UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

RELACIONES ENTRE MEDIDAS DE COMPLEJIDAD LINGÜÍSTICA

Germán Coloma

Septiembre 2018
Nro. 658

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>

RELACIONES ENTRE MEDIDAS DE COMPLEJIDAD LINGÜÍSTICA

Germán Coloma [*]

Resumen

Este trabajo sintetiza una serie de conceptos que han sido propuestos por la literatura lingüística para medir la complejidad de los distintos aspectos de los idiomas naturales (fonología, morfología, sintaxis, semántica). Se adentra también en el estudio de las relaciones que pueden establecerse entre dichas medidas de complejidad, y analiza la posible existencia de "efectos de compensación" entre tales medidas (es decir, situaciones en las cuales una característica más compleja de un idioma se ve contrarrestada por otra más simple). Para ello se van introduciendo distintos criterios estadísticos y empíricos, que van desde las correlaciones simples hasta correlaciones parciales, coeficientes de regresión y parámetros surgidos de sistemas de ecuaciones. La mayoría de los conceptos se ilustran apelando a dos bases de datos interlingüísticas que ya han sido utilizadas por el autor en trabajos anteriores.

Palabras clave: complejidad lingüística, correlación, efectos de compensación, ecuaciones simultáneas.

1. Introducción

Cuando una persona intenta intuitivamente calificar a cierto idioma como "simple" o "complejo", suele partir del marco de referencia que le da la lengua de la cual es hablante nativo. A los que hablamos español, por ejemplo, nos resulta natural pensar que el idioma portugués es relativamente simple, y que el chino mandarín, en cambio, es más complejo. Más aún, cuando nos topamos con el idioma gallego, solemos pensar que el mismo es más simple todavía que el portugués.

De hecho, juicios como los contenidos en el párrafo anterior derivan directamente de que el portugués y el gallego son idiomas parecidos al castellano, en tanto que el mandarín es totalmente distinto. Si el portugués nos parece simple y el mandarín complejo es porque aquel tiene muchas características comunes con nuestra lengua (que, por lo tanto, no tenemos necesidad de aprender), en tanto que el mandarín tiene muchas características diferentes. Pero la pregunta respecto de qué idioma es más complejo tendrá seguramente una respuesta distinta si, en vez de hacérsela a un hablante de español, se la hacemos a un hablante de chino shanghainés. A dicha persona el mandarín le parecerá sin duda mucho más simple que el portugués, entre otras cosas porque se escribe con los mismos símbolos de su propio idioma y porque tiene una gramática virtualmente idéntica.

Los ejemplos expuestos son evidencias directas de que el concepto de complejidad idiomática puede pensarse (y, de hecho, comúnmente se piensa) como algo relativo a determinado hablante o grupo de hablantes. Pero también es posible concebir a la complejidad de un idioma en términos absolutos, es decir, evaluar si un idioma es más

[*] Universidad del CEMA; Av. Córdoba 374, Buenos Aires, C1054AAP, Argentina. Teléfono: 6314-3000. Correo electrónico: gcoloma@cema.edu.ar. Las opiniones son personales y no representan necesariamente las de la Universidad del CEMA.

complejo que otro no porque sea más o menos parecido al idioma que uno habla sino por la presencia o ausencia de ciertas "complejidades objetivas". Comparemos, por ejemplo, la conjugación del verbo "amar" en portugués y en inglés. En portugués, este verbo se escribe igual que en castellano, y se pronuncia de una manera muy parecida. El inglés, en cambio, utiliza una palabra totalmente distinta (*love*). Sin embargo, para conjugar el verbo "amar", en inglés solamente necesitamos tres formas (*love – loves – loved*), y una vez que las aprendemos y las combinamos con algunas reglas generales que sirven para conjugar cualquier verbo, podemos amar en presente, pasado y futuro a cualquier persona singular o plural que queramos.

En portugués, en cambio, el verbo "amar" adopta 49 formas distintas ("*amo*", "*ama*", "*amei*", "*amavam*", "*amado*", etc.). Algunas de ellas se usan solo para una determinada función (por ejemplo, "*amo*" solo sirve para la primera persona del singular del presente del indicativo, igual que en castellano), y otras para más de una (por ejemplo, "*ama*" se usa para la tercera persona singular del presente del indicativo y para la segunda persona del singular del imperativo, también igual que en castellano).

La enorme diferencia que hay entre las pocas variantes verbales del inglés y las muchísimas que presenta el portugués (y nótese que lo que estamos describiendo es un verbo "regular") hacen que aun un hablante de español o de gallego, que encuentra que casi todas las formas portuguesas son muy similares a las de su propio idioma, concuerde que, en términos absolutos, la conjugación del verbo "amar" en portugués es más compleja que la conjugación del verbo equivalente en inglés. Partiendo de esa conclusión, solo hay un paso para llegar a otra que nos dice que la conjugación del verbo "amar" no solo es más simple en inglés que en portugués, sino que también es más simple en inglés que en castellano. Y eso es así porque en este tema estamos comparando cierto nivel de "complejidad absoluta" y no de "complejidad relativa".

Una definición posible de complejidad absoluta, que se utiliza para evaluar distintos aspectos de los idiomas, aparece en un artículo del lingüista finlandés Matti Miestamo (2008), y hace referencia al "número de partes de un sistema". Otra definición alternativa es la propuesta por el estadounidense John McWhorter (2001), quien considera que un idioma es más complejo que otro si posee "más distinciones y/o reglas explícitas". Ambas definiciones pretenden evaluar la complejidad de manera independiente de la persona que esté llevando a cabo la evaluación: la conjugación del verbo "amar" es más compleja en portugués que en inglés porque forma un sistema que consta de 49 "partes" (contra las tres que tiene el sistema equivalente en inglés), y también es más compleja porque surge como consecuencia de la aplicación de numerosas distinciones entre formas que sirven para los distintos tiempos y personas verbales (en oposición a las dos o tres distinciones que hay que hacer en inglés para saber cuándo usar "*love*", "*loves*" y "*loved*").

Pero en lingüística existe también un concepto de complejidad relativa que intenta librarse de la trampa que se genera cuando uno evalúa un idioma muy parecido al propio y lo compara con otro que es muy diferente. Dicho concepto considera que un fenómeno idiomático es más complejo si es "más difícil de procesar o de aprender", y su definición está asociada con la obra del holandés Wouter Kusters (2003). Para él, la mayor o menor complejidad relativa de un idioma debe analizarse desde la perspectiva de un "extranjero generalizado" (*generalized outsider*), o sea, de alguien que no habla la lengua en cuestión ni está familiarizado con la cultura de las personas que hablan dicha lengua como idioma

nativo, y que solo la necesita para comunicarse con esas personas. Dado eso, Kusters considera que la complejidad relativa de un idioma es el esfuerzo que dicho extranjero generalizado debe hacer para aprender el idioma en cuestión.

Los conceptos de complejidad absoluta y relativa pueden aplicarse a la evaluación de los idiomas en su conjunto o bien a la evaluación de algún componente de dichos idiomas. Cuanto más acotado sea dicho componente, más precisa será la evaluación, pero también es más probable que la misma tenga un nivel de generalidad menor. Si queremos, por ejemplo, comparar la complejidad de la conjugación del verbo "cantar" en español y en inglés, un punto que podemos observar es que todas las formas de dicho verbo ("canto", "cantas", "cantaba", "cantarían", etc.) comienzan en nuestro idioma con la expresión "cant-" (que es lo que se conoce como la "raíz" de todas esas palabras). En inglés, en cambio, el verbo equivalente presenta solo cuatro formas (*sing – sings – sang – sung*), pero las últimas dos cambian la raíz respecto de las dos primeras. Esto nos indica que, en cierto sentido, la conjugación del verbo "cantar" es más compleja en inglés que en castellano (que usa la misma raíz para todas las formas verbales). Pero esto es un fenómeno relacionado con cierta "complejidad local", aplicable únicamente al verbo que estamos analizando.

Cuanto más general es la comparación que realicemos, sin embargo, la evaluación de la complejidad se vuelve menos local y más "global". La comparación entre las conjugaciones de "amar" en portugués y en inglés que hicimos en el apartado anterior, por ejemplo, es más general que la que acabamos de hacer para el verbo "cantar", ya que "amar" es un verbo regular en los dos idiomas contrastados, en tanto que "cantar" es regular en castellano (y en portugués) e irregular en inglés.

Más general aún es la comparación entre las distintas formas del tiempo pasado que puede adoptar cualquier verbo en un idioma. En inglés, por ejemplo, se utiliza una única forma para varias situaciones en las que el castellano nos pide que distingamos entre pretérito perfecto del indicativo, pretérito imperfecto del indicativo y pretérito imperfecto del subjuntivo. Compárense, por ejemplo, las versiones en español y en inglés de los tres enunciados que aparecen más abajo:

(1)	*Ellos <u>estuvieron</u> en España*	–	*They <u>were</u> in Spain*
(2)	*Ellos <u>estaban</u> en España*	–	*They <u>were</u> in Spain*
(3)	*Ojalá <u>estuviesen</u> en España*	–	*I wish they <u>were</u> in Spain*

Los ejemplos expuestos nos sirven para ilustrar que, en términos mucho más globales, el sistema verbal del español es más complejo que el del inglés, y en eso concuerdan tanto una visión basada en el concepto de complejidad absoluta (porque el sistema verbal español tiene más "partes" y más "reglas", y hace más distinciones que el inglés) como una basada en el concepto de complejidad relativa (porque requiere más esfuerzo de aprendizaje por parte de un "extranjero generalizado").

El mismo tipo de comparaciones es susceptible de hacerse respecto de otros aspectos del lenguaje, como puede ser la pronunciación de las palabras. Supongamos que analizamos primero la complejidad de la palabra "sombrero" en español y en inglés (idioma en el cual la palabra equivalente es "*hat*"). Tanto desde el punto de vista absoluto como relativo, esta palabra parece ser más simple en inglés que en español, ya que consta de solo tres sonidos (contra ocho que tiene en español) y de una sola sílaba (contra tres de la versión española). Las tres sílabas de la palabra "som-bre-ro", además, son bastante

diferentes entre sí (la primera tiene una vocal en medio de dos consonantes, la segunda tiene dos consonantes y luego una vocal, y la tercera tiene una consonante y luego una vocal). Si bien algunos de los sonidos que se usan en español para pronunciar "sombrero" se repiten (la "o" y la "r" aparecen dos veces cada una), de cualquier modo el número de sonidos distintos que se necesitan para pronunciar dicha palabra es el doble de los que se necesitan para pronunciar "*hat*". Todo eso hace también que desde el punto de vista del esfuerzo de procesamiento y de aprendizaje que se requiere para dominar la pronunciación de esta palabra, casi cualquier extranjero encuentre más sencillo decir "*hat*" en vez de decir "sombrero".

Este análisis absolutamente local de la complejidad de la pronunciación, sin embargo, se invierte totalmente si nos movemos a un entorno más general. En inglés, por ejemplo, existen 10 u 11 sonidos vocálicos que sirven para distinguir entre el significado de las palabras, mientras que en castellano solo tenemos cinco vocales ([a], [e], [i], [o] y [u]). Para el oído de un inglés, las palabras "*hat*" ("sombrero"), "*heat*" ("calor"), "*hit*" ("golpear"), "*hut*" ("cabaña"), "*hurt*" ("dañar"), "*heart*" ("corazón") y "*hot*" ("caliente") solo difieren entre sí por la vocal que aparece entre [h] y [t], y existen además otras vocales adicionales (por ejemplo, las que aparecen en "*head*", "*hoard*", "*hood*" y "*who'ed*") que también sirven para distinguir palabras entre sí.

En inglés, asimismo, es relativamente común encontrar vocablos tales como "*sprints*" ("corridas"). Dicha palabra, que solo tiene una sílaba, consta de una vocal precedida por tres consonantes y seguida de otras tres (con lo cual la sílaba en cuestión está compuesta por siete sonidos). Nada similar existe en español, donde el número máximo de sonidos que puede tener una sílaba es cinco (por ejemplo, la primera sílaba de la palabra "transporte"), y aún en esos casos lo más común es que la pronunciemos omitiendo alguno de dichos sonidos (por ejemplo, que digamos "trasporte", sin pronunciar la "n").

Las particularidades mencionadas en los dos párrafos anteriores (y unas cuantas más) hacen que, en general, cualquier "extranjero generalizado" considere que la pronunciación del inglés es más compleja que la del español, y eso puede objetivizarse a través del recuento del número de vocales de ambos idiomas y de la caracterización de sus posibles estructuras silábicas. Cuanto más grande es el nivel en el cual uno quiere definir la complejidad de un idioma, sin embargo, más difícil se vuelve hacerlo. Esto ha llevado a que no sea habitual utilizar una medida única para la complejidad global de las distintas lenguas, y que en la mayoría de los casos se trabaje con medidas de complejidad local.

En su libro sobre complejidad gramatical, el lingüista estadounidense Peter Culicover (2013) considera que, a efectos de su uso en los análisis interlingüísticos (es decir, para la comparación entre varias lenguas) e intralingüísticos (es decir, para la comparación entre estructuras de una misma lengua) lo más útil es aplicar conceptos que se computen en términos relativamente locales pero que puedan definirse de manera general. Uno de dichos conceptos es el de "marcación" (*markedness*), que proviene de la lingüística teórica y cuyo uso se asocia con la obra de Noam Chomsky (1965). Según esta concepción, una forma es más compleja si está "marcada" y menos compleja si no lo está, y la propia lógica del idioma guía hacia el uso de la forma más simple (menos marcada). Por ejemplo, si partimos del sustantivo "cómputo", vemos que el mismo está menos marcado en su forma singular que plural ("cómputos"), porque el plural aparece

marcado con el sufijo "-s". Cuando el idioma español deriva un verbo en base a dicho sustantivo, parte de la forma menos marcada (singular) y no de la más marcada (plural), y construye la palabra "computar" (en vez de "computosar"). Lo mismo hace al volver a formar un sustantivo ("computadora"), y también al crear un nuevo verbo ("computarizar"). Cada una de estas formas, sin embargo, está más marcada que la anterior y es, por lo tanto, más compleja.

Otro concepto mencionado por Culicover como una forma de medir la complejidad local de ciertos fenómenos lingüísticos tiene que ver con la denominada "estructura jerárquica" de los mismos. Dicho concepto se ejemplifica normalmente con la oposición entre oraciones simples y complejas, en la cual las primeras aparecen aisladas y las segundas aparecen integradas por varios enunciados jerárquicamente ordenados. Cuanto más niveles jerárquicos haya dentro de una oración, más compleja será la misma, como puede comprobarse en los siguientes ejemplos que van incluyendo cada vez más enunciados "incrustados" dentro de la oración principal:

(4) *El perro come la comida.*

(5) *El perro come la comida [que le dio el hombre]*.

(6) *El perro come la comida [que le dio el hombre [que tenía otro perro]]*.

(7) *El perro come la comida [que le dio el hombre [que tenía otro perro [que ladraba mucho]]]*.

Nótese sin embargo que la mayor simplicidad de (4) respecto de las siguientes oraciones sucesivamente más complejas se obtiene a costa de un menor contenido informativo. Si, por ejemplo, quisiéramos decir lo mismo que expresa (7) utilizando exclusivamente oraciones simples, deberíamos escribir algo así:

(8) *El perro come la comida. Un hombre se la dio. Ese hombre tenía otro perro. Ese otro perro ladraba mucho.*

Si ahora comparamos (7) con (8), vemos que, en cierto sentido, este último texto es más complejo que el primero, ya que consta de cuatro oraciones distintas (en vez de una sola "gran oración"), que en total tienen 20 palabras (en vez de 17), 38 sílabas (en vez de 31) y 75 sonidos (en vez de 60). Desde el punto de vista de la "eficiencia informativa" de ambos textos, por lo tanto, el primero parece ser mejor que el segundo.

Es precisamente dicha idea de eficiencia la que se encuentra implícita en otro concepto que sirve para evaluar la complejidad relativa de las estructuras lingüísticas, y que tiene que ver con el procesamiento de los componentes de tales estructuras. Dicho concepto, propuesto por el inglés John Hawkins (2004), considera que una estructura es más compleja si, dado el contenido que quiere expresar, es menos eficiente. Dicha eficiencia, por su parte, tiene que ver con la cantidad de memoria que debe utilizarse para retener el contenido que se está procesando, la cual a su vez depende de la cantidad de componentes que se están utilizando y de las propiedades que se le asignan a dichos componentes. Comparemos, por ejemplo, estas dos oraciones:

(9) *Juan rompió el televisor [que me pidió prestado]*.

(10) *Juan rompió el [que me pidió prestado] televisor.**

Tal como puede observarse, (9) es una oración que en castellano consideramos

correcta, en tanto que (10) es una oración que consideramos incorrecta (y por ese hecho la estamos señalando con un asterisco). Sin embargo, tanto (9) como (10) están construidas usando exactamente las mismas palabras (es decir, los mismos componentes) y, en principio, buscan expresar el mismo significado. Pero mientras (9) es más "eficiente" para expresar dicho significado, (10) lo es menos, básicamente porque obliga a recordar por más tiempo que Juan rompió algo, antes de saber qué es exactamente lo que rompió. Es probablemente por eso que el idioma español nos fuerza a usar (9) en vez de (10), dado que esa última forma de expresar lo mismo es menos eficiente y, por ende, más compleja.

En el caso de la comparación de (7) con (8), en cambio, la mayor eficiencia (y, por ende, la menor complejidad de procesamiento) de la primera alternativa depende del hecho de que usa menos componentes que la segunda. Nótese además que, aunque (7) es una sola oración "gramaticalmente compleja", respeta la lógica implícita en el razonamiento de Hawkins de ir procesando los componentes sin necesidad de retenerlos por demasiado tiempo en nuestra memoria. La misma oración sería mucho menos eficiente (y más compleja) si la escribiéramos del siguiente modo:

(11) *El perro come la [que le dio el hombre [que tenía otro perro [que ladraba mucho]]] comida.**

por el mismo motivo por el que (10) es menos eficiente que (9).

2. Componentes del lenguaje

El lenguaje humano está compuesto esencialmente por sonidos, que se agrupan en palabras, que se agrupan a su vez en enunciados.[1] Cada uno de dichos elementos representa una unidad que se combina con otras unidades equivalentes, con el objeto de formar construcciones más complejas. Los sonidos, por ejemplo, son las unidades mínimas de pronunciación de un idioma, y se combinan entre sí de manera secuencial para formar palabras. Las palabras son las unidades mínimas con significado independiente, que también se combinan de manera secuencial para formar enunciados (que son, a su vez, las unidades mínimas de comunicación en las cuales se dividen los textos).

Entre los sonidos, las palabras y los enunciados existen ciertas unidades intermedias que tienen importancia para analizar distintos fenómenos lingüísticos. A mitad de camino entre el sonido y la palabra se encuentra la sílaba, que es un conjunto de sonidos que se pronuncian de manera relativamente simultánea y que están organizados en torno a un determinado núcleo (sonido básico), que es generalmente una vocal. También a mitad de camino entre el sonido y la palabra se encuentra el morfema, que es un conjunto de sonidos cuya unidad no tiene que ver con su pronunciación sino con su

[1] Esta caracterización, obviamente, se aplica a las lenguas habladas. En el caso de las lenguas de señas, no existe una unidad relacionada con el sonido, y la mayoría de los signos que se utilizan en ellas tienen un significado en sí mismo (y son, en ese sentido, más asimilables a palabras que a unidades de sonido). En las lenguas de señas, sin embargo, existen formas de descomponer a las señas en distintos elementos o "articuladores" (gestos manuales, posición de la mano, dirección del movimiento, etc.) que podrían cumplir las funciones que tienen los sonidos como elementos que forman las palabras en las lenguas habladas. Para un estudio más detallado de estos temas, puede consultarse el manual de lenguas de señas publicado en el año 2012 por la editorial De Gruyter, cuyos compiladores son Roland Pfau, Markus Steinbach y Bencie Woll.

significado. La palabra "inevitable", por ejemplo, está formada por diez sonidos (que en este caso corresponden cada uno a una letra), los cuales se agrupan de manera diferente en cinco sílabas y en tres morfemas. Mientras que la agrupación de los sonidos en sílabas nos indica el modo en el cual la palabra se pronuncia (i-ne-vi-ta-ble), la agrupación en morfemas nos indica cuáles son los componentes que hacen al significado final de dicha palabra (in-evit-able).

En cuanto a las unidades intermedias entre las palabras y los enunciados, vale la pena mencionar al sintagma. Este es un conjunto de palabras organizado jerárquicamente, que contiene una palabra que constituye su núcleo y, eventualmente, otras que operan como complementos de dicho núcleo o como especificadores del sintagma como un todo. Así como las sílabas y los morfemas pueden tener un solo sonido, y las palabras pueden tener una sola sílaba o un solo morfema, los sintagmas también pueden tener una sola palabra, y los enunciados tener un solo sintagma. Por ejemplo, el enunciado "Juan come" está formado por dos sintagmas: "Juan" (sintagma nominal) y "come" (sintagma verbal), y cada uno de ellos consta de una única palabra. En cambio, el enunciado "¡Qué mala suerte!" está formado por un único sintagma nominal, pero el mismo tiene tres palabras. Una de ellas es el núcleo ("suerte"); otra es un complemento de dicho núcleo ("mala"); y la tercera ("qué") actúa como especificador de todo el sintagma, indicando que se trata de una exclamación.[2]

2.1. Fonología y fonética

La descomposición del lenguaje humano en sonidos, palabras y enunciados (y su posterior agrupamiento y división en sílabas, morfemas y sintagmas) permite analizar a la lengua desde distintas perspectivas. Una de dichas perspectivas tiene que ver con la estructura del sistema de sonidos que se usa para construir las palabras, y recibe el nombre de fonología. Dicho campo de análisis estudia los sonidos como unidades que tienen valor distintivo para formar palabras, y los agrupa en "paquetes" más o menos similares que se pueden usar con dicha función. Cada uno de dichos paquetes recibe el nombre de "fonema", y dentro de él puede haber un único sonido o varios sonidos más o menos parecidos cuya distinción no es relevante para el idioma que se está analizando.

Tomemos, por ejemplo, la palabra española "endulzado". En dicha palabra, la letra "d" aparece dos veces, pero el sonido que se usa para pronunciarla es distinto en cada situación. En su primera aparición, la pronunciación estándar consiste en apoyar la lengua un poco por encima de los dientes y luego separarla, haciendo vibrar al mismo tiempo las cuerdas vocales. Dicho sonido se representa fonéticamente a través del símbolo [d]. Para la "d" que aparece por segunda vez (es decir, la que está entre medio de las vocales "a" y "o"), la pronunciación más común consiste en cambio en hacer vibrar las cuerdas vocales pero sin llegar a apoyar la lengua por encima de los dientes, dejando un pequeño espacio para que pase el aire. Ese segundo sonido se representa fonéticamente a través del símbolo [ð].

Si bien para los que hablamos español la diferencia entre los dos sonidos descriptos en el párrafo anterior nos parece algo completamente esotérico, hay idiomas en los cuales sirve para distinguir entre palabras con significados diferentes. Cuando en inglés alguien pronuncia la palabra "*day*" ("día"), utiliza un sonido parecido al de la

[2] Para una explicación más extensa de todos estos conceptos, véase Coloma (2017).

primera “d” de “endulzado”. En cambio, si la misma persona pronuncia la palabra “*they*” (“ellos”), usa un sonido bastante semejante al de la segunda “d” de “endulzado”.

