo	NA	28	16	0	1
Nepalí	IA	27	11	0	0
Noruego	GE	23	19	1	0
Nyangumarta	AU	17	3	1	0
Oromo	AA	24	5	0	0
Pashto	OI	30	7	1	0
Persa	OI	23	6	1	0
Polaco	ES	31	6	0	0
Portugués	LA	19	7	1	0
Punjabi	IA	30	10	0	0

Quechua Cusqueño	SA	25	3	0	0
Rumano	LA	22	7	1	0
Ruso	ES	21	6	1	0
Serbio-Croata	ES	25	5	1	0
Shilha	AA	33	3	0	0
Shona	BK	48	5	0	1
Sindhi	IA	46	10	0	0
Skou	AU	13	7	0	1
Somalí	AA	22	20	0	1
Sueco	GE	18	17	1	0
Swahili	BK	32	5	0	0
Tagalog	AU	16	5	1	0
Tailandés	AT	21	9	0	1
Taiwanés	ST	22	6	0	1
Tamil	DR	16	10	0	0
Telugu	DR	33	11	0	0
Tibetano	ST	28	8	0	1
Turco	AL	22	8	0	0
Ucraniano	ES	27	6	1	0
Uzbeco	AL	26	6	0	0
Vasco	OI	25	5	1	0
Vietnamita	AT	22	11	0	1
Wolof	AC	26	15	1	0
Wu	ST	27	6	0	1
Xhosa	BK	59	12	0	1
Yoruba	AC	18	11	0	1
Zapoteco	NA	20	5	1	1
Zulú	BK	52	12	0	1

Abreviaturas utilizadas: AA = Afroasiático; AC = Atlántico africano; AL = Altaico; AT = Austro-Tai; AU = Austronesio (incluye papuano y Pama-Nyungan); BK = Bantú-Khoisan; CU = Caucásico-Urálico; DR = Dravídico; ES = Eslavo; GE = Germánico; IA = Indoario; LA = Latino; NA = Norteamericano; OI = Otros idiomas indoeuropeos (incluye vasco); SA = Sudamericano; ST = Sino-tibetano.

Referencias bibliográficas

Atkinson, Quentin (2011): "Phonemic Diversity Supports Serial Founder Effect Model of Language Expansion from Africa", *Science*, vol 332, pp 346-349.

Becker-Kristal, Roy (2010): *Acoustic Typology of Vowel Inventories and Dispersion Theory: Insights from a Large Cross-Linguistic Corpus*. Los Ángeles, Universidad de California.

Bickel, Balthasar (2011): "Statistical Modeling of Language Universals", *Linguistic Typology*, vol 15(*2*), pp 401-413.

Bickel, Balthasar (2013): "Distributional Typology: Statistical Inquiries into the Dynamics of Linguistic Diversity". Zurich, Universidad de Zurich.

Brown, Keith (2006): *Encyclopedia of Language and Linguistics*, 2da edición.

Amsterdam, Elsevier.
Comrie, Bernard (2009): *The World's Major Languages*, 2da edición. Oxford, Routledge.
Coupé, Christophe, Egidio Marsico y François Pellegrino (2009): "Structural Complexity of Phonological Systems", en F. Pellegrino et al.: *Approaches to Phonological Complexity*. Berlín, Mouton De Gruyter.
Donohue, Mark y Johanna Nichols (2011): "Does Phoneme Inventory Size Correlate with Population Size?", *Linguistic Typology*, vol 15(*2*), pp 161-170.
Dryer, Matthew (2009): "Problems Testing Typological Correlations with the Online WALS", *Linguistic Typology*, vol 13(*1*), pp 121-135.
Fenk-Oczlon, Gertraud y August Fenk (2004): "Systemic Typologies and Crosslinguistic Regularities", en V. Solovyev y V. Polyakov: *Text Processing and Cognitive Technologies*. Moscú, MISA.
Gary, Jane y Rubino, Carl (2001): *Facts About the World's Languages: An Encyclopedia of the World's Major Languages*. Nueva York, H. W. Wilson.
Greene, William (2011): *Econometric Analysis*, 7ma edición. Nueva York, Prentice-Hall.
Hyman, Larry (2009): "How (Not) to Do Phonological Typology: The Case of Pitch-Accent", *Language Sciences*, vol 31, pp 213-238.
IPA (1999): *Handbook of the International Phonetic Association*. Cambridge, Cambridge University Press.
Jaeger, Florian, Peter Graff, William Croft y Daniel Pontillo (2011): "Mixed Effects Models for Genetic and Areal Dependencies in Linguistic Typology", *Linguistic Typology*, vol 15(*2*), pp 281-320.
Johnson, Keith (2008): *Quantitative Methods in Linguistics*. Oxford, Blackwell.
Kennedy, Peter (2008): *A Guide to Econometrics*, 6ta edition. Nueva York, Wiley.
Kramsky, Jiri (1959): "A Quantitative Typology of Languages", *Language and Speech*, vol 2(*2*), pp 72-85.
Lewis, Paul, Gary Simons y Charles Fennig (2013): *Ethnologue: Languages of the World*, 17ma edition. Dallas, SIL International.
Lindblom, Björn y Ian Maddieson (1988): "Phonetic Universals in Consonant Systems", en L. Hyman y C. N. Li (eds.): *Language, Speech and Mind*. Oxford, Routledge.
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 y M. Ohala: *Experimental Approaches to Phonology*. Nueva York, Oxford University Press.
Marsico, Egidio, Ian Maddieson, Christophe Coupé y François Pellegrino (2004): "Investigating the 'Hidden' Structure of Phonological Systems", *Proceedings of the 30th Meeting of the Berkeley Linguistics Society*, pp 256-267.
Shosted, Ryan (2006): "Correlating Complexity: A Typological Approach", *Linguistic Typology*, vol 10(*1*), pp 1-40.
Silnitsky, George (2003). "Correlation of Phonetic and Morphological Systems of Indo-European Languages", *Journal of Quantitative Linguistics*, vol 10(*2*), pp 129-141.
Wichmann, Soren, Taraka Rama y Eric Holman (2011): "Phonological Diversity, Word Length and Population Sizes Across Languages: The ASJP Evidence", *Linguistic Typology*, vol 15(*2*), pp 177-198.