La causa por la cual nosotros “no nos damos cuenta” de que pronunciamos distinto la letra “d” en diferentes posiciones es que para nosotros [d] y [ð] son dos “alófonos” del mismo fonema (es decir, dos sonidos que tenemos mentalmente agrupados dentro del mismo paquete). En el idioma inglés, en cambio, /d/ y /ð/ son dos fonemas distintos. El hecho de que puede haber sonidos distintos que están en un mismo idioma y sin embargo no tienen valor distintivo vuelve aconsejable utilizar una nomenclatura diferente para representar a los sonidos en sí (que usualmente se escriben entre corchetes) y para representar a los fonemas (que usualmente se escriben entre barras). El símbolo que se usa para representar a determinado fonema, sin embargo, se elige siempre entre los que están disponibles en el alfabeto fonético, y en general es el del sonido (alófono) más característico del fonema en cuestión.

La distinción entre fonema y sonido es en cierto modo la clave de la diferencia que existe entre la fonología y la fonética. Mientras la primera tiene que ver con la estructura de los sonidos de un idioma en función de formar palabras que luego tengan algún significado, la fonética es una disciplina que estudia la pronunciación en sí, y el conjunto de sonidos y combinaciones de ellos que se usan en el lenguaje. La fonética es, por lo tanto, una de las herramientas que usa la fonología para definir el sistema de sonidos de una lengua. Desde una perspectiva inversa, la fonología es una especie de “fonética funcional”, que solo se preocupa por analizar la pronunciación en tanto cumpla un papel en el sistema conjunto de la lengua.

Otros dos factores que influyen sobre la fonología de una lengua son el acento y el tono. El acento es un elemento que sirve para distinguir entre sílabas que se pronuncian de manera más o menos prominente, y en castellano es un elemento que usamos mucho para distinguir entre palabras que tienen los mismos fonemas ubicados en el mismo orden, pero que tienen significados diferentes (por ejemplo, “revolver” y “revólver”, o “esta” y “está”). El tono, en cambio, no se usa en español como un elemento para cambiar el significado de una palabra, pero sí sirve para indicar el uso que se le está dando a la misma (afirmación, interrogación, exclamación, etc.). De ese modo, por ejemplo, es posible distinguir entre distintos usos del mismo vocablo cuando decimos “mamá”, “¿mamá?” o “¡mamá!”.

Existen numerosas idiomas, sin embargo, en los cuales el tono sirve para modificar el significado de una palabra. En chino mandarín, por ejemplo, la palabra “*ma*” tiene un significado distinto según se la pronuncie con un tono “alto” (en cuyo caso significa “madre”) o en un tono “bajo” (en cuyo caso significa “caballo”). Ese tipo de lenguas en las cuales el tono tiene un carácter distintivo o “contrastante” suelen ser denominadas “lenguas tonales”.

2.2. Morfología y sintaxis

Nótese que, en el lenguaje humano, ni los sonidos ni los fonemas tienen en sí ningún significado, más que el de servir para combinarse y formar palabras. El estudio de la palabra desde el punto de vista de sus constituyentes mínimos con significado separable es, en cambio, el campo de la morfología. Dicha manera de analizar el lenguaje se concentra en identificar los distintos morfemas que pueden constituir una palabra, así como las relaciones entre los mismos. También sirve para distinguir entre morfemas

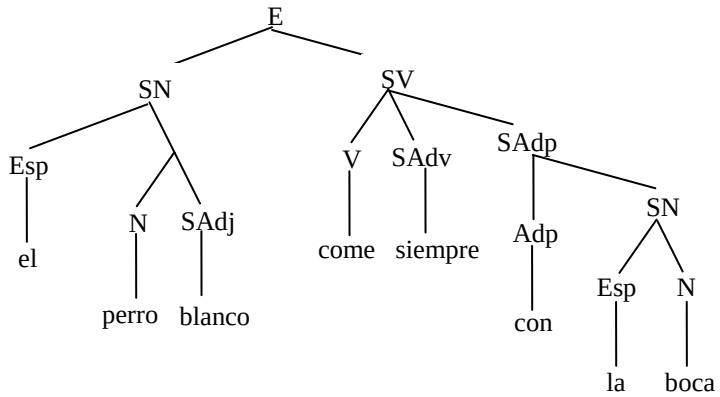
léxicos (es decir, los que tienen un significado conceptual) y morfemas gramaticales (que solo cumplen una función auxiliar).

Muchas veces, los morfemas gramaticales adoptan formas que solo les permiten aparecer como componentes de una palabra junto a un morfema léxico. Dichos morfemas gramaticales se unen a los morfemas léxicos en un proceso que se denomina "afijación", y que consiste en formar palabras con una raíz (morfema léxico) y con uno o varios afijos (morfemas gramaticales). En el caso de la palabra "in-evit-able", por ejemplo, "-evit-" es la raíz que hace referencia al verbo "evitar", en tanto que "in-" (negación) y "-able" (capacidad) son afijos que modifican el concepto principal denotado por la raíz.

La morfología estudia también casos en los cuales los morfemas se unen en una relación de composición, en la cual dos o más morfemas léxicos forman una nueva palabra (por ejemplo, "mal-trato"). Pueden darse también casos combinados, en los que hay al mismo tiempo composición y afijación (por ejemplo, "mal-trat-ador").

Si ahora pasamos a un nivel de análisis más agregado, y nos concentramos en el modo en el cual las palabras se combinan para formar enunciados, entramos en el campo de la sintaxis. Dicha rama del análisis lingüístico parte de la idea de que las palabras se combinan en sintagmas o "frases", y estos a su vez forman enunciados más o menos complejos. Si bien los enunciados pueden estar formados por un único sintagma nominal (por ejemplo, "Buenos días") o verbal (por ejemplo, "Llueve"), lo normal es que los mismos consistan en una combinación entre un sintagma nominal (SN) y un sintagma verbal (SV).[3] Dentro de dichos sintagmas, a su vez, pueden aparecer otros que forman parte de ellos, y que pueden ser sintagmas nominales, verbales, adjetivos, adverbiales y adposicionales. Todos ellos, sin embargo, se combinan dentro de una estructura que tiene una forma jerárquica, la cual se puede representar gráficamente a través de un "diagrama de árbol".

Tomemos, por ejemplo, el enunciado "El perro blanco come siempre con la boca", que puede representarse a través del siguiente diagrama:



Tal como puede observarse, este enunciado (E) se compone de un sintagma nominal ("el perro blanco") que se combina con un sintagma verbal ("come siempre con la boca"). Dentro del primero de ellos, sin embargo, aparece un adjetivo ("blanco") que forma un sintagma adjetivo (SAdj). Dentro del sintagma verbal, a su vez, aparece un adverbio ("siempre") que forma un sintagma adverbial (SAdv). El sintagma verbal

[3] A este tipo de enunciados con sintagma nominal y verbal se los suele denominar "cláusula" u "oración". Sobre esta terminología, véase Garrido (2009), capítulo 5.

contiene además un sintagma adposicional (SAdp) formado por la frase "con la boca". El núcleo de dicho sintagma es una adposición ("con"),[4] que se halla unida a un segundo sintagma nominal ("la boca"). Tanto en este sintagma como en el sintagma nominal principal ("el perro blanco") aparecen también especificadores (Esp), que aquí son los artículos "el" y "la".

2.3. Semántica y pragmática

Una perspectiva adicional que se utiliza para analizar el lenguaje es la semántica, que no tiene que ver con la estructura de los idiomas sino con su significado. Esta rama de la lingüística estudia las palabras en su carácter de signos que se refieren a algo, y que pueden ser meros "símbolos" (es decir, expresiones asociadas convencionalmente a cierto significado) o tener también un carácter icónico (es decir, que se parece a su significado, como puede ser el caso de las onomatopeyas) o indicativo (es decir, que cambia de significado según el contexto). Esta última situación puede ejemplificarse con el caso del pronombre personal "yo", que designa a una entidad diferente según lo use una persona u otra.[5]

El análisis semántico se concentra en buena medida en el estudio del vocabulario de los diferentes idiomas, y de cómo las palabras representan distintos "prototipos" más o menos amplios. El idioma español, por ejemplo, tiene dos palabras distintas para referirse a cierta relación de parentesco ("primo" y "prima"), que se usan según el referente sea varón o mujer. En inglés, en cambio, se utiliza una única palabra ("*cousin*"), y hay idiomas que, inversamente, tienen más palabras diferentes. Por ejemplo, en mandarín existen las palabras "*tangge*" (primo mayor por rama paterna), "*tangdi*" (primo menor por rama paterna), "*tangjie*" (prima mayor por rama paterna), "*tangmei*" (prima menor por rama paterna), "*biaoge*" (primo mayor por rama materna), "*biaodi*" (primo menor por rama materna), "*biaojie*" (prima mayor por rama materna) y "*biaomei*" (prima menor por rama materna).

Estas distintas estrategias de lexicalización de los conceptos pueden hacer también que un idioma tenga más de una palabra para el mismo significado (sinonimia), o una serie de palabras con significados más amplios y más estrechos (hiponimia). En castellano, por ejemplo, "lindo" y "bonito" son sinónimos, en tanto que "árbol" y "pino" tienen una relación de hiperónimo/hipónimo (ya que todos los pinos son árboles, pero solo algunos árboles son pinos). Hay también casos en los cuales a un idioma "le falta" una palabra que otros idiomas tienen. En tal caso, lo normal es reemplazar dicha palabra por un sintagma más complejo (por ejemplo, "arrollado de arroz japonés relleno") o bien incorporar directamente la palabra de un idioma que ya la posee (por ejemplo, "sushi").

La semántica tiene también relación con la pragmática, que es otro enfoque sobre el uso del lenguaje que tiene que ver con el efecto del contexto sobre el significado y la interpretación de las expresiones lingüísticas. Así, una expresión que, considerada aisladamente, tiene determinado significado, puede adoptar otro muy distinto (e, inclusive, opuesto) si es utilizada en determinado contexto. Analizado de ese modo, el contexto no solo ayuda a explicar lo que una expresión quiere decir (por ejemplo, la

[4] En español, las adposiciones son siempre preposiciones, porque se colocan al principio del sintagma adposicional. En otros idiomas (por ejemplo, vasco, turco, quechua) existen posposiciones (es decir, adposiciones que se colocan al final del sintagma).

[5] Para una explicación más completa de estos conceptos, véase Israel (2014).

palabra "aquí" se refiere a lugares distintos según dónde estén ubicados los hablantes) sino que también aporta ciertas "implicancias" que modifican su significado (por ejemplo, cuando una expresión se usa en sentido figurado en vez de usarse en sentido literal).

2.4. Patrones universales y tipología lingüística

El estudio de la complejidad de los idiomas está estrechamente relacionado con el hecho de que algunos fenómenos lingüísticos aparecen en todas las lenguas del mundo (o en la gran mayoría de ellas), en tanto que otros solo existen en algunos idiomas. Es precisamente en estos últimos casos en los cuales tiene interés analizar si un idioma es más o menos complejo que otro (y, eventualmente, elaborar un *ranking* que ordene a las lenguas de la más compleja a la más simple).

Para definir qué características pueden presentarse de manera más simple o más compleja en los distintos idiomas, resulta por lo tanto de utilidad definir primero qué otras características siguen un "patrón lingüístico universal". Estos patrones, llamados también "universales lingüísticos" (*language universals*), representan principios o resultados que parecen cumplirse de la misma manera en todos los idiomas. Algunos de ellos tienen un carácter general o "no implicativo" (por ejemplo, la idea de que todos los idiomas distinguen entre vocales y consonantes), en tanto que otros están expresados como reglas que dependen del cumplimiento de ciertas condiciones (por ejemplo, la idea de que si un idioma tiene "vocales nasalizadas", tiene que tener también vocales no nasalizadas –o sea, vocales comunes u "orales"– que correspondan a esas mismas vocales nasalizadas).[6]

Los universales lingüísticos pueden clasificarse también de acuerdo a la certeza respecto de su ocurrencia. Los ejemplos mencionados en el párrafo anterior parecen cumplirse en todos los casos en los que resultan aplicables, y serían, por lo tanto, "universales absolutos". En muchos otros casos, en cambio, los patrones que se observan en las distintas lenguas tienen un carácter probabilístico, y reciben el nombre de "tendencias" o "universales estadísticos". Siguiendo la clasificación sugerida por el inglés Bernard Comrie (1989), podemos decir entonces que existen cuatro tipos posibles de universales lingüísticos: absolutos no implicativos, absolutos implicativos, tendencias no implicativas y tendencias implicativas.

El estudio de los patrones lingüísticos universales difiere también según el enfoque metodológico que se adopte y la idea que se tenga respecto de su origen. En base a estos criterios, pueden distinguirse claramente dos escuelas de pensamiento, que son el "formalismo" y el "funcionalismo". El formalismo parte de la idea de que los universales lingüísticos se originan en factores innatos del ser humano, que tiene programado dentro de sí una determinada "gramática universal". Su enfoque metodológico, en general, tiende a ser deductivo, pues busca elaborar modelos teóricos formales de los cuales puedan derivarse los distintos patrones universales. Por el contrario, el funcionalismo enfatiza el carácter de los universales lingüísticos como medios que sirven para facilitar la comunicación (más que como resultados de aplicar una gramática genéticamente

[6] Un idioma cercano al español que tiene vocales nasalizadas es el portugués, lengua en la cual dichas vocales aparecen en palabras tales como "*são*" ("santo"), "*põe*" ("pone") y "*vim*" ("vine"). La versión no nasalizada de dichas vocales portuguesas es, por ejemplo, la que aparece en las palabras "*sal*" ("sal"), "*pôr*" ("poner") y "*vir*" ("venir").

programada). Su enfoque metodológico, por su parte, es netamente inductivo, ya que infiere a los patrones universales de la observación empírica de las distintas lenguas que se hablan en el mundo.[7]

Los enfoques formalista y funcionalista respecto de los universales lingüísticos se encuentran fuertemente asociados con los nombres de sus fundadores, que son los lingüistas estadounidenses Noam Chomsky (1965) y Joseph Greenberg (1966). Este último es considerado además como el principal iniciador de una rama de la lingüística que se conoce como "tipología", la cual se ocupa del estudio de las características que varían entre los distintos idiomas, de las limitaciones de dicha variación, y de las relaciones que pueden establecerse entre tales características. La propia naturaleza de esta rama del conocimiento lingüístico la pone casi totalmente dentro del enfoque funcionalista, ya que por definición la tipología trabaja con un número lo más grande y representativo posible de lenguas, y busca determinar las distintas características y las relaciones entre ellas aplicando –al menos en principio– un método inductivo.

Algunos autores (por ejemplo, los noruegos Halvor Eifring y Rolf Theil, 2005) señalan que los patrones universales y la tipología lingüística son en cierto sentido dos caras de la misma moneda, ya que el estudio de los primeros se refiere a los rasgos que las lenguas humanas tienen en común, en tanto que la tipología se enfoca en los rasgos en los cuales difieren. Dentro de estos últimos aparecen con singular importancia los fenómenos de complejidad, ya que una manera habitual que tienen los idiomas de diferir entre sí tiene que ver con aspectos que los vuelven más simples o más complejos. La propia lógica de que existan "universales implicativos" (absolutos o estadísticos) se entronca también con el análisis tipológico y con la evaluación de la complejidad de los idiomas. Si, por ejemplo, llegamos a la conclusión de que los idiomas que tienen palabras más cortas se caracterizan por tener enunciados más largos, esta regularidad puede considerarse como una tendencia implicativa. La misma, sin embargo, es también una relación entre características tipológicas de las lenguas, que a su vez implica una relación entre niveles de complejidad de sus distintos componentes.

Un último punto que vale la pena señalar respecto de la relación entre gramática universal y tipología de los idiomas es que, desde épocas relativamente recientes, parece estar produciéndose cierta convergencia entre los enfoques formalistas y funcionalistas. Esto se percibe, por ejemplo, en la obra del norteamericano Ray Jackendoff (2002), que es básicamente un intento de compatibilizar la idea de una gramática universal innata (que se encuentra programada en la mente de los hablantes) con factores del contexto (que tienen que ver con la función del lenguaje como medio de comunicación). También se ve en los intentos, cada vez más frecuentes, de construir modelos teóricos basados en hallazgos obtenidos a través de métodos inductivos, y en someter a los modelos hipotético-deductivos a diversos tipos de contrastación empírica.

3. Efectos de compensación

Hemos visto en la sección anterior que uno de los objetivos básicos de la tipología lingüística es hallar relaciones entre características de los distintos idiomas. Vimos también que, cuando una característica aparece siempre (o casi siempre) en conjunción con otra, eso puede revelar la existencia de un "universal implicativo" (absoluto o

[7] Para una buena comparación entre las características de estos dos enfoques, véase Mairal y Gil (2006).

estadístico). Algunas veces, sin embargo, no está claro cuál es la relación de implicación entre un fenómeno y otro, y no puede distinguirse bien si la presencia de cierto fenómeno implica la existencia de otro, o si es la presencia del segundo la que implica la existencia del primero.[8]

Cuando no es posible definir patrones universales que vayan claramente desde un fenómeno a otro, puede sin embargo ser útil detectar correlaciones entre distintas características lingüísticas. Dichas correlaciones implican que dos fenómenos aparecen en conjunto (correlación positiva), o bien que cuando uno de dichos fenómenos aparece el otro no lo hace (correlación negativa). En el caso de fenómenos que pueden medirse utilizando una escala de intensidad, también puede existir correlación entre un mayor nivel de cierto fenómeno y un mayor o menor nivel de otro. Por ejemplo, si medimos el número de fonemas vocálicos y consonánticos en una serie de idiomas, y hallamos que la cantidad de vocales está inversamente relacionada con la cantidad de consonantes, habremos encontrado una correlación negativa entre dos características que se miden usando variables numéricas.

Un caso particular, del cual puede ser un ejemplo la relación mencionada en el párrafo anterior, se da cuando la escala de medición que se usa para definir las características bajo estudio permite distinguir entre situaciones de mayor simplicidad y situaciones de mayor complejidad (absoluta o relativa). En tales casos, la presencia de correlación negativa puede verse como un indicio de la existencia de un "efecto de compensación" entre distintos aspectos de la complejidad de los idiomas (*language complexity trade-off*). Dicho efecto implica que un nivel de complejidad mayor para cierto componente de la lengua aparece en correspondencia con un nivel de complejidad menor para otro componente. Este tipo de relación puede ocurrir entre características que pertenezcan a la misma categoría lingüística (por ejemplo, fonología, morfología, sintaxis) o entre características que pertenezcan a distintas categorías.

3.1. La hipótesis de igual complejidad

La existencia o inexistencia de efectos de compensación ha sido utilizada en la literatura lingüística como un argumento a favor o en contra de una hipótesis según la cual "todos los idiomas son igualmente complejos" (*equal-complexity hypothesis*). Esta hipótesis tiene una tradición relativamente larga, que se remonta por lo menos a la primera mitad del siglo XX, a través de los autores de la llamada "lingüística descriptiva". Para estos autores, entre los que puede mencionarse a los estadounidenses Edward Sapir y Charles Hockett, la idea de igual complejidad tenía que ver con una ausencia de correlación entre la estructura de los idiomas y el nivel de civilización de los pueblos que habían creado dichos idiomas. Esta observación surgía en general del estudio de las lenguas amerindias norteamericanas, y de hecho los descriptivistas fueron los primeros que defendieron abiertamente la idea de que no existen "idiomas primitivos" (en el sentido de que todas las lenguas son igualmente capaces de expresar las distintas

[8] En la relación mencionada en la sección anterior entre vocales nasales y orales, por ejemplo, está claro que la presencia de vocales nasalizadas implica la existencia simultánea de vocales orales, pero que la inversa no es cierta (todos los idiomas conocidos tienen vocales orales, y solo una minoría tiene vocales nasalizadas). En la relación entre palabras y enunciados más cortos o más largos, en cambio, no está claro si la extensión de las palabras es la que impacta sobre la extensión de los enunciados o viceversa.

facetas de la cultura humana, y de expandirse y modificarse para hacerlo).[9]

Con la aparición de la corriente formalista o "generativista" de Noam Chomsky, y su concepción acerca de la existencia de una gramática universal, la idea de igual complejidad de los idiomas pasó a tener un fundamento basado en concepciones biológicas. Según esta corriente, la estructura de los idiomas está definida por características innatas del ser humano, y la diferencia entre las distintas lenguas tiene que ver con parámetros alternativos que cada idioma fija de manera diferente. Por ejemplo, así como en español los modificadores indirectos aparecen siempre detrás de los sustantivos que modifican (por ejemplo, "zapato de cuero"), en otros idiomas como el japonés aparecen adelante (por ejemplo, "*kawa no kutsu*", que literalmente significa "cuero-de zapato"). Pero esto no implica que una forma sea más compleja que la otra, sino que ambas son maneras alternativas de establecer reglas para cumplir con determinados principios gramaticales.

En su artículo sobre la evolución de la idea de igual complejidad entre los idiomas, el británico Geoffrey Sampson (2009) sostiene que la misma ha operado más como un "axioma" (es decir, como algo que se toma como dado, y que no intenta demostrarse) que como el resultado de analizar la estructura de las distintas lenguas. De hecho, los autores que han buscado evidencias empíricas de la hipótesis de igual complejidad se han encontrado más bien con lo contrario. Entre dichos autores sobresalen los de la llamada "tipología sociolingüística", tales como el inglés Peter Trudgill (2009). Esta corriente sostiene que los idiomas no solo no son igualmente complejos entre sí, sino que su complejidad depende esencialmente de dos factores socioculturales: el número de hablantes y el contacto con otras lenguas.

Para la lógica detrás de la tipología sociolingüística, los idiomas más complejos desde el punto de vista de su estructura suelen ser aquellos que habla poca gente, en lugares aislados y sin demasiado contacto con otros idiomas. Eso permite que se desarrollen complejidades que solo son útiles para la comunicación entre un conjunto reducido de personas que comparten la misma cultura. Cuando se da, en cambio, que un idioma pasa a ser usado por muchas personas de culturas distintas, y se vuelve además la segunda lengua de un grupo importante de hablantes, eso hace que su gramática tienda a simplificarse, y lo mismo suele ocurrir cuando el idioma entra en contacto con otras lenguas vecinas que le incorporan nuevas estructuras. Bajo esta concepción, esas estructuras nuevas solo se adoptan si sirven para reemplazar estructuras equivalentes más complejas que el idioma tenía antes de entrar en contacto con las otras lenguas, y todo eso conduce a un proceso de simplificación gramatical.[10]

Un elemento que, inversamente, parece jugar a favor de la complejidad de los idiomas es el tiempo. Esto lleva a que los idiomas más "nuevos" sean más simples que los más "viejos", y esa hipótesis ha sido desarrollada por el lingüista estadounidense John McWhorter (2001) para explicar la relativa simplicidad de los denominados "idiomas criollos" (*creole languages*). Estos idiomas han sido creados espontáneamente por grupos de hablantes de distintas lenguas que adoptan un idioma foráneo para entenderse entre sí, y que luego lo modifican hasta convertirlo en algo bastante diferente del original.

[9] Para una reseña de esta literatura, véase Joseph y Newmeyer (2012).

[10] En este punto existen, sin embargo, numerosas excepciones, sobre todo en el plano fonológico. Puede ocurrir, por ejemplo, que la incorporación de palabras de otros idiomas venga unida a la incorporación de fonemas que esos otros idiomas tienen y el idioma propio no.

En general, los idiomas criollos tienen un vocabulario basado en una "lengua madre", pero sus estructuras gramaticales suelen seguir principios (más simples) originados en otras lenguas. El palenquero, por ejemplo, es un idioma surgido a principios del siglo XVII en una zona de Colombia. Su base léxica es el castellano, pero tiene varias características estructurales de algunas lenguas africanas (por ejemplo, los pronombres posesivos se colocan después de los sustantivos) y carece de muchas de las sofisticaciones del español (por ejemplo, los adjetivos no tienen género ni número, y los verbos se conjugan de un modo mucho más simple).

También da la impresión de que el surgimiento de la lengua escrita ha tenido por efecto complejizar la gramática de los idiomas. En efecto, cuando un idioma empieza a escribirse, eso permite verbalizar ideas más complejas, porque se vuelve posible expresar pensamientos que no necesariamente tienen que ser comprendidos de manera inmediata (sino que pueden ser leídos y releídos varias veces hasta completar su comprensión). Para ello puede ser útil emplear de manera más frecuente ciertas estructuras más complejas (por ejemplo, oraciones subordinadas, o tiempos verbales pluscuamperfectos) y eso puede tener un efecto sobre la complejidad de la lengua como un todo.[11]

Por último, la complejidad de los idiomas parece ser también en cierta medida un fenómeno aleatorio, que depende de la historia particular de cada lengua y de variaciones probabilísticas que en unos idiomas se producen de determinada manera y en otros idiomas se producen de otra manera. Esta concepción implica que, en principio, cada aspecto de un idioma puede tomar valores más simples o más complejos, y tiene una probabilidad similar de adoptar dichos valores en cada lengua en particular. La realización efectiva de tales valores, sin embargo, será una sola para cada idioma, y resultará por lo tanto posible encontrarse con lenguas que sean complejas en varias dimensiones distintas, con lenguas que sean simples en esas mismas dimensiones, y con lenguas que combinen algunos aspectos simples con otros complejos.[12]

3.2. El enfoque sinergético

Lo expuesto en el apartado anterior parece indicar que la mayor parte de la evidencia empírica disponible debería llevarnos a descartar la hipótesis de igual complejidad de los idiomas. El hecho de que no todos los idiomas sean igualmente complejos, sin embargo, no necesariamente invalida la existencia de efectos de compensación. La pareja de lingüistas austríacos formada por Gertraud Fenk-Oczlon y August Fenk (2011), por ejemplo, ha mostrado que, si bien los idiomas no tienen por qué ser equivalentes en términos de ningún tipo de medida de complejidad global, existen varios aspectos en los cuales es posible detectar la presencia de *trade-offs*.

Una posible explicación para este tipo de efectos de compensación proviene del campo de la psicología cognitiva. Tal como se muestra en un artículo del británico Simon Kirby en conjunto con otros autores (2015), los idiomas pueden ser vistos como productos de cierto tipo de evolución cultural, en la cual operan presiones contrapuestas para que las lenguas sean al mismo tiempo lo más comprimidas posibles (a fin de facilitar su aprendizaje) y lo más expresivas posibles (a fin de facilitar su uso como medios de

[11] Para una explicación más completa de este punto, véase McWhorter (2003), capítulo 6.

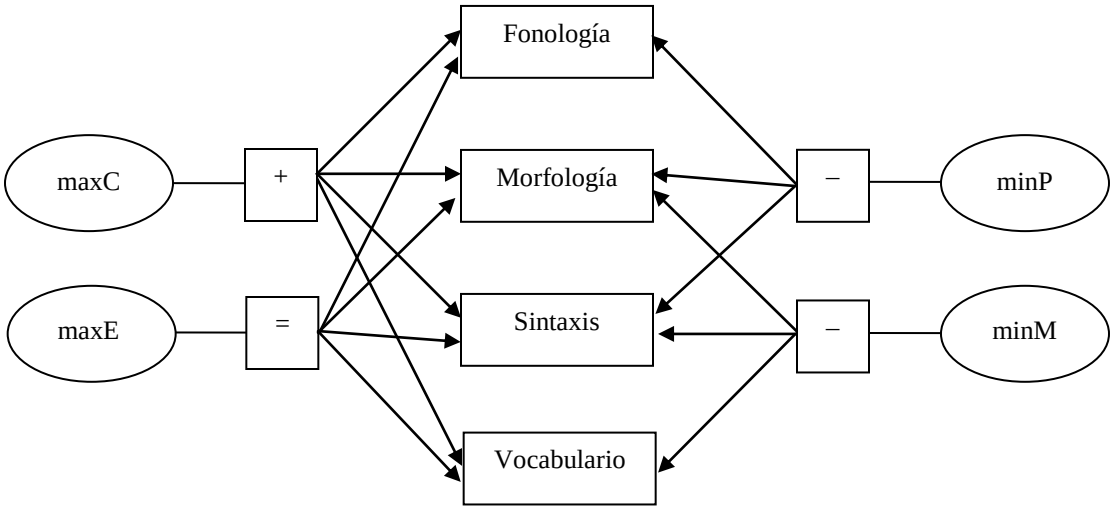
[12] Para una explicación más detallada de este tipo de teorías aplicadas a la diversidad de los idiomas, puede consultarse el libro del lingüista español José Luis Mendívil (2009).

comunicación). Es justamente dicha evolución la que genera la aparición de cierta estructura, que es en definitiva la que sirve para contrapesar los efectos de una tendencia a simplificar el lenguaje y de otra que busca complejizarlo.

Otra explicación alternativa es la que surge de aplicar el enfoque conocido como "lingüística sinergética" (*synergetic linguistics*), originado en la obra del alemán Reinhard Köhler (1987). Este enfoque, extraído de la teoría general de los sistemas, considera que el lenguaje es un sistema organizado y autorregulado, cuyas propiedades provienen de la interacción de distintos requisitos.

Si empleamos la lógica implícita en la lingüística sinergética, el tema de los efectos de compensación puede plantearse suponiendo que el "sistema agregado" conocido como lenguaje está compuesto por varios subsistemas, que serían sus distintos componentes (por ejemplo, la fonología, la morfología, la sintaxis, el vocabulario). Cada uno de dichos subsistemas tiene cierto nivel de complejidad, y dicho nivel de complejidad es útil para cumplir con algún requisito del sistema como un todo, pero inconveniente para el cumplimiento de algún otro.

Cuatro de los requisitos usualmente mencionados por la literatura sobre lingüística sinergética son la necesidad de proveer expresiones para distintos significados (codificación), la necesidad de utilizar el menor número posible de elementos para producir dichas expresiones (producción), la necesidad de ordenar dichos elementos de modo que sea fácil recordarlos (memorización) y la necesidad de modificar el idioma lo menos posible para que pueda ser comprendido (estabilidad).[13]



Lo expuesto en el párrafo anterior aparece representado en el gráfico adjunto, en el cual hemos supuesto que la complejidad de cada uno de los cuatro subsistemas identificados (fonología, morfología, sintaxis y vocabulario) contribuye a maximizar la facilidad de codificación (maxC), a minimizar el esfuerzo de producción del lenguaje (minP), a minimizar el esfuerzo de memorización del mismo (minM) y a maximizar la estabilidad de la lengua como sistema (maxE). La forma en la cual se cumplen estos requisitos, sin embargo, difiere según el subsistema y según la función. En este caso,

[13] Para una buena reseña de esta literatura, véase Köhler (2005).

puede suponerse que una mayor complejidad de cada componente ayuda a mejorar la codificación (y de ahí el signo "+" que aparece asociado con "maxC"), pero que si la fonología, la morfología o la sintaxis son más complejas, eso es malo en términos del esfuerzo de producción del lenguaje (y de ahí el signo "–" asociado con "minP"). Algo parecido ocurre con el esfuerzo de memorización, cuya minimización (minM) está asociada con menores niveles (–) de complejidad morfológica, sintáctica y del vocabulario utilizado. Por último, para contribuir a la estabilidad del sistema, lo ideal es que el mismo se modifique lo menos posible (y de ahí el signo "=" que aparece asociado con "maxE"). Esto último puede implicar, según los casos, mantener un nivel de complejidad relativamente alto o relativamente bajo.

Una forma de apreciar cómo un sistema de este tipo produce correlaciones negativas entre los niveles de complejidad de los distintos subsistemas es pensar que sus múltiples objetivos en términos de facilidad de codificación, esfuerzo de producción, esfuerzo de memorización y estabilidad determinan una especie de "función de valuación promedio" del lenguaje. Esta función debería depender positivamente de la facilidad de codificación y de la estabilidad, y negativamente del esfuerzo de producción y de memorización.[14]

Si, dentro de alguno de los elementos de la función de valuación, las variables correspondientes a nuestros cuatro componentes de la lengua (fonología, morfología, sintaxis y vocabulario) aparecen interactuando entre sí, la elección de un nivel de complejidad para determinado componente implica en general una relación inversa con la complejidad de los otros componentes. Dicha relación generará correlaciones negativas entre los distintos niveles de complejidad, ya que la forma óptima de lograr determinado objetivo parcial (por ejemplo, mejorar la codificación) sin perjudicar en exceso el logro de otro objetivo (por ejemplo, sin incrementar demasiado el esfuerzo de producción o de memorización) consiste en elegir niveles de complejidad parcial que no sean ni demasiado bajos ni demasiado altos.

4. Medición de la complejidad lingüística

Una distinción importante que puede hacerse entre las diferentes maneras de medir la complejidad idiomática es la que las divide en "medidas teóricas" (o tipológicas) y "medidas empíricas". Las medidas teóricas surgen de tomar en cuenta datos que provienen de las gramáticas de los diferentes idiomas, es decir, de las reglas que utilizan dichos idiomas para asignarles significados a los distintos sonidos. A fin de poder medirlos, tales datos se clasifican según los conceptos de la tipología lingüística, que nos permite definir en qué reglas difieren los idiomas, y si dichas reglas implican una complejidad mayor o menor.

Las medidas empíricas, en cambio, se basan en datos que surgen de palabras concretas o, más generalmente, de textos (es decir, de conjuntos de enunciados relacionados entre sí, que forman una unidad de sentido). Si bien su medición también sigue en cierto modo criterios tipológicos, dicha relación es indirecta, pues las medidas que se obtienen no se refieren a una lengua en sí sino a un recorte de dicha lengua, representado por un determinado grupo de palabras. Un ejemplo de medida teórica es el

[14] Para una explicación de la lógica detrás de este tipo de formulación, puede consultarse el artículo de Wimmer y Altmann (2007).

número de vocales que tiene el inventario de un idioma (por ejemplo, las cinco vocales del idioma español), en tanto que un ejemplo de medida empírica es el número promedio de fonemas por palabra (que no es el mismo para cualquier texto escrito en español, sino que varía según el ejemplo concreto que consideremos).

4.1. Medidas teóricas o tipológicas

Debido a la naturaleza del lenguaje, formado por una serie de componentes diversos, no existe (aún) una medida de su complejidad teórica que pueda considerarse como universalmente aceptada. Lo que hallamos en la literatura lingüística son más bien medidas parciales, que apuntan a cuantificar cierto componente del lenguaje (fonología, morfología, sintaxis, vocabulario) o bien a medir algún aspecto particular dentro de uno de dichos componentes. En el cuadro que figura a continuación aparecen algunos ejemplos, clasificados por componente y por tipo de medida (numérica o categórica). El cuadro menciona también referencias a trabajos representativos, en los cuales las distintas medidas han sido empleadas.

Medida	**Tipo**	**Referencia**
Fonología		
- Número de consonantes	Numérica	Maddieson (2007)
- Número de vocales	Numérica	Maddieson (2007)
- Complejidad silábica	Numérica	Shosted (2006)
Morfología		
- Número de inflexiones verbales	Numérica	Shosted (2006)
- Marcación del plural en los sustantivos	Categórica	Nichols (2009)
- Marcación del género en los sustantivos	Categórica	Nichols (2009)
Sintaxis		
- Orden de palabras fijo o variable	Categórica	Sinnemäki (2008)
- Existencia de sujetos tácitos	Categórica	Biberauer y otros (2014)
- Asimetría entre frase afirmativa y negativa	Categórica	Parkvall (2008)
Vocabulario		
- Pronombre de primera persona plural	Categórica	Nichols (2009)
- Número de contrastes en demostrativos	Numérica	Parkvall (2008)
- Existencia de artículos definidos e indefinidos	Categórica	Parkvall (2008)

Tal como puede observarse, algunas de las medidas propuestas son de fácil definición y conceptualización. El número de consonantes y de vocales, por ejemplo, surge directamente de examinar el inventario de fonemas de cada idioma, en tanto que ciertas variables categóricas como la marcación del género y del plural en los sustantivos es también algo fácilmente identificable consultando las reglas gramaticales de las distintas lenguas. Este último tipo de variables categóricas, además, sirve para clasificar a los idiomas, ya que un idioma que posee la característica bajo estudio (por ejemplo, que marca el plural en los sustantivos, como es el caso del castellano) es considerado complejo, y un idioma que no posee dicha característica (por ejemplo, que no indica si el sustantivo es singular o plural, como es el caso del chino mandarín) es considerado simple en ese aspecto.

Otras medidas teóricas de complejidad son, en cambio, más difíciles de definir y de calcular. El estadounidense Ryan Shosted, por ejemplo, publicó un artículo en el año

2006 en el cual comparó una medida teórica de complejidad silábica con otra medida teórica de complejidad verbal, buscando encontrar algún tipo de correlación entre ellas. Para medir la complejidad silábica, dicho autor realizó un cálculo de todos los ejemplos posibles de sílabas de cada lengua, basándose en el número de fonemas consonánticos, el número de fonemas vocálicos y los tipos de sílaba admisibles en cada idioma (por ejemplo, consonante+vocal, vocal+consonante, consonante+vocal+consonante, etc.). Llegó así a la conclusión de que algunas lenguas tienen más de 50.000 ejemplos posibles de sílabas (por ejemplo, el vietnamita, que tiene 52.007), en tanto que otros tienen menos de 100 casos (por ejemplo, el idioma koiari, que se habla en Papua Nueva Guinea, y que tiene solo 70 ejemplos posibles de sílabas).[15]

En cuanto a su medida de complejidad verbal, Shosted usó como criterio el número de posibles inflexiones, y tomó como base los cálculos realizados previamente por el alemán Balthasar Bickel y la norteamericana Johanna Nichols (2005). Estos autores calcularon el número de "categorías por palabra" que pueden indicarse mediante afijos verbales, y llegaron a la conclusión, por ejemplo, de que el inglés solo tiene dos categorías relevantes (tiempo y persona) y que, en cambio, el español tiene cinco (modo, tiempo, aspecto, persona y número). Por ese hecho el español es considerado morfológicamente más complejo que el inglés, pero menos complejo que otros idiomas tales como el turco (que tiene siete categorías de inflexión verbal: modo, tiempo, aspecto, persona, número, polaridad y evidencialidad).

Usando promedios de varias medidas teóricas, algunos autores han intentado elaborar también "índices de complejidad total". Johanna Nichols (2009), por ejemplo, utiliza un promedio de 17 indicadores, que agrupa a su vez en cinco categorías (fonología, morfología, tipos de palabras, sintaxis y vocabulario) y que usa para hacer un *ranking* de complejidad que abarca 68 idiomas. El primer lugar en dicho *ranking* es ocupado por el idioma ingusetio (que se habla en Chechenia y en otras zonas de Rusia cercanas a esa región), en tanto que el último lugar (es decir, el de mayor simplicidad) es ocupado por el idioma rama (que es una lengua indígena que se habla en Nicaragua). El español ocupa un lugar relativamente bajo en el ordenamiento de Johana Nichols, ya que está en el puesto 58.

Otro autor que ha elaborado un *ranking* de complejidad usando diferentes indicadores es el sueco Mikael Parkvall (2008), quien utilizó información de la primera edición del Atlas Mundial de Estructuras Lingüísticas (WALS, por su sigla en inglés). Para llevarlo a cabo, Parkvall seleccionó 47 características informadas en dicho atlas, más otras 6 características tomadas de otras fuentes, y llevó a cabo un ordenamiento de 155 idiomas. El idioma más complejo en ese *ranking* es el burushaski (que es una lengua que se habla en el norte de Pakistán), en tanto que el más simple es el sango (que es el idioma nacional de la República Centroafricana). A diferencia de lo que ocurre en el *ranking* elaborado por Nichols, en la lista de Parkvall el español ocupa un lugar relativamente alto (el puesto 11).

La última edición electrónica del Atlas Mundial de Estructuras Lingüísticas, compilada en 2013 por Matthew Dryer y Martin Haspelmath, y publicada por el Instituto Max Planck (Alemania), incluye una muestra de 100 idiomas que han sido seleccionados

[15] El castellano no figura entre las lenguas analizadas por Shosted en su artículo, pero basándonos en la información que figura en la gramática de la Real Academia Española (2011) podemos inferir que tiene 11.642 ejemplos posibles de sílabas.

como representativos de la diversidad lingüística y geográfica de las lenguas que se hablan en el mundo. En base a ello, los editores del WALS les han insistido a los autores de los distintos capítulos que incluyan esos idiomas en sus análisis "siempre que sea posible", lo cual hace que para tales idiomas haya una cobertura mucho mayor que para el resto.

Aprovechando ese hecho, en un trabajo anterior (Coloma, 2015a) decidimos tomar esa muestra de 100 idiomas del WALS y recopilar datos acerca de 60 características para las cuales la información disponible nos permitiera clasificar a los idiomas de acuerdo con alguna medida de complejidad. Siguiendo un método parecido a los utilizados por Parkvall y por Nichols, clasificamos dichas características en categorías (que en nuestro caso son cinco: fonología, morfología, sintaxis, sistema verbal y vocabulario), y calculamos índices de complejidad por categoría y también un índice de complejidad total. En el primer lugar del *ranking* basado en dicho índice de complejidad total quedó el abjasio (que es un idioma caucásico noroccidental que se habla en una zona cerca del Mar Negro) y en el último lugar quedó el tailandés. El español terminó en un puesto relativamente alto (el número 22), más cerca del resultado obtenido por Parkvall que del obtenido por Nichols.[16]

4.2. Medidas empíricas

Si bien las medidas teóricas o tipológicas tienen la ventaja de que representan un conjunto de elementos que puede considerarse relativamente general dentro de la estructura de cada idioma, su principal desventaja es que no pueden ser observadas de manera directa, sino que necesitan de alguien que previamente las extraiga de la gramática del idioma en cuestión. Alguien que no habla español, por ejemplo, no puede deducir que en nuestro idioma los sustantivos están clasificados por género (y que solo tenemos dos categorías: masculino y femenino), porque para eso necesita conocer el "funcionamiento" de la lengua (por ejemplo, saber que los sustantivos masculinos van precedidos por los artículos "el" y "un", y los femeninos por los artículos "la" y "una", que cuando son modificados por un adjetivo este puede cambiar de terminación según el sustantivo sea masculino o femenino, etc.).

Parte de esos problemas pueden obviarse si, en vez de utilizar medidas teóricas, tratamos de evaluar la complejidad de un idioma usando medidas empíricas. Por ejemplo, si tenemos un texto escrito en cierto idioma, y el mismo está dividido en enunciados, palabras y fonemas, resulta relativamente simple calcular el número de palabras por enunciado o el número de fonemas por palabra, aunque no conozcamos el significado de dichas palabras. Usar medidas empíricas en vez de medidas teóricas tiene además la ventaja de que, al menos en principio, existen menos dudas respecto de cuán real es la información que se está utilizando, y menos problemas de interpretación cuando hay varios criterios alternativos para definir el mismo fenómeno.

Para poner un ejemplo del último punto señalado en el párrafo anterior, tomemos

[16] Como veremos más adelante, estas diferencias en los *rankings* dependen de la lista de idiomas que uno incluya, pero sobre todo del peso que se le dé a cada componente dentro del índice de complejidad total. En el caso del español, la mayoría de los ordenamientos coinciden en destacar su relativa simplicidad fonológica, pero también la relativa complejidad de su sistema verbal y de algunos aspectos ligados con su vocabulario. Según se le otorgue más peso a uno u otro factor, el castellano termina en una posición global más alta o más baja.

una variable teórica común para medir la complejidad de la fonología de una lengua, como es el número de vocales que utiliza. En el caso del idioma ruso, por ejemplo, hay dos escuelas fonológicas tradicionales (una de las cuales está asociada con la ciudad de Moscú y la otra con la ciudad de San Petersburgo) que dan respuestas distintas para la definición de dicha variable. Una de dichas escuelas (la de Moscú) sostiene que el ruso tiene cinco vocales (las mismas que el castellano), en tanto que la otra (la de San Petersburgo) sostiene que tiene seis (las cinco del castellano, más una sexta vocal parecida a la "y" del idioma guaraní). La causa de esta divergencia es que los fonólogos moscovitas suelen considerar a la sexta vocal del ruso como un alófono del fonema /i/, en tanto que los petersburgueses la consideran como un fonema independiente. Nótese que esto no tiene que ver con que la gente pronuncie distinto las vocales en Moscú y en San Petersburgo, sino con diferencias teóricas en la clasificación de los sonidos dentro del concepto de fonema.[17]

El principal problema de utilizar medidas empíricas para evaluar la complejidad de los idiomas, sin embargo, es la representatividad de los textos utilizados como base para calcular dichas medidas, así como la comparabilidad de los mismos. En general, los autores tratan de solucionar el problema de comparabilidad empleando el mismo texto en distintos idiomas (si es que eso resulta posible). En cuanto a la representatividad del texto como manifestación de determinado idioma, en cambio, la decisión suele tener que ver con el tema que se está analizando. Para algunos estudios resulta importante que se trate de textos obtenidos originalmente de manera oral, en tanto que para otros puede servir igualmente bien que se trate de un texto escrito. Muchas veces, además, el requisito de representatividad puede entrar en conflicto con el de comparabilidad: un texto escrito en un idioma y traducido luego a otro puede ser muy representativo de su lengua de origen, pero poco representativo de la lengua a la cual ha sido traducido.[18]

Las medidas empíricas de complejidad que resultan más fáciles de calcular son los denominados "cocientes lingüísticos" (*linguistic ratios*). Los mismos no son otra cosa que divisiones entre medidas del tamaño de determinado texto, expresadas en distintas unidades (fonemas, sílabas, palabras, enunciados, etc.). Los austríacos Gertraud Fenk-Oczlon y August Fenk (1999), por ejemplo, utilizaron este tipo de cocientes para evaluar distintos aspectos de la complejidad de un grupo de 34 idiomas, tomando como base un conjunto de 22 enunciados originalmente escritos en alemán. Los cocientes que emplearon fueron los siguientes: fonemas por sílaba, sílabas por palabra, sílabas por enunciado, y palabras por enunciado.

Estas medidas empíricas de complejidad pueden relacionarse de manera general con algún campo del análisis lingüístico. Por ejemplo, el número de fonemas por sílaba puede verse como una medida de cierto aspecto de la complejidad fonológica de los idiomas, en tanto que el número de sílabas por palabra debería estar fuertemente correlacionado con alguna medida de complejidad morfológica (ya que, cuanto mayor es el número de sílabas por palabra, mayor debería ser el número de morfemas por palabra).

[17] Para una discusión un tanto airada de estas diferencias (que incluye además una toma de posición en el tema), véase Canepari (2005), capítulo 8.

[18] Este problema, por ejemplo, ha sido señalado en relación a los textos bíblicos. Por el gran número de traducciones que tienen, dichos textos son candidatos naturales para ser usados en el cálculo de medidas empíricas, en contextos de comparación entre distintas lenguas. En algunos de los idiomas a los cuales la Biblia ha sido traducida, sin embargo, las referencias que en ella aparecen pueden resultar totalmente extrañas y antinaturales.

La cantidad de palabras por enunciado, por su parte, puede servir como medida aproximada de la complejidad sintáctica de un idioma, en tanto que la cantidad de sílabas por enunciado debería ser una especie de medida global de complejidad morfosintáctica.

Si introducimos algo más de estructura gramatical en la definición de las medidas empíricas de complejidad, podemos calcular también otros indicadores que hagan uso de dicha estructura. Los canadienses Connie Adsett y Yannick Marchand (2010), por ejemplo, tomaron en cuenta la idea de que la sílaba más común en la generalidad de los idiomas es la que consiste en juntar una consonante con una vocal, y computaron la frecuencia relativa de dicho tipo de sílaba en una serie de "corpus" (es decir, de bases de datos formadas por numerosos textos), correspondientes a nueve idiomas europeos. Para calcular esa frecuencia relativa, hicieron el cociente entre la cantidad de sílabas formadas por "consonante + vocal" y el total de sílabas de cada corpus, obteniendo así un porcentaje que funciona como medida de la "simplicidad silábica" de cada idioma.

Aplicando un criterio similar, el lingüista estadounidense Max Bane (2008) tomó un conjunto de textos escritos en 20 idiomas (de distintas procedencias) y aplicó un algoritmo a través del cual calculó el cociente entre la cantidad de afijos y la cantidad total de morfemas (es decir, de "afijos + raíces"). Dicho cociente puede interpretarse como una medida empírica de la complejidad morfológica, que toma un valor más elevado si el idioma le inserta a las palabras muchos afijos, y un valor que tiende a cero si utiliza solamente palabras con un único morfema. Otro autor que postuló una serie de indicadores empíricos destinados a medir la complejidad en una categoría en particular es el alemán Benedikt Szmrecsányi (2004), quien analizó el tema de la complejidad sintáctica comparando tres medidas: palabras por enunciado, sintagmas por enunciado, y un "índice de complejidad sintáctica" que busca reflejar la existencia de distintos tipos de elementos.

En cuanto a la complejidad léxica o del vocabulario utilizado, la medida empírica más común es la que se conoce con el nombre de "cociente entre tipos y ocurrencias" (*type-token ratio*). Esta medida se obtiene contando el número total de palabras de determinado texto (ocurrencias) y el número de palabras distintas que aparecen en dicho texto (tipos). Por ejemplo, la oración "el pueblo español habla el idioma español" tiene 7 palabras (ocurrencias) pero solo 5 tipos, ya que las palabras "el" y "español" aparecen dos veces cada una. Su *type-token ratio*, por lo tanto, es igual a 0,7143 (número este que surge de hacer el cociente entre 5 y 7).

En un artículo publicado en el año 2014, el finlandés Kimmo Kettunen señaló que el cociente entre tipos y ocurrencias no solo brinda información para evaluar la complejidad léxica de un texto, sino también la complejidad morfológica del mismo. Comparando los valores obtenidos para dicho indicador, aplicado al texto de la Constitución de la Unión Europea publicado en 21 idiomas distintos, Kettunen llegó a la conclusión de que los *type-token ratios* presentaban una alta correlación positiva con ciertas medidas de complejidad morfológica, tales como el cociente entre afijos y morfemas sugerido por Bane (2008), y la medida teórica de complejidad verbal calculada por Bickel y Nichols (2005) y utilizada por Shosted (2006).

El cociente entre tipos y ocurrencias también puede ser usado para medir relaciones a nivel de sílabas y de fonemas. Por ejemplo, la oración transcripta anteriormente ("el pueblo español habla el idioma español") tiene 15 sílabas y 34 fonemas, pero el número de tipos silábicos y fonémicos es respectivamente igual a 11 y a

12. Sus cocientes entre tipos y ocurrencias, por ende, son iguales a 0,73 (para las sílabas) y a 0,35 (para los fonemas). Estas medidas pueden verse como indicadores de distintos aspectos de la complejidad del texto analizado, en una escala que va de 0 (mayor simplicidad) a 1 (mayor complejidad).

A efectos de ilustrar con un ejemplo concreto algunas de las medidas empíricas reseñadas en los párrafos anteriores, utilizaremos el texto de una fábula de Esopo cuyo título es "El viento norte y el sol". Dicha fábula tiene la característica de que viene siendo empleada desde hace más de 100 años por la Asociación Fonética Internacional como "espécimen" o texto ejemplificador de los sonidos de un gran número de lenguas que se hablan en el mundo. El texto en castellano más utilizado en esa función es el que aparece en un artículo de los fonetistas españoles Eugenio Martínez Celdrán, Ana Fernández Planas y Josefina Carrera (2003), y es el siguiente:

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

Si contamos el número de enunciados, palabras, sílabas y fonemas en este texto, hallaremos que el mismo está compuesto por 9 enunciados, 97 palabras, 193 sílabas y 425 fonemas. Esto nos permite calcular distintos cocientes lingüísticos, entre los cuales pueden mencionarse el número de fonemas por sílaba (igual a 2,2021), el número de sílabas por palabra (igual a 1,9897), el número de fonemas por palabra (igual a 4,3814) y el número de palabras por enunciado (igual a 10,78).

Como la descripción fonética del texto, y el conocimiento que uno tiene respecto de la lengua española, nos permiten individualizar las sílabas incluidas en cada una de las palabras de esta fábula, podemos computar también de manera relativamente directa el porcentaje de sílabas constituidas por "consonante + vocal". Dicho porcentaje es igual al 47,54%, porcentaje que se eleva al 54,10% si le sumamos las sílabas conformadas por una única vocal (sin ninguna consonante). Esto puede verse como un indicador de la simplicidad silábica del texto (y, de manera indirecta, también de la simplicidad silábica de la lengua castellana).

Si ahora analizamos las palabras de "El viento norte y el sol" en términos de los morfemas que las constituyen, veremos que las 97 palabras del texto pueden dividirse en un total de 137 morfemas. Esto se debe a que hay 70 palabras que tienen un único morfema, 14 palabras que tienen dos morfemas (una raíz y un afijo), y 13 palabras que tienen tres morfemas (una raíz y dos afijos). El cociente entre afijos y morfemas, por lo tanto, es igual a 0,2920, lo cual representa un indicador de la complejidad morfológica del texto analizado. Otro posible indicador para medir el mismo fenómeno sería hacer el cociente entre morfemas y palabras, que en este caso da un número igual a 1,4124.

En lo que se refiere a la complejidad sintáctica del texto, un primer indicio es el que nos da el cociente entre palabras y enunciados. Si analizamos sintácticamente cada enunciado, sin embargo, podemos computar también el número de sintagmas, que en este caso es igual a 58 (22 sintagmas nominales, 14 verbales, 15 preposicionales, 5 adjetivos y 2 adverbiales). Si dividimos dicho número por los 9 enunciados identificados, obtenemos un cociente igual a 6,44 sintagmas por enunciado.

En lo que se refiere a la complejidad léxica, la misma puede evaluarse a través del cociente entre tipos y ocurrencias de las distintas palabras. Para ello debemos observar que, entre las 97 palabras del texto, hay 37 que son repeticiones de otras palabras que figuran en el mismo texto ("el", "viento", "sol", "viajero", "en", "que", etc.). Esto hace que el número de "tipos" sea igual a 60, y que el *type/token ratio* en términos de palabras sea igual a 0,6186. Calculado en términos de fonemas, dicho cociente es igual a 0,0494, ya que los 425 sonidos que se usan al leer esta fábula son en realidad repeticiones de 21 fonemas (todos los que tiene la lengua castellana menos el que corresponde a la letra "ñ", que no aparece en ninguna palabra).

4.3. La complejidad de Kolmogorov

El paso siguiente en cuanto a sofisticación en la medición empírica de la complejidad de los idiomas es el concepto de "complejidad de Kolmogorov", nombre que hace referencia a la obra del matemático ruso Andrei Kolomogorov (1963). Este concepto busca medir la información contenida en determinada secuencia de caracteres a través del algoritmo más simple que sea capaz de generar la serie de caracteres en cuestión. Cuanto más breve sea dicho algoritmo, más simple será la secuencia, y cuanto más largo, más compleja.

Supongamos, por ejemplo, que queremos comparar la complejidad de las palabras "lalala", "runrun" y "salero". Las tres palabras están compuestas por seis letras, pero mientras la primera consiste en la repetición de la sílaba "la" tres veces consecutivas, y la segunda consiste en la repetición de la sílaba "run" dos veces consecutivas, la tercera está compuesta por tres sílabas distintas (que, además, no repiten ninguna letra). La palabra "lalala", por lo tanto, puede escribirse de manera compacta como "3*la" (es decir, 3 veces "la"), en tanto que la palabra "runrun" puede escribirse como "2*run" (es decir, 2 veces "run"), pero no hay forma de escribir "salero" de una manera más compacta que "salero". Si ahora contamos el número de caracteres que hay que usar para escribir "3*la" y "2*run", vemos que para la primera secuencia necesitamos 4 caracteres y para la segunda necesitamos 5. La complejidad de Kolmogorov de "lalala", por lo tanto, puede calcularse como el cociente entre el número de caracteres de la versión comprimida y el número total de caracteres de la palabra (es decir, 4 dividido 6), lo cual da un resultado igual a 0,67. La complejidad de Kolmogorov de la palabra "runrun", en cambio, es igual a 0,83 (que es el cociente entre 5 y 6), y la de "salero" es igual a 1 (ya que tanto la versión comprimida como la original tienen 6 caracteres).

Cuando uno quiere aplicar el concepto de complejidad de Kolmogorov a una serie de caracteres más larga que una palabra, el cómputo manual de este índice se vuelve virtualmente imposible. Sin embargo, la aparición de programas informáticos especializados en la compresión de archivos ha permitido llevar a cabo dicho cómputo de manera relativamente sencilla. La manera de hacerlo es crear primero un archivo de texto con el material que se quiere analizar, y luego compactar dicho archivo. La complejidad de Kolmogorov del texto en cuestión no es otra cosa que el cociente entre el tamaño del archivo compactado y el tamaño del archivo original, medidos ambos en las mismas unidades (por ejemplo, en *bytes*).

Tal como ha sido definida en los párrafos anteriores, la complejidad de Kolmogorov es una medida global de la complejidad de un texto, pero existe una variación de dicho concepto, propuesta por Katharina Ehret y Benedikt Szmrecsányi

(2015), que sirve para medir aspectos específicos de dicha complejidad. Para aplicarla deben utilizarse, además del texto original, dos "versiones alteradas" del mismo: una que elimina al azar el 10% de las letras (versión "morfológicamente alterada"), y otra que elimina al azar el 10% de las palabras (versión "sintácticamente alterada"). Estas versiones alteradas, a su vez, son sometidas a un procedimiento de compresión, y los elementos que se utilizan luego para medir la complejidad son los tamaños de los respectivos archivos que contienen las versiones alteradas comprimidas.

Una sofisticación aún más grande que la implícita en las medidas de complejidad de Kolmogorov, y en las variaciones existentes en base a la misma, es la que surge de computar el periodograma de un texto. El mismo implica llevar a cabo un cálculo de la correlación que existe entre los distintos elementos de una serie (es decir, de la frecuencia con la cual se repiten dichos elementos) agrupados en conjuntos de tamaño creciente. De ese modo, el periodograma empieza calculando con qué frecuencia se repiten las letras una detrás de la otra (es decir, con qué probabilidad hallamos dos letras iguales seguidas), pasa luego a calcular con qué frecuencia se repiten los conjuntos de dos letras, luego de tres letras, y así sucesivamente. Esto genera un diagrama de dichas repeticiones, que de algún modo representa una "radiografía de la estructura del lenguaje".

En un trabajo que utiliza periodogramas para comparar las características implícitas en una serie de corpus escritos en 15 lenguas distintas, el español Fermín Moscoso (2011) llegó a la conclusión de que la estructura de dichas lenguas es relativamente similar. Encontró así que, en general, los idiomas repiten poco sus elementos cuando se los computa a nivel de secuencias cortas (letras, sílabas, morfemas), que tienen un nivel de repetición intermedio cuando se los analiza a nivel de secuencias intermedias (palabras, sintagmas) y que, en cambio, tienen un nivel de repetición más alto cuando se los analiza a nivel de secuencias largas (enunciados y párrafos). Dichos niveles "bajos", "intermedios" y "altos" están definidos respecto de la estructura que se supone que tendrían los textos si constaran de elementos puramente aleatorios, lo cual lleva a concluir que el lenguaje se caracteriza por ser un modo de componer textos que ordena sus símbolos de modo de se repitan poco cuando los mismos están unos al lado de otros (es decir, cuando se usan para formar sílabas) pero que sí exhibe comparativamente muchas repeticiones ligadas con la estructura de sus enunciados y sus párrafos (relacionadas, por ejemplo, con el orden de las palabras y con el tamaño de las oraciones).

4.4. La ley de Menzerath

Las medidas cuantitativas del lenguaje, tanto teóricas como empíricas, han sido (desde hace ya muchos años) objeto de estudio por parte de investigadores que buscaron encontrar regularidades estadísticas entre ellas. Esto ha dado pie al descubrimiento de algunas "leyes" interesantes, que encontraron relaciones que en algunos casos parecen extenderse a dominios que no son exclusivamente lingüísticos.

Uno de los primeros resultados de ese tipo es la llamada "ley de Grimm", que originalmente se aplicó a un fenómeno que tuvo lugar en las lenguas germánicas a comienzos de la Edad Media, pero que después se descubrió que tenía alcances mucho más generales. Esa ley, cuyo nombre hace referencia a la obra de Jacob Grimm (1822), expresa la idea de que el cambio fonético es regular, y que si un grupo de palabras de cierto idioma modifica su pronunciación en determinado sentido, dicho cambio se

extenderá a todas las palabras en su mismo contexto fonético. En el castellano, por ejemplo, un fenómeno de ese tipo se produjo a comienzos del siglo XVII, cuando todas las palabras que se escriben con la letra "j", y que en esa época se pronunciaban con un sonido parecido al de la "j" francesa o al de la "sh" inglesa, pasaron a pronunciarse con el sonido que se utiliza actualmente.[19]

La primera regularidad lingüística puramente cuantitativa, sin embargo, fue la postulada por el norteamericano George Zipf (1935), y tiene que ver con la relación entre la frecuencia de las palabras de un texto (f) y el *ranking* en el cual se ordenan esas palabras en virtud de dicha frecuencia (r). La "ley de Zipf", tal como se la conoce desde entonces, postula que ambas variables tienen una relación del tipo "$f = a \cdot r^b$", y que el valor de "b" tiende a ser igual a -1 cuando el número de palabras del texto aumenta.

Si bien la ley de Zipf hace referencia a una propiedad estructural de las lenguas, relacionada con la repetición de las palabras (y ha sido usada también para estudiar otros fenómenos, tales como la repetición de las sílabas y de los fonemas), la misma no tiene que ver directamente con fenómenos asociados con la complejidad de los idiomas. La ley cuantitativa más conocida que sí se relaciona con esos fenómenos es la llamada "ley de Menzerath", cuyo nombre se origina en la obra del alemán Paul Menzerath (1954). Esta ley predice que la medida de un elemento lingüístico (y) debería estar negativamente correlacionada con la medida de los componentes de dicho elemento (x). Tal correlación surge de una forma funcional entre ambas variables que resulta similar a la de la ley de Zipf, y puede escribirse del siguiente modo:

$$y = a \cdot x^b \quad ;$$

donde "a" y "b" son parámetros (es decir, números que se supone que permanecen más o menos constantes).

La primera formulación de la ley de Menzerath a través de una ecuación como la expuesta en el párrafo anterior se debe al lingüista eslovaco Gabriel Altmann (1980), quien la utilizó para explicar la relación entre el número de sílabas por palabra y el número de fonemas por sílaba de un determinado texto. A partir de ese momento, la ley de Menzerath (conocida también como "ley de Menzerath-Altmann") fue aplicada en numerosos contextos,[20] pero por la forma en la cual está definida implica siempre una relación entre una medida de la complejidad de un elemento (por ejemplo, de las palabras) y una medida de complejidad de los componentes de ese elemento (por ejemplo, de las sílabas).

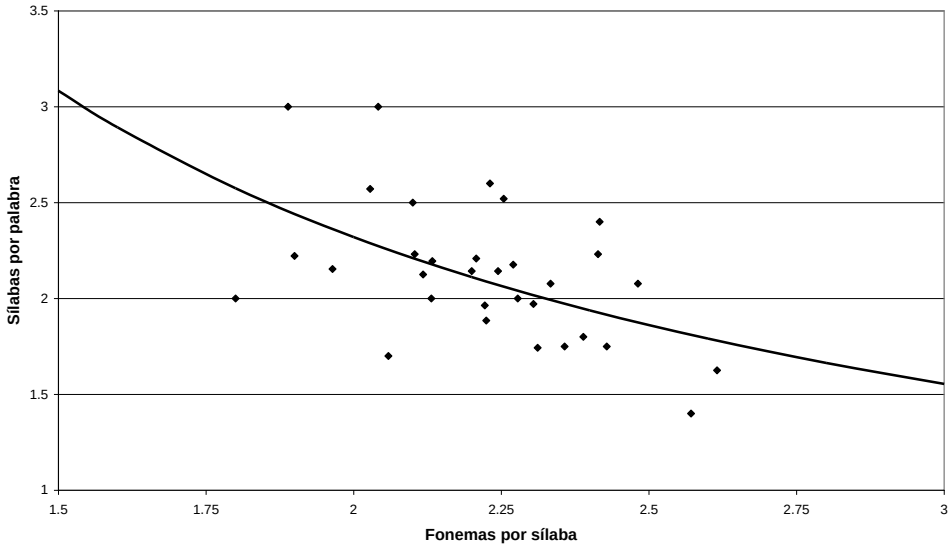
Para chequear si dos medidas de la complejidad de los idiomas tienen una relación como la predicha por la ley de Menzerath resulta necesario llevar a cabo un análisis de regresión estadística. Dicho análisis implica un primer paso en el cual deben recopilarse datos de un texto (o de distintos textos) que impliquen ocurrencias de las medidas que se quieren correlacionar (por ejemplo, sílabas por palabra y fonemas por sílaba). En la oración "el puñal se entibiaba contra su pecho, y debajo latía la libertad agazapada", las 13 palabras que la forman tienen 2,2 sílabas en promedio, y dichas

[19] Para una buena explicación sobre este y otros cambios en la pronunciación del español a lo largo del tiempo, véase Pharies (2006), capítulo 7.

[20] El propio Altmann, por ejemplo, usó esta ley para estudiar fenómenos de complejidad en las obras musicales (Boroda y Altmann, 1991). Más recientemente, también ha sido usada por especialistas en biología para estudiar la complejidad de los genomas (por ejemplo, Ferrer y Forns, 2010).

sílabas tienen a su vez 2,1 fonemas en promedio.

La relación entre estos dos números puede explicarse de distintas maneras, pero una posibilidad es suponer que la misma surge de una función como la escrita más arriba. Pero la oración puesta como ejemplo en el párrafo anterior pertenece al cuento "Continuidad de los parques", de Julio Cortázar, que forma parte del libro *Final del juego* (1956). Si ahora tomamos los 32 enunciados que componen ese cuento y calculamos los cocientes "sílabas/palabras" y "fonemas/sílabas" que corresponden a cada uno de ellos, tendremos dos series de 32 valores cada una que constituyen un conjunto de 32 "observaciones". Cada una de dichas observaciones puede representarse como un punto en el espacio de fonemas por sílaba versus sílabas por palabra, tal cual aparece en el gráfico reproducido a continuación.



¿Qué es entonces lo que predice la ley de Menzerath en este caso? Pues básicamente que la medida del elemento lingüístico "sílabas por palabra" tendrá una relación con la medida de cada uno de sus componentes (fonemas por sílaba) que puede representarse a través de la línea gruesa que hemos dibujado en el gráfico. Dicha línea no es otra cosa que una "tendencia", cuya fórmula es la que surge de suponer una función del tipo "$y = a.x^b$", para la cual se da que "a = 4,6" y "b = -1".

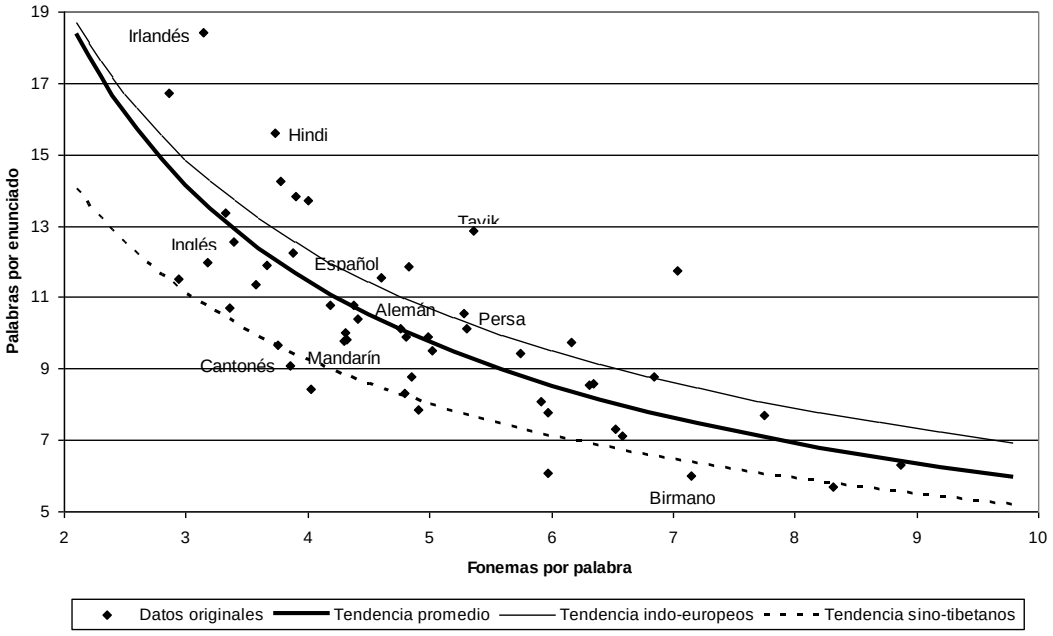
¿Pero de dónde surgen esos números que determinan la posición de esta línea de tendencia en el gráfico? Pues de aplicar el segundo paso del análisis de regresión estadística, que consiste en un procedimiento que se conoce como "estimación por mínimos cuadrados". En este caso, el mismo implica hallar la línea que hace más pequeño el promedio de los cuadrados de las distancias verticales que hay entre cada uno de los puntos del gráfico y el lugar por donde pasa la línea en cuestión.[21]

Hay veces, sin embargo, en las que la línea que uno encuentra aplicando un procedimiento de regresión como el descripto no da una información demasiado buena

[21] El lector que quiera profundizar los detalles de la técnica de regresión descripta, puede consultar alguno de los numerosos libros que existen al respecto en la literatura sobre estadística. Uno que resulta particularmente útil para quienes estén interesados en aplicaciones de la estadística a la lingüística es el de Hernández Campoy y Almeida (2005), así como también el de Stefan Gries (2013).

acerca del comportamiento esperado de una variable ante cambios en otra con la cual se supone que está relacionada. Esto ocurre cuando la distancia entre los puntos y la línea que mejor los ajusta es demasiado grande, cosa que suele pasar si la "nube" que forman los puntos en cuestión tiene una forma que no permite predecir una relación positiva o negativa entre las variables. Las causas por las cuales puede producirse este problema son básicamente dos: o bien las variables que uno quiere relacionar no están verdaderamente conectadas una con la otra, o bien existen también otras variables que influyen en el problema y que no están siendo tenidas en cuenta en el análisis.

Tomemos el caso de otro ejemplo posible de aplicación de la ley de Menzerath, que se refiere a un contexto interlingüístico. Cuando vimos la fábula "El viento norte y el sol", dijimos que la misma era utilizada por la Asociación Fonética Internacional para ejemplificar los sonidos de un gran número de lenguas. En un trabajo anterior nuestro (Coloma, 2015b) usamos 50 versiones distintas de dicha fábula para ver si la ley de Menzerath servía para explicar la relación entre la complejidad de los enunciados (medidos en palabras) y la complejidad de las palabras (medidas en fonemas). Como cada una de las versiones está en un idioma distinto, resultó posible definir no solo las variables "palabras/enunciados" y "fonemas/palabras" para cada uno de los textos estudiados, sino también incorporar variables relacionadas con las familias lingüísticas y las áreas geográficas a las que corresponden los diferentes idiomas.



En el gráfico adjunto aparecen los puntos correspondientes a nuestras 50 versiones de "El viento norte y el sol", y cada punto representa el texto escrito en un determinado idioma. Por una cuestión de disponibilidad de los textos, la muestra está relativamente sesgada a favor de los idiomas eurasiáticos, pero de cualquier modo contiene también diez lenguas amerindias, diez lenguas africanas, dos lenguas austronesias y una lengua australiana. Si uno lleva a cabo un análisis de regresión para tratar de obtener una línea como la predicha por la ley de Menzerath, halla en este caso un resultado según el cual "a = 3,45" y "b = -0,73", que es el representado por la línea

sólida más gruesa del gráfico. La regresión nos da un ajuste relativamente bueno, pero hay algunos idiomas cuyos valores en términos de palabras por enunciado versus fonemas por palabra quedan muy lejos de la línea de tendencia promedio estimada.

Para mejorar el poder explicativo de nuestra regresión, entonces, lo que hicimos fue incorporar ocho variables binarias que adoptan un valor igual a uno cuando un idioma pertenece a determinada área o familia lingüística, y cero cuando no pertenece. Con ese artilugio, lo que se consigue es trabajar con un valor diferente del parámetro "a" para cada subgrupo de observaciones (en este caso, para cada subgrupo de idiomas que pertenece a una determinada región y/o familia lingüística), y estimar distintas líneas de tendencia para dichos subgrupos.

En el gráfico hemos representado dos de las líneas de tendencia estimadas, que corresponden a los trece idiomas indo-europeos de nuestra muestra (alemán, bengalí, español, francés, griego, hindi, inglés, irlandés, nepalí, persa, polaco, portugués y tayik), y a las tres lenguas sino-tibetanas que incluimos (mandarín, cantonés y birmano). Dichas líneas tienen distintos valores del parámetro "a", que son respectivamente iguales a 3,40 y a 3,11, y rigen para un contexto en el cual el valor estimado para el parámetro "b" ya no es más igual a -0,73 sino que es igual a -0,64. Esto último se debe a que, al controlar por factores filogenéticos y geográficos, las líneas de tendencia ya no tienen la misma pendiente que antes, sino que (en este caso) tienen una pendiente menor.

5. Relación entre medidas de complejidad

5.1. Tablas de contingencia y coeficientes de correlación simple

Hemos mencionado en la sección 2.1 que existen lenguas tales como el castellano que utilizan el acento para distinguir entre significados de diferentes palabras que tienen los mismos fonemas en el mismo orden ("revolver" vs. "revólver") y otras como el chino mandarín que usan el tono con un objetivo parecido ("*má*" vs. "*mà*"). Una idea relativamente tradicional de la tipología fonológica era que las lenguas que tienen un sistema tonal no distinguen entre sílabas acentuadas y no acentuadas, y que, en cambio, los idiomas no tonales sí efectúan dicha distinción. Con el tiempo, sin embargo, se han encontrado numerosos casos de sistemas tonales que también usan el acento para diferenciar entre sílabas (generalmente de un modo "predecible").[22] Y hay también casos de idiomas como el nepalí y el tamil, que no tienen un sistema tonal y que tampoco distinguen entre sílabas acentuadas y no acentuadas.

Lo que sí parece ser bien raro es que haya idiomas tonales en los cuales el acento sea distintivo. Por ejemplo, en la muestra de 50 idiomas que usamos en la sección 4.4 para ilustrar la ley de Menzerath, no hay ningún idioma que tenga un sistema de tonos y, al mismo tiempo, use el acento para distinguir entre el significado de las palabras. Tal situación puede apreciarse en el siguiente cuadro.

Acento / Tono	Distintivo	No Distintivo	Total
Distintivo	0	14	14
No Distintivo	17	19	36
Total	17	33	50

[22] Sobre este tema, véase Hyman (2009).

El cuadro que hemos elaborado con las distintas combinaciones posibles de acento y tono es lo que en estadística se conoce como una "tabla de contingencia". Dicha tabla resulta muy útil para calcular la correlación entre dos variables, en especial en un caso como este en el cual cada una de ellas solo puede tomar dos valores: "distintivo" y "no distintivo". Para hacer eso existe una fórmula que utiliza los números que aparecen en cada una de las cuatro celdas principales de la tabla de contingencia (en este caso, 0, 14, 17 y 19) y también los números que aparecen en los "bordes de la tabla" (o sea, 17, 33, 14 y 36). Con todo eso, se llega al siguiente "coeficiente de correlación simple" (r) entre acento y tono:

$$r = \frac{0 \cdot 19 - 17 \cdot 14}{\sqrt{17 \cdot 33 \cdot 14 \cdot 36}} = -0,4476 \text{ ;}$$

cuyo valor indica que la relación entre estas dos variables es negativa.

Para evaluar si la correlación que encontramos es importante o no, hay que tener en cuenta que el coeficiente hallado puede en principio tomar valores que van desde 1 (correlación positiva perfecta) hasta -1 (correlación negativa perfecta). Otra cosa que es importante considerar es que dicho coeficiente no surge de utilizar datos para todos los idiomas que existen en el mundo, sino solamente para una muestra de dichos idiomas. Resulta por lo tanto posible que, por casualidad, hayamos elegido idiomas que tienen una correlación entre acento y tono que sea mayor o menor que la "verdadera correlación" que habríamos encontrado si hubiéramos podido usar los datos de todas las lenguas existentes.

La teoría estadística ha establecido una serie de criterios para determinar cuándo un coeficiente puede ser considerado "significativo", y dichos criterios dependen esencialmente del tamaño de la muestra que uno está usando. En el caso de una muestra de 50 datos, se supone que un coeficiente de correlación cuyo valor absoluto es mayor que 0,282 es significativo, y uno menor que dicho número no lo es. En este caso, "significativo" quiere decir que la probabilidad de que en la realidad el "verdadero coeficiente de correlación" sea nulo (es decir, que no haya ninguna relación entre las variables que estamos relacionando) es menor que 5%.

Como a nosotros el coeficiente nos dio igual a -0,4476, podemos decir con cierta confianza que la correlación entre acento y tono es negativa y significativa (ya que 0,4476 es mayor que 0,282). Más aún, el mismo criterio estadístico por el cual el valor límite para que una correlación sea significativa en una muestra de 50 observaciones es 0,282 nos indica que, si nuestro coeficiente nos dio igual a -0,4476, la probabilidad de que en realidad su valor sea cero es de solamente el 0,11%. Y eso es otra medida de cuán significativo es este valor, puesto que 0,11% no solo es menor que 5%, sino que también es menor que 1%. Podemos decir por ende que la correlación negativa hallada entre acento y tono, igual a -0,4476, es "significativa al 1% de probabilidad".

Los sistemas en los cuales el tono o el acento son elementos que sirven para distinguir entre diferentes palabras pueden verse también como indicativos de que ciertos idiomas son más complejos en algún aspecto de su fonología. La correlación negativa entre acento y tono, además, sería una manifestación de un efecto de compensación entre dos medidas distintas de complejidad (ya que una mayor complejidad en el sistema de asignación del acento se corresponde con una menor complejidad en el sistema de tonos,

y viceversa).

5.2. Coeficientes de correlación parcial

Para explicar el concepto de correlación simple, cosa que hicimos en el apartado anterior, partimos de un ejemplo basado en la fonología comparada, aplicado a la relación entre acento y tono. Algo parecido vamos a hacer aquí para explicar el concepto de correlación parcial. Supongamos, por ejemplo, que contamos con una serie de datos referidos al número de fonemas que tienen ciertos idiomas, al uso del tono como elemento para distinguir entre palabras con significados diferentes, y a la mayor o menor complejidad de las sílabas (en términos de la cantidad máxima de sonidos que tales sílabas pueden tener). Partamos, a manera de ilustración, de la información que aparece en el cuadro que está a continuación, que ha sido extraída del Atlas Mundial de Estructuras Lingüísticas (WALS).

Idioma / Variable	**Fonemas**	**Tonal**	**Sílabas**
Árabe	Muchos	No	Complejas
Birmano	Muchos	Sí	Simples
Griego	Pocos	No	Complejas
Indonesio	Pocos	No	Complejas
Japonés	Pocos	Sí	Simples
Karok	Pocos	Sí	Complejas
Quechua	Muchos	No	Simples
Zulú	Muchos	Sí	Simples

Tal como puede observarse, en dicho cuadro hay datos de ocho idiomas (árabe, birmano, griego, indonesio, japonés, karok, quechua y zulú), que hacen referencia a la cantidad de fonemas (muchos o pocos), a si cada lengua es o no tonal, y a la estructura silábica de la misma (simple o compleja).[23] Supongamos también que, en ese contexto, nos interesa saber si existe alguna relación entre la complejidad fonémica de estos idiomas y su respectiva complejidad tonal.

Si observamos el cuadro con cierto detenimiento, veremos que nuestros ocho idiomas pueden dividirse en cuatro grupos. El primero de ellos (formado por el árabe y el quechua) consta de dos lenguas con muchos fonemas y sin sistema tonal. El segundo (formado por el birmano y el zulú) también tiene dos lenguas, que se caracterizan por tener muchos fonemas y hacer, al mismo tiempo, distinción entre tonos. El tercer grupo (griego e indonesio) está formado por dos idiomas con relativamente pocos fonemas que no tienen tonos distintivos; y el cuarto grupo (japonés y karok) también consta de dos idiomas que tienen pocos fonemas, pero que sí distinguen entre diferentes tonos.

El hecho de que cada uno de los cuatro grupos esté formado por exactamente dos idiomas hace que la correlación simple entre las variables "fonemas" y "tonal" sea nula. Efectivamente, si calculamos el coeficiente de correlación para este ejemplo, usando el método visto en el apartado anterior, el mismo nos dará un número igual a cero. ¿Quiere

[23] Para llevar a cabo esta última clasificación, se consideró que un idioma tiene una estructura silábica compleja si admite sílabas formadas por seis o más fonemas, y una estructura simple en caso contrario.

esto decir que nuestros datos indican que no existe ninguna relación entre fonemas y tonos en esta muestra de idiomas? No necesariamente.

En efecto, la correlación entre complejidad fonémica y tonal para nuestra muestra resulta nula si la analizamos de manera aislada, pero esto se modifica sustancialmente si tenemos en cuenta la posible interacción entre estas dos variables y la tercera característica que nos informa el cuadro: la complejidad silábica. Para incorporar dicho elemento al análisis, lo más fácil es dividir la muestra en dos: una que involucre solamente a los idiomas que tienen sílabas complejas (árabe, griego, indonesio y karok), y otra que incluya a las lenguas cuya estructura silábica es simple (birmano, japonés, quechua y zulú). Tal como se observa en el cuadro que aparece a continuación, la primera submuestra tiene a los dos idiomas no tonales con pocos fonemas (griego e indonesio), y no tiene ninguna lengua que tenga a la vez varios tonos y muchos fonemas. Esas dos lenguas (birmano y zulú) se encuentran ambas en la segunda submuestra, que es la que corresponde a los idiomas con estructura silábica simple.

Concepto	Tonal	No Tonal	Total
Estructura silábica compleja			
Muchos fonemas	0	1	1
Pocos fonemas	1	2	3
Total	1	3	4
Estructura silábica simple			
Muchos fonemas	2	1	3
Pocos fonemas	1	0	1
Total	3	1	4

Si ahora calculamos los correspondientes coeficientes de correlación para cada una de estas dos submuestras, veremos que en ambos casos el coeficiente nos da igual a -0,33. Esto implica que, controlando por el hecho de que los idiomas pueden tener una estructura silábica simple o compleja, la correlación entre fonemas y tonos es negativa en vez de cero. A este tipo de correlación se la conoce en estadística como "correlación parcial", y surge de tener en cuenta el efecto de otras características (en este caso, de la complejidad silábica) sobre las variables que estamos correlacionando.

Lo que el concepto de correlación parcial nos dice en este caso puede ponerse en palabras de la siguiente manera. Que un idioma tenga sílabas complejas se relaciona con que el mismo sea simple tanto en sus fonemas como en sus tonos, y por eso es que los dos ejemplos de lenguas con pocos fonemas y sin tonos entran dentro del grupo de idiomas con sílabas complejas. Del mismo modo, que un idioma tenga sílabas simples se relaciona con que el mismo tenga muchos fonemas y varios tonos, y por eso es que las dos lenguas con esas características están en el grupo de idiomas de estructura silábica simple. Ahora bien, si nos concentramos en las otras dos lenguas de cada submuestra, veremos que en la primera nos aparece un idioma con muchos fonemas y sin tonos (árabe) junto con otro con pocos fonemas y varios tonos (karok). De la misma manera, en la segunda submuestra hay otro idioma con muchos fonemas y sin tonos (quechua) junto a otra lengua tonal con pocos fonemas (japonés). Es justamente esa contraposición la que hace que ahora tengamos cierta correlación negativa entre fonemas y tonos, que surge una vez que depuramos el efecto de la complejidad silábica sobre la complejidad de las

otras dos variables.

La forma mediante la cual calculamos la correlación parcial entre fonemas y tonos en nuestra muestra de ocho idiomas nos resultó efectiva porque las dos submuestras que nos quedan al dividir la muestra total son absolutamente simétricas. En la mayoría de los casos en los cuales uno quiere calcular correlaciones parciales, sin embargo, eso no es así, y resulta necesario utilizar un procedimiento más sofisticado. El mismo consiste en construir primero una matriz con las correlaciones simples entre todas las variables que uno está relacionando (que en este caso serían "fonemas", "tonos" y "sílabas"), y hallar luego la inversa de dicha matriz.[24] Con los números que aparecen en esa matriz inversa, se aplica luego una fórmula para cada coeficiente de correlación parcial que uno quiere calcular, obteniendo de ese modo un resultado para cada par de variables.

Concepto	Fonemas	Tonos	Sílabas
Correlación simple			
Fonemas	1,00	0,00	-0,50
Tonos	0,00	1,00	-0,50
Sílabas	-0,50	-0,50	1,00
Matriz inversa			
Fonemas	1,50	0,50	1,00
Tonos	0,50	1,50	1,00
Sílabas	1,00	1,00	2,00
Correlación parcial			
Fonemas	1,00	-0,33	-0,58
Tonos	-0,33	1,00	-0,58
Sílabas	-0,58	-0,58	1,00

Tal como puede observarse en el cuadro adjunto, los coeficientes de correlación simple forman una matriz simétrica en la cual la diagonal principal está constituida por números iguales a uno. Esto se debe a que, por definición, cada variable tiene una correlación positiva perfecta consigo misma. También se observa que la correlación de cualquier variable "x" respecto de cualquier otra variable "y" (por ejemplo, "fonemas" respecto de "tonos") es por definición igual a la correlación de "y" respecto de "x" (por ejemplo, "tonos" respecto de "fonemas").

En cuanto a los coeficientes de correlación parcial, los mismos surgen de tomar el número que aparece en la matriz inversa, cambiarle el signo, y dividirlo por la raíz cuadrada del producto entre los números que aparecen en la diagonal principal de dicha matriz inversa. Por ejemplo, para calcular la correlación parcial entre fonemas y tonos, lo que hay que hacer es tomar el número 0,50, ponerle signo negativo, y dividirlo por la raíz cuadrada de "1,50*1,50". Esta operación da como resultado -0,33, que es el número que calculamos antes para esta correlación parcial. Del mismo modo, nuestro cuadro nos dice que, si bien la correlación simple entre fonemas y sílabas (y entre tonos y sílabas) es igual

[24] Dos matrices son inversas cuando, si uno las multiplica entre sí, obtiene otra matriz que se conoce como "matriz identidad". Dicha matriz identidad se caracteriza por tener números uno en su diagonal principal, y cero en el resto de la matriz. El procedimiento para obtener una matriz inversa es relativamente complicado de llevar a cabo manualmente, pero es sencillo de realizar utilizando un programa informático de hoja de cálculo (por ejemplo, Microsoft Excel).

a -0,50, la correspondiente correlación parcial es -0,58. Este número surge de hacer "-1,00" dividido por la raíz cuadrada de "1,50*2,00".

Por supuesto, los coeficientes de correlación parcial que uno obtiene cuando hace estos cálculos están basados pura y exclusivamente en los datos que se incluyen en el cómputo de los mismos. Si en vez de tomar los ocho idiomas que elegimos para hacer nuestra explicación tomáramos más idiomas, los resultados nos darían distinto. Más aún, si en vez de incluir solamente tres variables aplicáramos el procedimiento a un conjunto más grande de características, los resultados también serían diferentes.

Por ejemplo, en un trabajo que ya mencionamos en la sección 4.1 (Coloma, 2015a) tomamos en consideración 60 variables (entre características fonológicas, morfológicas, sintácticas, verbales y léxicas) y las aplicamos a los 100 idiomas de la muestra principal del WALS. Con eso pudimos calcular coeficientes de correlación simple para 1770 pares distintos de variables, de los cuales solo 85 resultaron ser a la vez negativos y estadísticamente significativos. Al aplicar el procedimiento de inversión matricial descripto más arriba, sin embargo, el resultado nos cambió considerablemente, y el número de coeficientes de correlación parcial que resultaron ser negativos y significativos se elevó a 162.

Pero el cálculo de los coeficientes de correlación parcial no necesariamente debe limitarse a la inclusión de las variables que uno está interesado en analizar, sino que también puede incorporar variables complementarias. En el trabajo citado en el párrafo anterior, por ejemplo, hicimos un cálculo adicional incluyendo otras doce variables relacionadas con factores filogenéticos (familias lingüísticas), geográficos (áreas lingüísticas) y poblacionales (idiomas con muchos hablantes versus idiomas con pocos hablantes). Como consecuencia de ello, el número de coeficientes de correlación parcial negativos y significativos correspondiente a nuestras 60 variables originales pasó a ser igual a 243.

5.3. Correlación entre variables numéricas

Todos los coeficientes de correlación que hemos calculado hasta ahora se refirieron a casos en los cuales las variables que queríamos relacionar podían tomar solamente dos valores: simple y complejo. Esta no es, por supuesto, la única situación posible, y ni siquiera es la situación más habitual para explicar el concepto estadístico de correlación. Lo más común es que, cuando uno busca correlacionar dos variables, las mismas adopten valores numéricos, y que dichos valores puedan ser numerosos.

Este tipo de variables con muchos valores numéricos posibles nos ha aparecido ya cuando vimos la posibilidad de calcular medidas empíricas de complejidad. Para ilustrar la ley de Menzerath en la sección 4.4, por ejemplo, computamos el cociente entre fonemas y palabras del texto de la fábula de "El viento norte y el sol" traducido a distintos idiomas, y lo comparamos con el cociente entre palabras y enunciados correspondiente al mismo texto. Si, además, contamos el número de sílabas que tienen las palabras en las distintas versiones de "El viento norte y el sol", podemos también calcular los cocientes entre fonemas y sílabas, y entre sílabas y palabras.

En el cuadro que aparece más abajo hemos reproducido, justamente, los valores de esos cocientes para una muestra de ocho idiomas (árabe, birmano, español, igbo, inglés, malayo, quechua y vietnamita). Tal como puede observarse, los mismos toman en general valores que deben representarse a través de números decimales o fraccionarios,

ya que solo por casualidad es que dichos cocientes producen un número entero. De la variación entre los valores correspondientes a unos idiomas y otros puede sacarse también un valor promedio para cada cociente, que aparece informado en la última fila del cuadro.

Idioma / Variable	**Fonemas por sílaba**	**Sílabas por palabra**	**Palabras por enunciado**
Árabe	2,25	2,55	9,44
Birmano	2,29	3,12	6,00
Español	2,20	1,99	10,78
Igbo	1,71	1,94	13,38
Inglés	2,68	1,27	12,56
Malayo	2,30	2,68	9,75
Quechua	2,19	2,88	8,55
Vietnamita	2,85	1,00	16,71
Promedio	**2,31**	**2,18**	**10,90**

¿Cómo se hace en este caso para calcular, por ejemplo, la correlación entre fonemas por sílaba y sílabas por palabra? Pues se usa una fórmula que combina los valores de cada una de estas variables para cada uno de los idiomas, comparando a su vez tales valores con sus respectivos promedios. La fórmula en cuestión es la siguiente:

$$r(f,s) = \frac{\sum\left(f_i - \overline{f}\right)\cdot\left(s_i - \overline{s}\right)}{\sqrt{\sum\left(f_i - \overline{f}\right)^2 \cdot \sum\left(s_i - \overline{s}\right)^2}} = \frac{\sum\left(f_i - 2{,}31\right)\cdot\left(s_i - 2{,}18\right)}{\sqrt{\sum\left(f_i - 2{,}31\right)^2 \cdot \sum\left(s_i - 2{,}18\right)^2}} = -0{,}5182 \quad ;$$

donde "f_i" es "fonemas por sílaba", "s_i" es "sílabas por palabra", y "$\overline{f}$" y "$\overline{s}$" son los valores promedio para dichas variables. Tal como puede observarse, la correlación que se obtiene en este caso es negativa e igual a -0,5182. Esto indica que, en nuestra muestra de ocho idiomas, parece haber una relación inversa entre el número de fonemas por sílaba y el número de sílabas por palabra.

Sin embargo, como vimos en la sección anterior, la correlación simple no siempre nos da una información totalmente confiable respecto de la posible relación que existe entre dos variables. También en este caso es posible calcular coeficientes de correlación parcial, usando el procedimiento de inversión matricial. Para este ejemplo concreto, la correlación parcial nos da resultados bastante similares a la correlación simple si analizamos la relación entre fonemas por sílaba versus sílabas por palabra, y la relación entre sílabas por palabra versus palabras por enunciado. Dichos resultados, sin embargo, se modifican de manera notable si comparamos los coeficientes correspondientes a la relación entre fonemas por sílaba y palabras por enunciado.

En el cuadro adjunto puede verse así que, mientras la correlación entre fonemas por sílaba y sílabas por palabra pasa de ser igual a -0,5182 (correlación simple) a ser igual a -0,5429 (correlación parcial), y la que se verifica entre sílabas por palabra y palabras por enunciado pasa de -0,9276 (correlación simple) a -0,9302 (correlación parcial), el coeficiente de correlación simple entre fonemas por sílaba y palabras por enunciado (que es positivo e igual a 0,3542) se transforma en un coeficiente de

correlación parcial negativo cuyo valor es -0,3959.

Concepto	Fonemas por sílaba	Sílabas por palabra	Palabras por enunciado
Correlación simple			
Fonemas por sílaba	1,0000	-0,5182	0,3542
Sílabas por palabra	-0,5182	1,0000	-0,9276
Palabras por enunciado	0,3542	-0,9276	1,0000
Matriz inversa			
Fonemas por sílaba	1,6212	2,2026	1,4689
Sílabas por palabra	2,2026	10,1539	8,6382
Palabras por enunciado	1,4689	8,6382	8,4922
Correlación parcial			
Fonemas por sílaba	1,0000	-0,5429	-0,3959
Sílabas por palabra	-0,5429	1,0000	-0,9302
Palabras por enunciado	-0,3959	-0,9302	1,0000

La causa por la cual se produce este cambio resulta en este caso relativamente fácil de explicar. Como los fonemas por sílaba están negativamente correlacionados con las sílabas por palabra, y estas últimas están negativamente correlacionadas con las palabras por enunciado, entonces se produce un efecto muy fuerte de transmisión de un fenómeno a otro cuando uno calcula la correlación simple entre fonemas por sílaba y palabras por enunciado. Así, un valor elevado en la cantidad de fonemas por sílaba suele traer aparejado un valor bajo en el la cantidad de sílabas por palabra. Pero justamente ese valor bajo en la cantidad de sílabas por palabra se encuentra correlacionado con un valor relativamente alto en la cantidad de palabras por enunciado.

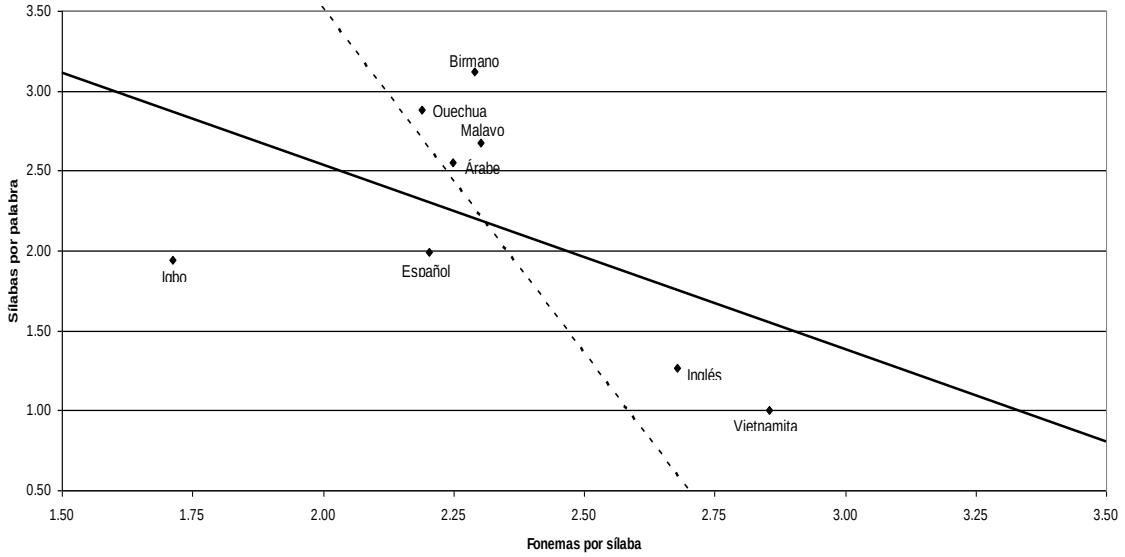
Debido a este fenómeno cruzado, cuando uno compara los valores de la variable "fonemas por sílaba" con los de la variable "palabras por enunciado" se lleva la falsa impresión de que ambas tienen una correlación positiva, pero dicha correlación desaparece totalmente (y, antes bien, se vuelve negativa) si uno logra eliminar el efecto que sobre ella está teniendo de manera concomitante la variable "sílabas por palabra". Eso es justamente lo que hace el procedimiento de cálculo de la correlación parcial: depurar dicho efecto y concentrarse en calcular la relación entre fonemas por sílaba y palabras por enunciado que no se encuentra explicada a través de la interacción con la variable "sílabas por palabra".

En el ejemplo que vimos en la sección anterior, la correlación parcial nos sirvió para ver que algo que permanecía oculto para la correlación simple se volvía patente cuando uno depuraba las variables bajo análisis del efecto de otras variables. Pero la correlación parcial también puede servir para ver si algo que aparece como estadísticamente significativo en el análisis de correlación simple no es en realidad una "correlación espuria". Esto es lo que ocurre en nuestro caso para la correlación entre fonemas por sílaba y palabras por enunciado (que parece ser positiva, pero en realidad no lo es).

5.4. Correlación y regresión

El análisis de correlación que hicimos en la sección anterior, para tratar de ver qué relación existía entre los fonemas por sílaba, las sílabas por palabra y las palabras por enunciado, consistió esencialmente en un cálculo basado en los datos del texto que tomamos como muestra en ocho idiomas distintos. Dichos datos, obviamente, también pueden representarse en gráficos, en los cuales cada idioma es un punto cuya posición depende del valor de los cocientes que estemos estudiando.

Tomemos, por ejemplo, la relación entre fonemas por sílaba y sílabas por palabra. Si nuestro gráfico nos muestra al primero de dichos conceptos en el eje de las abscisas y al segundo en el eje de las ordenadas, lo que queda es algo como lo que aparece más arriba. Tal como puede observarse, en dicho diagrama aparecen los nombres de los idiomas a los que corresponde cada punto, y también aparecen dos líneas rectas con pendiente negativa. Las mismas surgen de hacer regresiones por mínimos cuadrados, similares a las que hicimos en la sección 4.4 cuando ilustramos el concepto de "ley de Menzerath".



El lector quizás se pregunte por qué en el gráfico hay dos líneas rectas en vez de una, y cómo puede ser que ambas sean representaciones de un análisis de regresión llevado a cabo con los mismos datos. La respuesta es que la primera de ellas (la línea gruesa llena) corresponde a la siguiente función:

*Sílabas/Palabra = 4,8485 – 1,1558*Fonemas/Sílaba* ;

en tanto que la segunda (la línea punteada) corresponde a esta otra:

*Fonemas/Sílaba = 2,8157 – 0,2323*Sílabas/Palabra* .

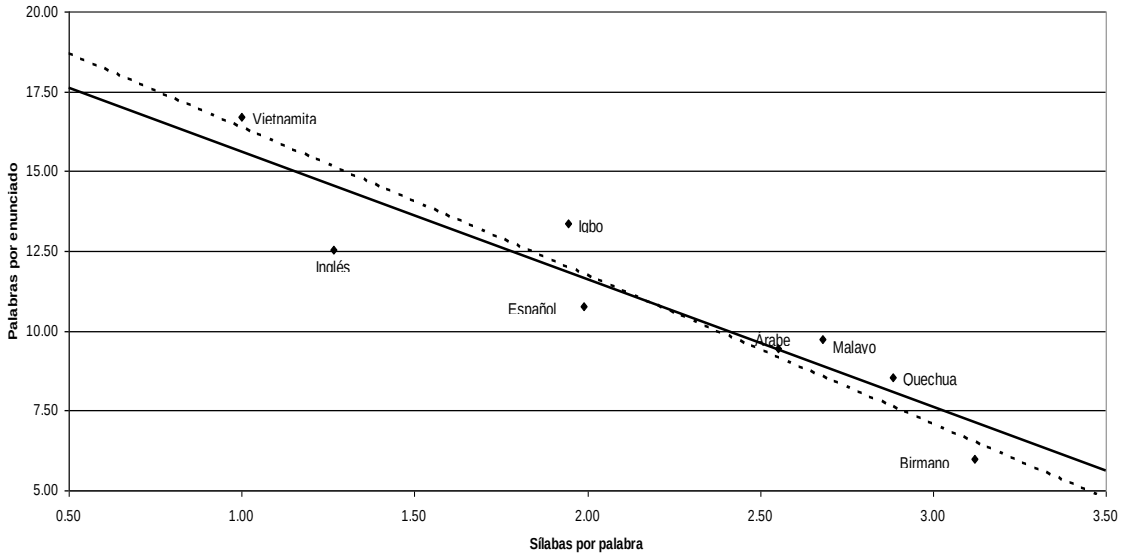
Ambas funciones son el resultado de sendas regresiones por mínimos cuadrados que buscan hallar líneas rectas que ajusten los ocho puntos del gráfico de la mejor manera posible. Pero mientras la primera se preocupa por minimizar las "distancias verticales" entre los puntos y la línea recta, la segunda se preocupa por minimizar las "distancias horizontales". Y, salvo que todos los puntos estén perfectamente alineados, no es lo

mismo una cosa que la otra.

Pero estos resultados, ¿tienen algo que ver con los coeficientes de correlación calculados en la sección anterior para estos ocho idiomas? Claro que sí. Pruebe el lector tomar el número que corresponde a la pendiente de la línea gruesa llena (b_1 = -1,1558), y multiplíquelo por la pendiente de la línea punteada (b_2 = -0,2323). Ahora sáquele la raíz cuadrada a dicha multiplicación, y póngale signo negativo. El resultado que obtendrá no será otra cosa que "r = -0,5182", que es exactamente el coeficiente de correlación simple que hallamos en la sección anterior para las variables "fonemas por sílaba" y "sílabas por palabra".

El lector notará también que las líneas de regresión que dibujamos en nuestro gráfico, si bien tienen ambas pendiente negativa, están bastante separadas una de la otra. Dicha separación tiene que ver con que las variables representadas tienen una correlación relativamente poco significativa, y es por lo tanto posible que aparezcan observaciones (como son, en nuestro caso, las que corresponden a los idiomas inglés y vietnamita) que queden por debajo de una de las líneas pero por encima de la otra.

Veamos, en cambio, qué pasa con la relación entre sílabas por palabra y palabras por enunciado (representada en el gráfico que aparece más abajo). Recordemos que, en nuestra muestra de ocho idiomas, estas variables tienen un coeficiente de correlación igual a -0,9276 (bastante más elevado que el de fonemas por sílaba vs. sílabas por palabra). Esto genera que los puntos de nuestro nuevo gráfico terminen estando mucho más alineados que los del gráfico anterior.



Nótese también que ahora las dos líneas de regresión están mucho más cerca una de la otra. En efecto, en este caso la regresión de palabras por enunciado versus sílabas por palabra (línea gruesa llena) nos queda representada por:

*Palabras/Enunciado = 19,5954 – 3,9924*Sílabas/Palabra* ;

en tanto que la de sílabas por palabra versus palabras por enunciado (línea punteada) nos genera una función igual a:

*Sílabas/Palabra = 4,5272 – 0,2155*Palabras/Enunciado* .

Si ahora calculamos el producto entre las respectivas pendientes y le sacamos la raíz cuadrada negativa, obtendremos un resultado igual a -0,9276, que es exactamente el mismo que obtuvimos en la sección anterior para el coeficiente de correlación entre sílabas por palabra y palabras por enunciado.

La relación entre regresión y correlación no se limita solo al caso de los coeficientes de correlación simple, sino que se extiende también a los coeficientes de correlación parcial. En este caso el punto de contacto aparece cuando uno hace un análisis de regresión múltiple, en el cual se supone que cada variable depende de más de un factor explicativo. En nuestro ejemplo con datos referidos a tres cocientes lingüísticos es posible llevar a cabo tres regresiones diferentes, en las cuales los fonemas por sílaba sean función de las sílabas por palabra y de las palabras por enunciado, las sílabas por palabra sean función de los fonemas por sílaba y de las palabras por enunciado, y estas últimas estén explicadas por los valores de los fonemas por sílaba y de las sílabas por palabra. Si hacemos dichas regresiones aplicando el método de mínimos cuadrados, lo que obtenemos es lo siguiente:

*Fonemas/Sílaba = 4,6651 – 0,6091*Sílabas/Palabra – 0,0944*Palabras/Enunciado* ;

*Sílabas/Palabra = 5,4500 – 0,4838*Fonemas/Sílaba – 0,1977*Palabras/Enunciado* ;

*Palabras/Enunciado = 24,2709 – 1,6605*Fonemas/Sílaba – 4,3782*Sílabas/Palabra* ;

y ahora nos resultará posible multiplicar los respectivos coeficientes de regresión tomados de a pares y sacar las correspondientes raíces cuadradas de los resultados de dichas multiplicaciones.

¿Y qué cree el lector que saldrá de aplicar dichos procedimientos? Pues sí, aparecerán los mismos coeficientes de correlación parcial que calculamos en la sección anterior utilizando el método de inversión de matrices. En efecto, si tomamos el coeficiente correspondiente a sílabas por palabra en la regresión explicativa de fonemas por sílaba (b_1 = -0,6091), lo multiplicamos por el coeficiente correspondiente a fonemas por sílaba en la regresión explicativa de sílabas por palabra (b_1 = -0,4838) y calculamos la raíz cuadrada negativa, el resultado nos da "r = -0,5429", que es exactamente el mismo coeficiente de correlación parcial que obtuvimos en la sección anterior. Y lo mismo pasa con los otros dos pares de coeficientes de regresión, que generan coeficientes de correlación parcial iguales a "r = -0,9302" (para sílabas por palabra versus palabras por enunciado) y a "r = -0,3959" (para fonemas por sílaba versus palabras por enunciado).

Todos estos resultados, por supuesto, son curiosidades matemáticas originadas en la forma en la cual trabajan los distintos algoritmos que se usan para calcular coeficientes de regresión y coeficientes de correlación. Desde el punto de vista conceptual, sin embargo, pueden servirnos también para observar ciertos fenómenos que tienen lugar de manera simultánea, sobre todo si partimos de la idea de que el lenguaje puede ser visto como un sistema en el cual los distintos componentes cumplen funciones interrelacionadas. Con ese enfoque, y con las herramientas estadísticas que hemos visto hasta aquí, nos adentraremos a continuación en el mundo de la lingüística sinergética, en el cual intentaremos bucear para ver si sacamos alguna conclusión respecto de qué tipo de relaciones es posible hallar entre las diferentes medidas teóricas y empíricas de la complejidad de los idiomas.

5.5. Sistemas de ecuaciones y lingüística sinergética

Las tres ecuaciones que usamos en la sección anterior para representar las relaciones entre fonemas por sílaba, sílabas por palabra y palabras por enunciado pueden verse como partes integrantes de un "sistema de ecuaciones". En matemática, se le llama así a un conjunto de ecuaciones que comparten las mismas variables y que puede ser resuelto de manera simultánea, tratando de hallar valores que satisfagan al mismo tiempo las distintas ecuaciones del sistema.

En general, cuando uno usa un sistema de ecuaciones, lo hace porque supone que las relaciones que se establecen entre las distintas variables de dicho sistema tienen un origen común, o porque las mismas sirven para explicar distintas partes de un proceso que confluye en un resultado común. En economía, por ejemplo, uno de los casos más típicos en los que se utilizan sistemas de ecuaciones es el que se refiere al modelo de equilibrio de un mercado, en el cual se supone que tanto el precio como la cantidad de un producto se determinan por la interacción de la oferta y la demanda. El sistema de ecuaciones que sirve para representar eso es, por lo tanto, un sistema en el cual hay una ecuación de oferta (que relaciona la cantidad vendida por los productores con el precio del producto) y una ecuación de demanda (que relaciona dicho precio con la cantidad comprada por los consumidores). El precio de equilibrio del sistema es aquel para el cual la cantidad vendida se iguala con la cantidad comprada, y dicha cantidad no es otra cosa que la cantidad de equilibrio del mercado.

¿Pero cómo podemos hacer para encontrar, en un contexto en el cual las variables son los niveles de complejidad de los idiomas, un sistema semejante al descripto en el párrafo anterior? Una posibilidad es hacer lo que hicimos en la sección 7.3, en la cual tomamos como variables los datos obtenidos empíricamente de un texto traducido a diferentes idiomas. Pero también podemos hacer algo parecido utilizando variables tipológicas, y más aún, podemos tratar de construir el sistema usando un modelo teórico de la relación que se supone que existe entre tales variables.

Esta utilización de un modelo teórico para justificar las variables que incorporemos (y para interpretar las relaciones que obtengamos) nos acerca de algún modo al uso que hacen otras ciencias de los sistemas de ecuaciones. En el caso ya mencionado del sistema de oferta y demanda, por ejemplo, lo que hay detrás es una teoría que sostiene que la función de demanda surge de las decisiones de los consumidores de un producto, y que la función de oferta surge de las decisiones de los productores que venden dicho producto. En el caso de la lingüística, tenemos la posibilidad de basar nuestro sistema de ecuaciones en un modelo sinergético como el que esbozamos en la sección 3.2, en el cual las variables a utilizar son los niveles de complejidad de los subsistemas fonológico, morfológico, sintáctico y léxico.

5.5.1. Un modelo con variables tipológicas

La lógica implícita en el enfoque de la lingüística sinergética supone que los idiomas adoptan distintos niveles de complejidad con el objetivo de satisfacer simultáneamente ciertos requisitos de codificación, economía y estabilidad. La idea detrás de ese razonamiento es que cada uno de dichos niveles tiene una relación con los demás, que puede ser o no significativa. El sistema de ecuaciones que terminaremos teniendo, por lo tanto, será algo que podrá representarse del siguiente modo:

*Fonología = c(1) +c(2)*Morfología +c(3)*Sintaxis +c(4)*Vocabulario* ;

*Morfología = c(5) +c(6)*Fonología +c(7)*Sintaxis +c(8)*Vocabulario* ;

*Sintaxis = c(9) +c(10)*Fonología +c(11)*Morfología +c(12)*Vocabulario* ;

*Vocabulario = c(13) +c(14)*Fonología +c(15)*Morfología +c(16)*Sintaxis* ;

donde los coeficientes *c(1)* a *c(16)* son los valores de los parámetros que relacionan a cada medida de complejidad con las otras.[25]

Supongamos ahora que queremos estimar el valor de dichos parámetros usando un procedimiento de regresión como el que empleamos en la sección 7.3. Para ello es necesario definir primero una muestra de idiomas con ciertos niveles de complejidad para cada uno de sus subsistemas lingüísticos, así como una medida de tales niveles de complejidad. Tomemos, por ejemplo, la muestra de 100 idiomas del WALS, y definamos la complejidad de cada uno de los subsistemas de la siguiente manera:

a) Fonología: Se consideran complejos a los idiomas que tienen más de 25 consonantes, o más de 6 tipos de vocales, o que usan al tono como rasgo distintivo; y simples a los idiomas que no cumplen con ninguno de esos requisitos. Esto hace que el 60% de los idiomas de nuestra muestra sea complejo en esta dimensión, y que el 40% sea simple.

b) Morfología: Se consideran complejos a los idiomas polisintéticos (32% del total), y simples a los idiomas analíticos y sintéticos (que abarcan el 68% restante).

c) Sintaxis: Se consideran complejos a los idiomas que no tienen un orden dominante para el sujeto, el verbo y el objeto, o que usan pronombres relativos para formar proposiciones subordinadas adjetivas (22%), y simples a los demás (78%).

d) Vocabulario: Se consideran complejos a los idiomas que tienen artículos definidos y que distinguen entre los verbos "ser" y "estar" (33%), y simples a los que no cumplen con alguna de esas dos condiciones (67%).

Si ahora construimos variables binarias con estas definiciones, tendremos una matriz de 100 filas (una por idioma) y 4 columnas (una por variable). La fila correspondiente al idioma español, por ejemplo, tendrá un valor simple para las variables "fonología" y "morfología" (ya que el castellano tiene solo 22 consonantes y 5 vocales, no tiene tonos distintivos y no es polisintético), y un valor complejo para las variables "sintaxis" y "vocabulario" (ya que el español usa pronombres relativos para formar proposiciones subordinadas adjetivas, tiene artículos definidos, y distingue entre los verbos "ser" y "estar").

El paso siguiente para poder operar con un sistema como el propuesto es convertir a estas variables conceptuales en variables numéricas. En este caso eso es muy sencillo, ya que implica simplemente asignarle un valor igual a 1 a las observaciones complejas y un valor igual a 0 a las observaciones simples.

Una vez hecho eso, lo siguiente es correr regresiones que utilicen al mismo tiempo los valores de las cuatro variables numéricas definidas. Como las mismas se refieren a ecuaciones que supuestamente están relacionadas dentro del sistema del lenguaje humano, un método útil para aplicar es el que se conoce como "regresión de ecuaciones simultáneas" (*simultaneous-equation regression*). Cuando uno hace eso, no solamente busca los valores de los coeficientes que aproximan mejor las relaciones entre las variables bajo análisis, sino que aprovecha también las correlaciones que aparecen

[25] Este modelo es una reproducción del que aparece en Coloma (2016a).

entre los errores de estimación de las distintas ecuaciones, y las utiliza para mejorar la precisión de los coeficientes estimados. Esto implica usar información que surge de estimar cada ecuación individual en la estimación final de las demás ecuaciones.[26]

Los resultados de correr el sistema en cuestión utilizando dicho método, y tomando como base la muestra de 100 idiomas del WALS, son los siguientes:

*Fonología = 0,8580 –0,3149*Morfología –0,1532*Sintaxis –0,3743*Vocabulario* ;

*Morfología = 0,5499 –0,2928*Fonología +0,0586*Sintaxis –0,2032*Vocabulario* ;

*Sintaxis = 0,3174 –0,1146*Fonología +0,0472*Morfología –0,1324*Vocabulario* ;

*Vocabulario = 0,6414 –0,3496*Fonología –0,2040*Morfología –0,1652*Sintaxis* .

Como algunos de los coeficientes obtenidos toman valores que no son significativamente distintos de cero desde el punto de vista estadístico, este sistema puede reestimarse dejando solamente los coeficientes que sí son significativos. Si hacemos eso los resultados se modifican, y nos quedan expresados del siguiente modo:

*Fonología = 0,8591 –0,3161*Morfología –0,1532*Sintaxis –0,3745*Vocabulario* ;

*Morfología = 0,5686 –0,2989*Fonología –0,2099*Vocabulario* ;

*Sintaxis = 0,3421 –0,1263*Fonología –0,1404*Vocabulario* ;

*Vocabulario = 0,6424 –0,3498*Fonología –0,2053*Morfología –0,1671*Sintaxis* .

Como hemos visto en la sección 5.4, los coeficientes de una regresión como esta pueden ser reinterpretados en términos de coeficientes de correlación parcial entre las variables involucradas. Si hacemos eso, lo que obtenemos es una matriz como la que aparece a continuación.

Correlación parcial	Fonología	Morfología	Sintaxis	Vocabulario
Fonología	1,0000			
Morfología	-0,3074	1,0000		
Sintaxis	-0,1405	0,0000	1,0000	
Vocabulario	-0,3620	-0,2076	-0,1532	1,0000

Tal como puede observarse, la idea detrás de esta matriz (y del sistema de ecuaciones que la genera) es que la complejidad de cada subsistema tiene cierta correlación negativa con la complejidad de por lo menos otros dos subsistemas.

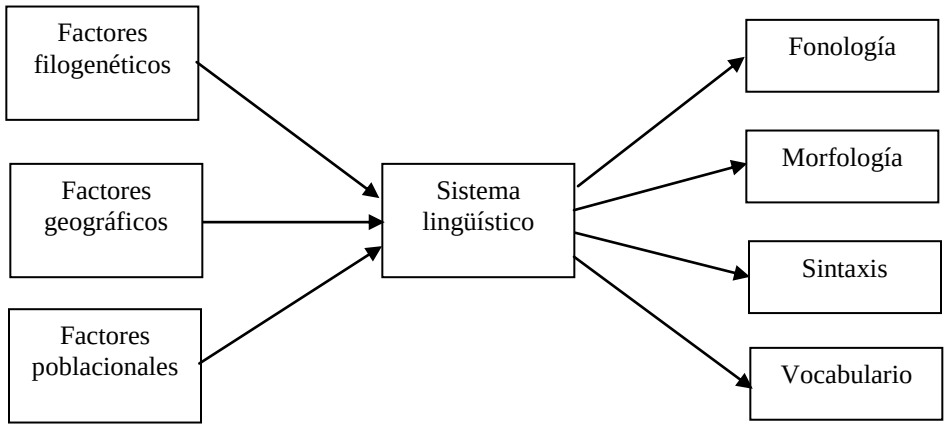
5.5.2. Incorporación de factores extralingüísticos

Tal como puede observarse en el modelo descripto en el apartado anterior, nuestra forma simplificada de representar el funcionamiento del sistema lingüístico toma como variables básicas los niveles de complejidad de los diferentes componentes del lenguaje. Pero si pensamos que los idiomas están influidos por otros elementos adicionales (por

[26] Este procedimiento fue inventado por el estadístico estadounidense Arnold Zellner, quien lo expuso por primera vez en un artículo del año 1962. Desde ese momento ha sido utilizado en numerosas aplicaciones, en especial en ciencias sociales tales como economía, sociología y ciencia política.

ejemplo, factores geográficos, filogenéticos y poblacionales), también podemos incluir algunos de esos elementos dentro de la explicación de las correlaciones encontradas.

Una forma de pensar la interacción entre variables lingüísticas y extralingüísticas en un contexto como el que estamos analizando es suponer que estas últimas forman un entorno en el cual se produce el surgimiento de las primeras. Un idioma como el español, por ejemplo, surgió en un determinado continente (Europa), es parte de una familia (indo-europea) que tiene muchas otras lenguas, y alcanzó un determinado número de hablantes (más de 400 millones) que lo llevaron a expandirse por prácticamente todo el mundo conocido. El idioma mapuche, en cambio, tuvo su origen en un continente totalmente distinto (Sudamérica), quedó como único exponente de su familia lingüística (araucana) y sus hablantes son menos de medio millón de personas que ocupan una zona bastante más reducida (que se limita básicamente a ciertas regiones de Chile y de la Argentina).



Así pensado el asunto, el mismo puede representarse a través de un gráfico como el que aparece más arriba, y eso implica que cada una de las variables lingüísticas (fonología, morfología, sintaxis y vocabulario) ha sido influida por factores extralingüísticos, pero que la inversa no es cierta. Esto quiere decir, por ejemplo, que las características del idioma español no han tenido incidencia en la existencia de Europa como continente, ni han influido en el surgimiento de la familia indo-europea, ni han sido un factor determinante en el crecimiento poblacional del mundo hispanoparlante (que probablemente se habría desarrollado igual si los conquistadores españoles hubieran hablado en vasco en vez de hacerlo en castellano).

Esta relación desde las variables extralingüísticas hacia las lingüísticas, representada por la dirección que adoptan las flechas del gráfico, permite suponer que tales variables extralingüísticas son "explicativas" de las variables lingüísticas. Esto implica que, desde un punto de vista estadístico, es posible construir un sistema en el cual se establezcan relaciones como estas:

*Fonología = c(1) +c(2)*Geografía +c(3)*Filiación +c(4)*Población* ;

*Morfología = c(5) +c(6)*Geografía +c(7)*Filiación +c(8)*Población* ;

*Sintaxis = c(9) +c(10)*Geografía +c(11)*Filiación +c(12)*Población* ;

*Vocabulario = c(13) +c(14)*Geografía +c(15)*Filiación +c(16)*Población* ;

en el que los coeficientes a estimar tengan que ver con la influencia de las variables extralingüísticas sobre cada uno de los niveles de complejidad lingüística que estamos estudiando.

¿Pero qué uso puede tener un sistema como ese y, más aún, qué relación puede tener dicho sistema con el que estimamos anteriormente? Una posible respuesta a esa pregunta es que este nuevo sistema puede servir para solucionar un problema que tiene el sistema anterior, y que hasta aquí hemos ignorado. Dicho problema se conoce en la literatura estadística como "problema de la endogeneidad", y hace referencia a que, si estamos buscando relaciones entre variables que se determinan al mismo tiempo mediante el mismo mecanismo (como serían, en nuestro contexto, los niveles de complejidad de los distintos componentes del lenguaje), entonces los resultados que surgen de regresiones en las cuales algunas de esas variables aparecen como variables independientes pueden estar distorsionados.

Una manera de arreglar esas distorsiones es usar "variables instrumentales". Las mismas sirven para reemplazar a las variables endógenas por otras variables parecidas a ellas, que tienen la propiedad de ser totalmente exógenas al sistema que se quiere estimar. En nuestro caso, por ejemplo, una forma simple de construir variables instrumentales es correr una regresión con un sistema de ecuaciones como el que escribimos unos párrafos más arriba. Lo que surge de ello es lo siguiente:

*Fonología = 0,5919 +0,2445*Africa +0,0747*Norteamérica –0,1318*Papunesia –*
*0,4862*Sudamérica –0,5919*Australia +0,5193*Amazonas +0,2117*Sudestasia*
*+0,0000*Mesoamérica –0,4706*Austronesia –0,2026*Indoeuropea +*
*0,0626*Nigercongo +0,2357*Población ;*

*Morfología = 0,2839 –0,1906*Africa +0,5495*Norteamérica +0,0834*Papunesia +*
*0,1720*Sudamérica +0,2876*Australia +0,2941*Amazonas –0,1674*Sudestasia*
*–0,3333*Mesoamérica –0,2806*Austronesia –0,1441*Indoeuropea –*
*0,0333*Nigercongo –0,1398*Población ;*

*Sintaxis = 0,2471 –0,2413*Africa +0,2529*Norteamérica –0,2281*Papunesia –*
*0,0157*Sudamérica +0,0386*Australia +0,0186*Amazonas –0,1817*Sudestasia*
*–0,5000*Mesoamérica +0,2621*Austronesia +0,5814*Indoeuropea*
*+0,0279*Nigercongo –0,0785*Población ;*

*Vocabulario = 0,3461 +0,3149*Africa +0,2373*Norteamérica +0,0316*Papunesia +*
*0,3958*Sudamérica –0,2032*Australia –0,7418*Amazonas +0,2827*Sudestasia*
*–0,4167*Mesoamérica +0,4699*Austronesia +0,1335*Indoeuropea –*
*0,2233*Nigercongo –0,3546*Población .*

Tal como puede observarse, en estas ecuaciones los factores geográficos aparecen a través de ocho variables binarias (que representan diferentes regiones), los factores de filiación surgen por medio de tres variables binarias correspondientes a las familias lingüísticas con más idiomas dentro de la muestra del WALS (austronesia, indo-europea y Niger-Congo), y los factores poblacionales se incorporan a través de una única variable (que toma un valor igual a uno cuando un idioma tiene más de 5 millones de hablantes, y cero en caso contrario).

Con estos resultados es posible construir cuatro variables instrumentales (que llamaremos *Fonôlogia*, *Morfôlogia*, *Sintâxis* y *Vocabûlario*), formadas por los valores

estimados por nuestras regresiones para las correspondientes variables lingüísticas.[27] Dichas variables instrumentales son combinaciones lineales de las doce variables extralingüísticas, pero tienen la particularidad de ser estimadores de las cuatro variables lingüísticas y de estar basados en datos que son exógenos a ellas. Es entonces posible utilizarlas para estimar el sistema de ecuaciones que relaciona a las variables lingüísticas entre sí, y de dicha estimación surge ahora lo siguiente:

*Fonología = 1,1276 –0,5125*Morfôlogia –0,6073*Sintâxis –0,6969*Vocabûlario* ;

*Morfología = 0,7413 –0,5560*Fonôlogia –0,2657*Vocabûlario* ;

*Sintaxis = 0,5538 –0,4416*Fonôlogia –0,2088*Vocabûlario* ;

*Vocabulario = 0,7729 –0,5569*Fonôlogia –0,1947*Morfôlogia –0,2112*Sintâxis* .

Con estos resultados podemos calcular nuevos coeficientes de correlación parcial, que son los que aparecen en el cuadro que está a continuación. Tal como puede observarse, dichos coeficientes son más elevados que los hallados anteriormente, pero el signo de los mismos (negativo en todos los casos) y el *ranking* entre ellos no sufre ninguna alteración.

Correlación parcial	Fonología	Morfología	Sintaxis	Vocabulario
Fonología	1,0000			
Morfología	-0,5338	1,0000		
Sintaxis	-0,5179	0,0000	1,0000	
Vocabulario	-0,6230	-0,2275	-0,2100	1,0000

Los resultados de las regresiones llevadas a cabo para construir las variables instrumentales, además, tienen cierto interés para detectar relaciones entre medidas de complejidad y factores extralingüísticos. De ellos surge, por ejemplo, que los idiomas con mayor número de hablantes tienden a tener un vocabulario más simple, que los idiomas surgidos en América del Norte tienen una morfología más compleja, y que los idiomas australianos y austronesios tienen en general una fonología más simple.

5.6. Medidas empíricas parciales y complejidad global

En el modelo expuesto en la sección anterior empleamos los conceptos de la lingüística sinergética para explicar las relaciones entre la complejidad de los distintos subsistemas del lenguaje. Sin embargo, por la estructura de dicho modelo, no nos fue posible analizar cómo es que estas relaciones entre diferentes componentes pueden servir para cumplir con los requerimientos que tiene el sistema de la lengua en términos de codificación, economía y estabilidad. En buena medida, esto es así porque las variables que construimos, y que pudimos medir y observar, son medidas que se refieren a la complejidad de cada subsistema (fonología, morfología, sintaxis y vocabulario), pero no

[27] El uso del acento circunflejo o "sombrero" (^) para indicar que se trata de una variable instrumental es una convención habitual en estadística. En este caso, como los nombres de las variables tienen muchas letras, hemos optado por poner dicho símbolo arriba de la vocal que está más cerca del medio de cada palabra.

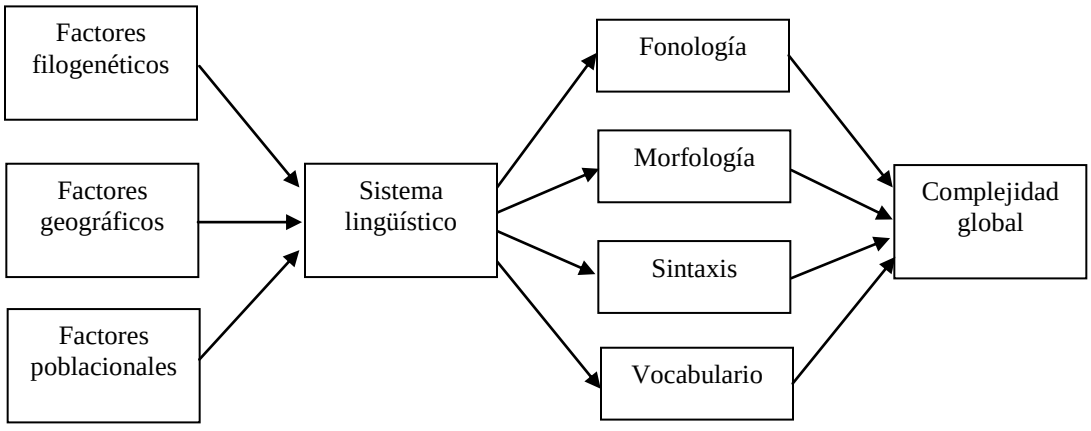
tuvimos manera de introducir ninguna variable adicional que sirviera para medir, de manera independiente, a la complejidad del sistema como un todo.

Dentro de los indicadores de complejidad que vimos en la sección 4, sin embargo, existe uno que por definición tiene un carácter global, y es la denominada "complejidad de Kolmogorov". Tal como mencionamos en dicha sección, la complejidad de Kolmogorov es una medida empírica que se calcula como el cociente entre el tamaño de una versión compactada de cierto texto y el tamaño de la correspondiente versión original.

Para poder utilizar a la complejidad de Kolmogorov como variable en una comparación interlingüística, resulta necesario aplicarla a un conjunto de textos representativos de distintos idiomas. Tal como venimos haciendo desde el apartado 4.4, en esta sección utilizaremos como ejemplo al texto de la fábula "El viento norte y el sol". Dicho texto nos permite calcular la complejidad de Kolmogorov para cada una de sus versiones en diferentes idiomas, y también nos permite relacionar dicha complejidad con algunas medidas empíricas parciales (y con otras variables de carácter geográfico, filogenético y poblacional).

5.6.1. Un modelo con variables empíricas

El modelo que utilizaremos para explicar las relaciones entre la complejidad de Kolmogorov y las distintas medidas empíricas parciales de complejidad está extraído de un artículo que publicamos en 2016 en la revista *Glottometrics*, y es bastante parecido al que vimos en la sección anterior. Gráficamente, el mismo puede representarse utilizando un esquema similar al ya visto, con el agregado de la relación entre las medidas parciales de complejidad y nuestra medida de complejidad global (que en este caso es la complejidad de Kolmogorov).



En el gráfico adjunto puede verse así que los distintos factores filogenéticos, geográficos y poblacionales que influyen sobre el sistema lingüístico hacen que dicho sistema elija determinados niveles de complejidad para sus diferentes componentes (fonología, morfología, sintaxis y vocabulario) y que, a su vez, dichos componentes son los que determinan el nivel de complejidad global del sistema. Las relaciones entre este último nivel y las medidas parciales de complejidad, entonces, se verán reflejadas a

través de ciertos coeficientes estimados empíricamente, en tanto que las relaciones entre complejidad y factores extralingüísticos provendrán del empleo de variables adicionales que se refieran a dichos factores (filogenéticos, geográficos y poblacionales).

La forma en la cual el sistema lingüístico elige los niveles de complejidad parcial puede interpretarse como un proceso en el cual se tienen en cuenta objetivos de economía, codificación y estabilidad. Una manera de modelar eso es pensar que los valores de las variables (que denotaremos con las letras "f", "m", "s" y "v") se eligen con el objetivo de minimizar la complejidad global de todo el sistema (o sea, de satisfacer el requerimiento de economía), y que los niveles deseados de cumplimiento de los requisitos de codificación y estabilidad se toman como dados.

Supongamos, por ejemplo, que los niveles deseados de codificación y estabilidad están determinados por factores geográficos, filogenéticos y poblacionales (que denotaremos con los símbolos "gg", "fg" y "pb"). Esto puede representarse a través de una función "D(gg,fg,pb)" que le asigna cierto nivel a cada idioma, el cual debe a su vez satisfacerse a través de determinados valores de las variables de complejidad parcial. La forma en la cual tales valores consiguen el nivel requerido de codificación y estabilidad tiene lugar a través de una relación entre las variables "f", "m", "s" y "v" y el valor de "D(gg,fg,pb)", que puede representarse por medio de la siguiente igualdad:

$D(gg, fg, pb) = R(f, m, s, v)$.

Ahora bien, como por otro lado sabemos que "f", "m", "s" y "v" sirven también para determinar el nivel de complejidad global (que denotaremos utilizando la letra "k"), nuestro problema de elección de las variables de complejidad parcial puede escribirse sintéticamente del siguiente modo:

$k(min) = C(f, m, s, v)$ sujeto a: $D(gg, fg, pb) = R(f, m, s, v)$;

donde la palabra "min" indica que lo que queremos es minimizar el nivel de "k", y "C(f,m,s,v)" es una función que relaciona dicho nivel de complejidad global con los distintos posibles niveles de las variables de complejidad parcial.

Para resolver un problema como el expuesto, la matemática tiene desarrollado desde hace varios siglos un procedimiento que se conoce con el nombre de "método de Lagrange", en honor al matemático francés Joseph Louis de Lagrange (1736-1813). El mismo consiste en incorporar la restricción "D(gg,fg,pb) = R(f,m,s,v)" dentro de la función que uno quiere minimizar, construyendo una nueva función que se conoce como "lagrangeano" y que, en un caso como el nuestro, adopta la siguiente forma:

$k(min) = L = C(f, m, s, v) + \mu \cdot [D(gg, fg, pb) - R(f, m, s, v)]$;

donde "μ" es una "variable artificial" (también llamada "multiplicador de Lagrange" o "precio sombra"), que sirve para "traducir" las unidades en las cuales está expresada la restricción a unidades semejantes a las de la función que uno quiere minimizar.

Una vez construido el lagrangeano, el procedimiento de minimización consiste en calcular las derivadas de L respecto de cada una de las variables que uno está eligiendo o determinando (que en este caso son "f", "m", "s", "v" y "μ"), e igualar dichas derivadas a cero. A esas igualdades se las conoce con el nombre de "condiciones de primer orden", y de ellas surgen los valores de los distintos niveles de complejidad parcial que resultan

óptimos para minimizar la variable "k".[28]

Nótese que las variables de complejidad parcial aparecen en dos partes distintas del lagrangeano. Por un lado, determinan de manera directa el nivel de complejidad (a través del componente que hemos denotado con la letra "C", que aparece sumando). Por otro, también permiten cumplir con los requerimientos de codificación y estabilidad (a través de la función "R", que aparece restando). Menores niveles de "f", "m", "s" y "v" hacen por lo tanto que el nivel de "C" disminuya, pero también repercuten negativamente sobre el nivel de "R" (y, por ende, sobre el cumplimento de los requisitos de codificación y estabilidad). Por eso la elección de los niveles óptimos para estas medidas de complejidad parcial debe hacerse teniendo en cuenta el efecto conjunto sobre ambas funciones.

7.5.2. Estimación del modelo

Para poder estimar un modelo como el expuesto en el apartado anterior, resulta necesario definir primero cómo vamos a medir las variables de complejidad parcial que necesitamos utilizar. En este caso lo que haremos será usar variables numéricas, basadas en los textos de "El viento norte y el sol" que tenemos en nuestra muestra de 50 idiomas. Las mismas quedarán definidas del siguiente modo:

a) Fonología (*f*): Su complejidad se mide en base a un índice basado en el número de fonemas consonánticos, el número de fonemas vocálicos, el número de tonos, y el carácter distintivo del acento para cada idioma. Dicho índice surge de aplicar la siguiente fórmula:

$f = Consonantes + Vocales*(Tonos+Acento)$;

y nos genera una variable cuyo valor promedio es igual a 42,28, y que tiene un valor máximo igual a 110 (correspondiente al idioma vietnamita) y un valor mínimo igual a 20 (correspondiente al idioma tausug, hablado en las Islas Filipinas).

b) Morfología (*m*): Su complejidad se mide a través del cociente entre fonemas y palabras de cada una de las versiones del texto. Dicho cociente tiene un valor promedio igual a 5,02, siendo su máximo de 9,16 (correspondiente al idioma amazónico peruano yine) y su mínimo de 2,86 (correspondiente al vietnamita).

c) Sintaxis (*s*): Su complejidad se mide a través del cociente entre palabras y enunciados de cada una de las versiones del texto. El valor máximo de ese cociente es igual a 18,43 (para el idioma irlandés), el valor mínimo es igual a 5,70 (para el idioma chickasaw, hablado en la zona centro-oeste de Estados Unidos) y el promedio es igual a 10,31.

d) Vocabulario (*v*): Su complejidad se mide a través del cociente entre tipos y ocurrencias de las palabras en cada una de las versiones del texto. Ese indicador tiene un máximo igual a 0,9048 (para el idioma birmano), un mínimo igual a 0,4459 (para el idioma seri, que se habla en el norte de México) y un promedio igual a 0,6521.

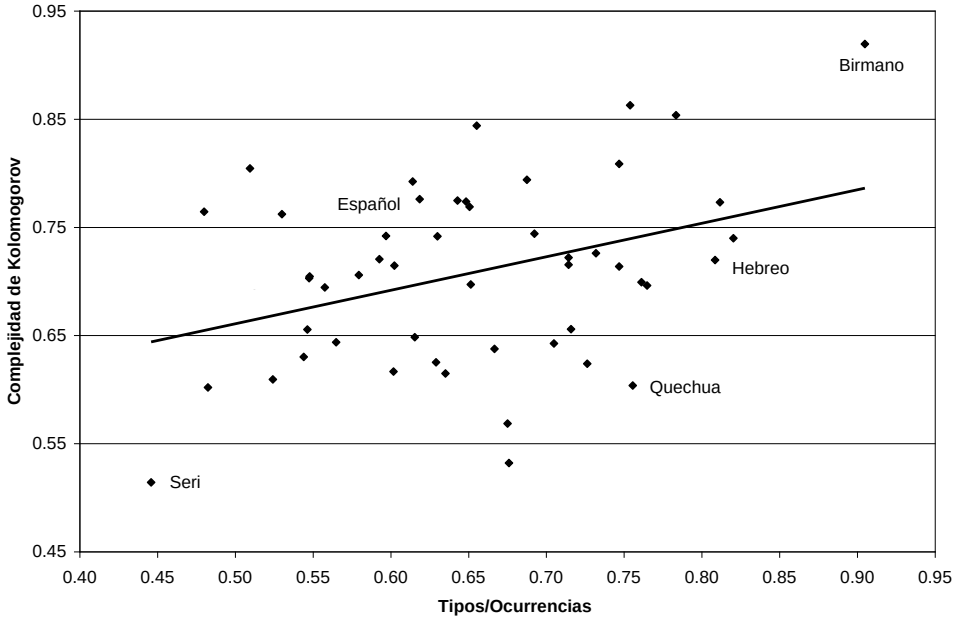
Así definidas, estas variables de complejidad parcial presentan entre sí los siguientes coeficientes de correlación simple:

[28] Esto es un procedimiento matemático estándar, que se usa mucho en disciplinas tales como la física, la ingeniería y la economía. El lector interesado en estos temas puede consultar alguna obra general sobre métodos de optimización, como por ejemplo el libro del matemático indio Rangarajan Sundaram (1996).

Correlación simple	Fonología	Morfología	Sintaxis	Vocabulario
Fonología	1,0000			
Morfología	-0,3720	1,0000		
Sintaxis	0,2136	-0,7265	1,0000	
Vocabulario	-0,0428	0,5581	-0,5046	1,0000

Tal como puede observarse, estas correlaciones son negativas y significativas en tres casos: fonología vs. morfología, morfología vs. sintaxis, y sintaxis vs. vocabulario.[29] No parece en cambio haber una relación negativa entre fonología y sintaxis, ni tampoco entre morfología y vocabulario. La correlación entre fonología y vocabulario, por su parte, sí es negativa pero poco significativa desde el punto de vista estadístico.

En lo que atañe a la complejidad de Kolmogorov, hemos visto que la misma puede calcularse como el cociente entre el tamaño de la versión compactada y el tamaño de la versión original de un archivo correspondiente al texto que se quiere evaluar. Aplicando dicho procedimiento para la muestra de "El viento norte y el sol", obtenemos como resultado que el valor promedio de la complejidad de Kolmogorov es igual a 0,708, en un contexto en el cual la mayor complejidad corresponde al texto escrito en birmano (k = 0,9195) y la menor complejidad aparece para el texto escrito en la lengua seri (k = 0,5142).



El resto de los idiomas se ubica en niveles intermedios entre dichos valores, tal como puede apreciarse en el gráfico adjunto. En él hemos representado a la complejidad

[29] Estos valores son semejantes a los hallados por otros autores que han realizado estudios de lingüística cuantitativa. La correlación entre nuestras medidas de complejidad morfológica y sintáctica, por ejemplo, puede verse como una aplicación particular de la ley de Menzerath (Altmann, 1980). La correlación entre complejidad fonológica y morfológica, por su parte, es semejante a la descubierta originalmente por el lingüista inglés Daniel Nettle (1995), y profundizada luego por otros autores tales como Wichmann, Taraka y Holman (2011) o Moran y Blasi (2014).

de Kolmogorov junto con el cociente entre tipos y ocurrencias (es decir, junto con nuestra medida de complejidad léxica), para cada uno de los 50 idiomas. Tal como puede apreciarse, la relación entre complejidad léxica y complejidad de Kolmogorov es en promedio positiva. Esto parece razonable si pensamos en la complejidad del vocabulario como uno de los componentes de la complejidad global de los idiomas, y en la complejidad de Kolmogorov como una medida empírica de dicha complejidad global.

En nuestra muestra de 50 idiomas aplicada al texto de "El viento norte y el sol", sin embargo, no todas las medidas de complejidad parcial que hemos postulado tienen una correlación positiva con la complejidad de Kolmogorov. De hecho, dicha correlación positiva aparece solo para el vocabulario (r = 0,3639) y para el índice de complejidad fonológica (r = 0,2543), en tanto que las medidas de complejidad morfológica (r = -0,1395) y sintáctica (r = -0,1108) tienen una correlación negativa (no significativa) con la complejidad de Kolmogorov. Esto resulta contraintuitivo, ya que dichas medidas de complejidad parcial también deberían comportarse como elementos constitutivos de un concepto más amplio de complejidad global.

Una posible explicación para esto surge de suponer que los valores de *f*, *m*, *s* y *v* provienen de un modelo sinergético como el que postulamos en el apartado anterior, según el cual la complejidad global (k) queda determinada por una función "C(f,m,s,v)", cuya minimización se encuentra sujeta a la restricción "D(gg,fg,pb) = R(f,m,s,v)". Si uno cuenta con datos referidos a observaciones de diferentes idiomas (como es nuestro caso, aplicado al texto de "El viento norte y el sol"), una forma relativamente simple de aproximar las funciones que hemos postulado es escribir algo como lo siguiente:

$$C(f,m,s,v) = c(1)*f + c(2)*m + c(3)*s + c(4)*v \quad ;$$

$$R(f,m,s,v) = a(1)*\ln(f) + a(2)*\ln(m) + a(3)*\ln(s) + a(4)*\ln(v) \quad ;$$

donde las distintas variables de complejidad parcial aparecen influyendo linealmente sobre la complejidad global y logarítmicamente sobre los requerimientos de codificación y estabilidad.[30]

Los coeficientes que forman parte de estas funciones (*c(1), c(2), c(3), c(4), a(1), a(2), a(3)* y *a(4)*) pueden estimarse a través de un análisis de regresión, que incluya una ecuación de complejidad global y varias ecuaciones derivadas de las condiciones de primer orden de minimización de dicha complejidad. Aplicadas a las funciones que hemos escrito más arriba, dichas condiciones de minimización son las siguientes:

$$\frac{\partial L}{\partial f} = c(1) - \mu \cdot \frac{f}{a(1)} = 0 \; ; \qquad \frac{\partial L}{\partial m} = c(2) - \mu \cdot \frac{m}{a(2)} = 0 \; ;$$

$$\frac{\partial L}{\partial s} = c(3) - \mu \cdot \frac{s}{a(3)} = 0 \; ; \qquad \frac{\partial L}{\partial v} = c(4) - \mu \cdot \frac{v}{a(4)} = 0 \; ;$$

y, operando en ellas, resulta posible llegar a una serie de relaciones entre las variables de complejidad parcial. En nuestro caso, eso implica:

[30] Esto es una manera de suponer que las medidas de complejidad parcial tienen un efecto positivo sobre la expresividad de los idiomas, pero que dicho efecto es "menos que proporcional". También implica que los distintos sub-sistemas del lenguaje son en cierta medida "sustitutos entre sí" (en el sentido de que una menor complejidad en uno de ellos puede ser reemplazada por una mayor complejidad en otro).

$$f \cdot 3 = \frac{a(1)}{a(2)} \cdot \frac{c(2)}{c(1)} \cdot m + \frac{a(1)}{a(3)} \cdot \frac{c(3)}{c(1)} \cdot s + \frac{a(1)}{a(4)} \cdot \frac{c(4)}{c(1)} \cdot v \quad ;$$

$$m \cdot 3 = \frac{a(2)}{a(1)} \cdot \frac{c(1)}{c(2)} \cdot f + \frac{a(2)}{a(3)} \cdot \frac{c(3)}{c(2)} \cdot s + \frac{a(2)}{a(4)} \cdot \frac{c(4)}{c(2)} \cdot v \quad ;$$

$$s \cdot 3 = \frac{a(3)}{a(1)} \cdot \frac{c(1)}{c(3)} \cdot f + \frac{a(3)}{a(2)} \cdot \frac{c(2)}{c(3)} \cdot m + \frac{a(3)}{a(4)} \cdot \frac{c(4)}{c(3)} \cdot v \quad ;$$

$$v \cdot 3 = \frac{a(4)}{a(1)} \cdot \frac{c(1)}{c(4)} \cdot f + \frac{a(4)}{a(2)} \cdot \frac{c(2)}{c(4)} \cdot m + \frac{a(4)}{a(3)} \cdot \frac{c(3)}{c(4)} \cdot s \quad .$$

El siguiente paso consiste en escribir estas relaciones como parte de un sistema de ecuaciones que sirva para estimar los valores de los coeficientes. Esto puede hacerse del siguiente modo:

$$k = c(1)*f + c(2)*m + c(3)*s + c(4)*v \quad ;$$

$$f*3 = c(5)*m + c(6)*s + c(7)*v \quad ;$$

$$m*3 = [1/c(5)]*f + [c(6)/c(5)]*s + [c(7)/c(5)]*v \quad ;$$

$$s*3 = [1/c(6)]*f + [c(5)/c(6)]*m + [c(7)/c(6)]*v \quad ;$$

$$v*3 = [1/c(7)]*f + [c(5)/c(7)]*m + [c(6)/c(7)]*s \quad ;$$

donde se da que "*c(5) = [a(1)·c(2)]/[a(2)·c(1)]*", "*c(6) = [a(1)·c(3)]/[a(3)·c(1)]*" y "*c(7) = [a(1)·c(4)]/[a(4)·c(1)]*".

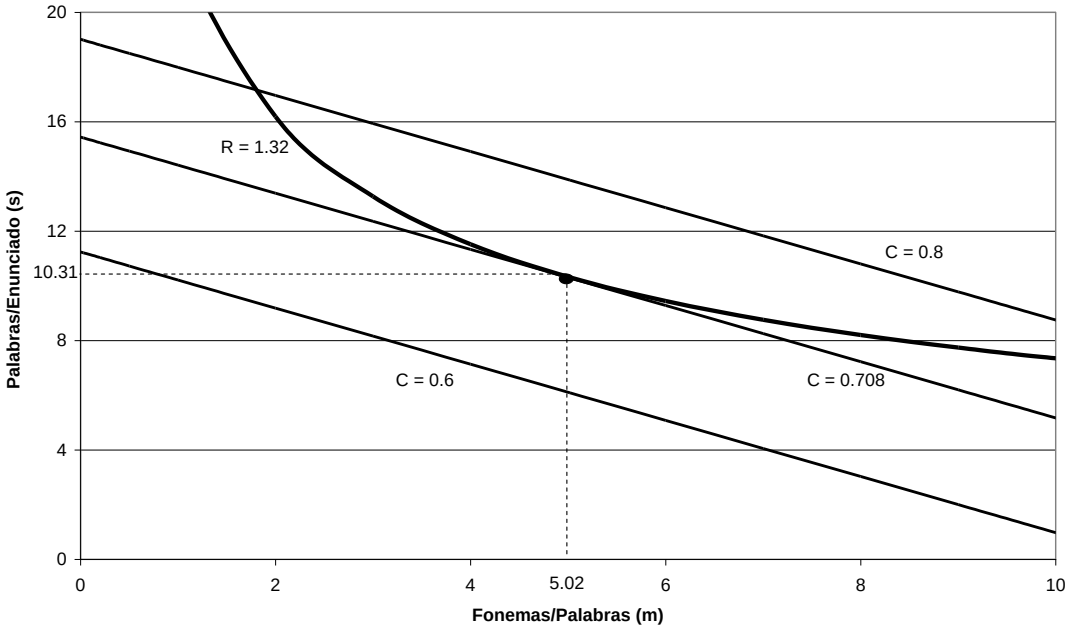
Si ahora estimamos los coeficientes *c(1)* a *c(7)* utilizando un procedimiento de regresión de ecuaciones simultáneas, podemos reemplazar a *f*, *m*, *s* y *v* por variables instrumentales ($\hat{f}$, $\hat{m}$, $\hat{s}$, $\hat{v}$) que dependan de factores geográficos, filogenéticos y poblacionales. Con ello obtuvimos los siguientes valores: "*c(1) = 0,0013*", "*c(2) = 0,026399*", "*c(3) = 0,025719*", "*c(4) = 0,392505*", "*c(5) = 9,087457*", "*c(6) = 4,342966*" y "*c(7) = 68,87271*", y con los últimos tres números de esta lista pudimos a su vez calcular los valores de *a(1)*, *a(2)*, *a(3)* y *a(4)*. Suponiendo que "*a(1) +a(2) +a(3) +a(4) = 1*", dichos valores son: "*a(1) = 0,0821*", "*a(2) = 0,1836*", "*a(3) = 0,3742*" y "*a(4) = 0,3601*".[31]

Las relaciones a las que hemos llegado pueden representarse a través de un gráfico como el que aparece más abajo, en el cual hemos dibujado una curva que corresponde a un valor específico de la función *R* (igual a 1,32, que es lo que se verifica cuando *f*, *m*, *s* y *v* adoptan los valores promedio para nuestra muestra de idiomas) y tres casos particulares de la función *C*. Estos últimos corresponden al valor promedio de la complejidad de Kolmogorov (*C = 0,708*) y a otras dos situaciones alternativas ("*C = 0,6*" y "*C = 0,8*").

Tal como puede observarse, el diagrama ha sido dibujado en el espacio de

[31] Este supuesto es puramente convencional, ya que nuestra medida de la restricción de codificación y estabilidad del lenguaje no está atada a ninguna magnitud directamente observable. Eso hace que sea posible suponer cualquier conjunto de valores para *a(1)*, *a(2)*, *a(3)* y *a(4)*, siempre que los mismos mantengan entre sí las relaciones dadas por los valores de los coeficientes *c(5)*, *c(6)* y *c(7)*.

fonemas por palabra versus palabras por enunciado (es decir, de las variables "*m*" y "*s*"), y en él se ve cómo la curva correspondiente a "*R = 1,32*" es tangente con la recta que corresponde al valor promedio de *k*. Eso implica que dicho valor de *k* es el mínimo que se puede conseguir si uno le exige al lenguaje que cumpla con el requisito de codificación y estabilidad establecido por los valores de *f*, *m*, *s* y *v* que se observan en nuestra muestra de idiomas. Nótese, por ejemplo, que los valores de *m* y *s* que se determinan en el punto de tangencia ("*m = 5,02*" y "*s = 10,31*") no son otra cosa que los promedios que tales variables adoptan en el conjunto de las observaciones de la muestra de 50 idiomas correspondientes al texto de "El viento norte y el sol".



El modelo de optimización que hemos desarrollado también puede servirnos para calcular coeficientes de correlación parcial entre *f*, *m*, *s* y *v*. Para ello resulta necesario hacerle una pequeña modificación, y re-estimar un sistema de ecuaciones que incluya a las relaciones entre las variables de complejidad parcial pero no a la relación entre estas y la complejidad global. Dicho sistema es el siguiente:

$$f*3 = c(5)*\hat{m} + c(6)*\hat{s} + c(7)*\hat{v} \quad ;$$

$$m*3 = c(8)*\hat{f} + c(6)*c(8)*\hat{s} + c(7)*c(8)*\hat{v} \quad ;$$

$$s*3 = c(9)*\hat{f} + c(5)*c(9)*\hat{m} + c(7)*c(9)*\hat{v} \quad ;$$

$$v*3 = c(10)*\hat{f} + c(5)*c(10)*\hat{m} + c(6)*c(10)*\hat{s} \quad ;$$

y en él no necesariamente se da que "*c(8) = 1/c(5)*", "*c(9) = 1/c(6)*" ni "*c(10) = 1/c(7)*".

Justamente por el hecho de que en esta nueva estimación no estamos imponiendo las restricciones de igualdad mencionadas en el párrafo anterior es que resulta posible estimar coeficientes de correlación parcial. Los mismos surgen de efectuar cálculos usando la siguiente fórmula:

$$r_{xy} = -\sqrt{\frac{c_{xy} \cdot c_{yx}}{9}} \quad ;$$

en la cual r_{xy} es la correlación parcial entre dos variables *x* e *y* cualesquiera, y c_{xy} y c_{yx} son los coeficientes de regresión correspondientes a cada variable en la ecuación explicativa de la otra variable.

Si ahora aplicamos esta fórmula a los seis pares posibles de medidas de complejidad parcial que estamos utilizando, los coeficientes de correlación parcial pasan a ser los que aparecen en el siguiente cuadro:

Correlación parcial	Fonología	Morfología	Sintaxis	Vocabulario
Fonología	1,0000			
Morfología	-0,2322	1,0000		
Sintaxis	-0,3720	-0,2592	1,0000	
Vocabulario	-0,3947	-0,2750	-0,4406	1,0000

Tal como puede observarse, estas correlaciones son negativas para todas las combinaciones de variables de complejidad parcial incluidas en nuestro modelo. La más grande en valor absoluto es la que relaciona a las medidas de complejidad sintáctica y léxica, seguida por la que relaciona a esta última con la complejidad fonológica. Nótese además que los coeficientes entre fonología y sintaxis, y entre morfología y vocabulario (que eran positivos cuando los calculábamos utilizando un procedimiento de correlación simple) se han vuelto negativos y significativos. Esto concuerda con la idea de que todas estas variables representan aspectos parciales de la complejidad global de los idiomas y tienen por lo tanto cierta capacidad de sustituirse entre sí, a fin de cumplir con los requisitos de codificación y estabilidad implícitos en el uso del lenguaje.

Referencias bibliográficas

Adsett, Connie y Yannick Marchand (2010). "Syllabic Complexity: A Computational Evaluation of Nine European Languages"; *Journal of Quantitative Linguistics*, vol 17, pp 269-290.

Altmann, Gabriel (1980). "Prolegomena to Menzerath's Law"; *Glottometrika*, vol 2, pp 1-10.

Bane, Max (2008). "Quantifying and Measuring Morphological Complexity"; *Proccedings of the 26th West Coast Conference on Formal Linguistics*, pp 69-76. Somerville, Cascadilla.

Biberauer, Theresa, Anders Holmberg, Ian Roberts y Michelle Sheenan (2014). "Complexity in Comparative Syntax: The View from Modern Parametric Theory"; en F. Newmayer y L. Preston (eds.): *Measuring Grammatical Complexity*, pp 217-240. Nueva York, Oxford University Press.

Bickel, Balthasar y Johanna Nichols (2005). "Inflectional Synthesis of the Verb"; en M. Haspelmath, M. Dryer, D. Gil y B. Comrie (eds.): *The World Atlas of Language Structures*, 1ª edición, pp 94-97. Oxford, Oxford University Press.

Boroda, Moisei y Gabriel Altmann (1991). "Menzerath's Law in Musical Texts"; *Musikometrica*, vol 3, pp 1-13.

Canepari, Luciano (2005). *A Handbook of Pronunciation*. Munich, Lincom Europa.
Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, MIT Press (hay versión en español: *Aspectos de la teoría de la sintaxis*; Barcelona, Gedisa).
Coloma, Germán (2015a). "Efectos de compensación entre indicadores de la complejidad de los idiomas"; Documento de Trabajo 569. Buenos Aires, Universidad del CEMA.
Coloma, Germán (2015b). "The Menzerath-Altmann Law in a Cross-Linguistic Context"; *SKY Journal of Linguistics*, vol 28, pp 139-159.
Coloma, Germán (2016a). "A Simultaneous-Equation Regression Model of Language Complexity Trade-Offs"; Documento de Trabajo 597. Buenos Aires, Universidad del CEMA.
Coloma, Germán (2016b). "An Optimization Model of Global Language Complexity"; *Glottometrics*, vol 35, pp 49-63.
Coloma, Germán (2017). *La complejidad de los idiomas*. Oxford, Peter Lang.
Comrie, Bernard (1989). *Language Universals and Linguistic Typology*, 2ª edición. Oxford, Blackwell (hay versión en español: *Universales del lenguaje y tipología lingüística*; Madrid, Gredos).
Cortázar, Julio (1956). *Final del juego*. Buenos Aires, Sudamericana.
Culicover, Peter (2013). *Grammar and Complexity*. Oxford, Oxford University Press.
Dryer, Matthew y Martin Haspelmath (2013). *The World Atlas of Language Structures Online*. Leipzig, Instituto Max Planck.
Ehret, Katharina y Benedikt Szmrecsányi (2015). "An Information-Theoretic Approach to Assess Linguistic Complexity"; en R. Baechler y G. Seiler (eds.): *Complexity, Variation and Isolation*. Berlín, De Gruyter.
Eifring, Halvor y Rolf Theil (2005). *Linguistics for Students of Asian and African Languages*. Oslo, Universidad de Oslo.
Fenk-Oczlon, Gertraud y August Fenk (1999). "Cognition, Quantitative Linguistics and Systemic Typology"; *Linguistic Typology*, vol 3, pp 151-177.
Fenk-Oczlon, Gertraud y August Fenk (2011). "Complexity Trade-Offs in Language Do Not Imply an Equal Overall Complexity"; en V. Solovyev y V. Polyakov (eds.): *Text Processing and Cognitive Technologies*, pp 145-148. Kazan, Kazan State University Press.
Ferrer, Ramón y Nuria Forns (2010). "The Self-Organization of Genomes"; *Complexity*, vol 15, pp 34-36.
Garrido, Joaquín (2009). *Manual de lengua española*. Madrid, Castalia.
Greenberg, Joseph (1966). *Language Universals*. La Haya, Mouton.
Gries, Stefan (2013). *Statistics for Linguistics with R*, 2ª edición. Berlín, De Gruyter.
Grimm, Jacob (1822). *Deutsche Grammatik*, 2ª edición. Gotinga, Dieterich.
Hawkins, John (2004). *Efficiency and Complexity in Grammars*. Nueva York, Oxford University Press.
Hernández Campoy, Juan y Manuel Almeida (2005). *Metodología de la investigación sociolingüística*. Málaga, Comares.
Hyman, Larry (2009). "How (Not) to Do Phonological Typology: The Case of Pitch-Accent"; *Language Sciences*, vol 31, pp 213-238.
Israel, Michael (2014). "Semantics: How Language Makes Sense"; en C. Genetti (ed.): *How Languages Work*, pp 150-179. Nueva York, Cambridge University Press.

Jackendoff, Ray (2002). *Foundations of Language*. Nueva York, Oxford University Press (hay versión en español: *Fundamentos del lenguaje*; México, Fondo de Cultura Económica).

Joseph, John y Frederick Newmeyer (2012). "All Languages Are Equally Complex: The Rise and Fall of a Consensus"; *Historiographia Linguistica*, vol 39, pp 341-368.

Kettunen, Kimmo (2014). "Can Type-Token Ratios Be Used to Show Morphological Complexity of Languages?"; *Journal of Quantitative Linguistics*, vol 21, pp 223-245.

Kirby, Simon, Monica Tamariz, Hannah Cornish y Kenny Smith (2015). "Compression and Communication in the Cultural Evolution of Linguistic Structure"; *Cognition*, vol 141, pp 87-102.

Köhler, Reinhard (1987). "System Theoretical Linguistics"; *Theoretical Linguistics*, vol 14, pp 241-257.

Köhler, Reinhard (2005). "Synergetic Linguistics"; en G. Altmann, R. Köhler y R. Piotrowski (eds.): *Quantitative Linguistics*, pp 760-774. Berlín, De Gruyter.

Kolmogorov, Andrei (1963). "On Tables of Random Numbers"; *Sankhya*, vol 25, pp 369-376.

Kusters, Wouter (2003). *Linguistic Complexity*. Utrecht, LOT.

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 (eds.): *Experimental Approaches to Phonology*, pp 93-103. Nueva York, Oxford University Press.

Mairal, Ricardo y Juana Gil (2006). "A First Look at Universals"; en R. Mairal y J. Gil (eds.): *Linguistic Universals*, pp 1-45. Cambridge, Cambridge University Press.

Martínez Celdrán, Eugenio, Ana Fernández Planas y Josefina Carrera (2003). "Illustrations of the IPA: Castilian Spanish"; *Journal of the International Phonetic Association*, vol 33, pp 255-260.

McWhorter, John (2001). "The World's Simplest Grammars Are Creole Grammars"; *Linguistic Typology*, vol 5, pp 125-166.

McWhorter, John (2003). *The Power of Babel*. Nueva York, Harper.

Mendívil, José Luis (2009). *Origen, evolución y diversidad de las lenguas*. Frankfurt, Peter Lang.

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

Miestamo, Matti (2008). "Grammatical Complexity in a Cross-Linguistic Perspective"; en M. Miestamo, K. Sinnemäki y F. Karlsson (eds.): *Language Complexity: Typology, Contact and Change*, pp 23-41. Amsterdam, John Benjamins.

Moran, Steven y Damián Blasi (2014). "Cross-Linguistic Comparison of Complexity Measures in Phonological Systems"; en F. Newmayer y L. Preston (eds.): *Measuring Grammatical Complexity*, pp 217-240. Nueva York, Oxford University Press.

Moscoso, Fermín (2011). "The Universal 'Shape' of Human Language: Spectral Analysis Beyond Speech"; *Nature Precedings*, 6097-2.

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

Nichols, Johanna (2009). "Linguistic Complexity: A Comprehensive Definition and Survey"; en G. Sampson, D. Gil y P. Trudgill (eds.): *Language Complexity as an Evolving Variable*, pp 110-125. Oxford, Oxford University Press.

Parkvall, Mikael (2008). "The Simplicity of Creoles in a Cross-Linguistic Perspective"; en M. Miestamo, K. Sinnemäki y F. Karlsson (eds.): *Language Complexity: Typology, Contact and Change*, pp 265-285. Amsterdam, John Benjamins.

Pfau, Roland, Markus Steinbach y Bencie Woll (2012). *Sign Language: An International Handbook*. Berlín, De Gruyter.

Pharies, David (2006). *Breve historia de la lengua española*. Chicago, University of Chicago Press.

Real Academia Española (2011). *Nueva gramática de la lengua española: fonética y fonología*. Madrid, Espasa.

Sampson, Geoffrey (2009). "A Linguistic Axiom Challenged"; en G. Sampson, D. Gil y P. Trudgill (eds.): *Language Complexity as an Evolving Variable*, pp 1-18. Oxford, Oxford University Press.

Shosted, Ryan (2006). "Correlating Complexity: A Typological Approach"; *Linguistic Typology*, vol 10, pp 1-40.

Sinnemäki, Kaius (2008). "Complexity Trade-Offs in Core Argument Marking"; en M. Miestamo, K. Sinnemäki y F. Karlsson (eds.): *Language Complexity: Typology, Contact and Change*, pp 67-88. Amsterdam, John Benjamins.

Sundaram, Rangarajan (1996). *A First Course in Optimization Theory*. Nueva York, Cambridge University Press.

Szmrecsányi, Benedikt (2004). "On Operationalizing Syntactic Complexity"; en *Anales de la 7ª Jornada Internacional de Análisis de Datos Textuales* (JADT 2004), pp 1031-1038.

Trudgill, Peter (2009). "Sociolinguistic Typology and Complexification"; en G. Sampson, D. Gil y P. Trudgill (eds.): *Language Complexity as an Evolving Variable*, pp 98-109. Oxford, Oxford University Press.

Wichmann, Soren, Rama Taraka y Eric Holman (2011). "Phonological Diversity, Word Length and Population Sizes across Languages: The ASJP Evidence"; *Linguistic Typology*, vol 15, pp 177-197.

Wimmer, Gezja y Gabriel Altmann (2007). "Towards a Unified Derivation of Some Linguistic Laws"; en P. Grzybek (ed.): *Contributions to the Science of Text and Language*, pp 329-338. Dordrecht, Springer.

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

Zipf, George (1935). *The Psycho-Biology of Language*. Boston, Houghton Mifflin.