REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2023 PUCV, Chile 56(113) 545-571

Análisis de Lenguas con Valores Extremos en los Índices de Relatividad, Densidad y Eficiencia Informativa: La Tipología Morfológica y Genética y la Complejidad del Sistema Fonético-Fonológico en el Estudio del Número y Longitud de Palabras y Fonemas*

Analysis of Languages with Extreme Values in the Indices of Relativity, Density and Informative Efficiency: The Morphological and Genetic Typology and the Complexity of the Phonetic-Phonological System in the Study of the Number and Length of Words and Phonemes

Enrique J. Vercher García
UNIVERSIDAD COMPLUTENSE DE MADRID
ESPAÑA
evercher@ucm.es

Manuel Bullejos Lorenzo
UNIVERSIDAD DE GRANADA
ESPAÑA
bullejos@ugr.es

Recibido: 19-III-2021 / **Aceptado:** 16-XI-2022
DOI: 10.4067/S0718-09342023000300545

Resumen

El presente artículo analiza las correlaciones matemáticas entre las lenguas que presentan valores extremos en los llamados 'índice de relatividad informativa', 'índice de densidad informativa', 'índice de eficiencia informativa léxica' e 'índice de eficiencia informativa fónica'. Dichos índices expresan los coeficientes resultantes de dividir el número de 'tokens' y el número de 'unidades fónicas convencionales de token', empleados para expresar una misma información. En el presente trabajo nos centramos muy especialmente en aquellas lenguas que muestran valores extremos en dichos indíces y analizamos en qué modo afectan la tipología morfológica o las características fonético-fonológicas de las lenguas a cuestiones como número total de palabras y fonemas, longitud de palabras o economía del lenguaje.

Palabras Clave: Índice de relatividad informativa, índice de densidad informativa, índice de eficiencia informativa léxica, índice de eficiencia informativa fónica, tokens.

Abstract

This article analyses the mathematical correlations between languages which present extreme values in the so-called 'index of informative relativity', 'index of informative density', 'lexical informative efficiency index', and 'phonic informative efficiency index'. These indices express the coefficients resulting from dividing the number of 'tokens' and number of 'token conventional phonic units', used to express the same information. In the present work we focus very especially on those languages that show extreme values in aforesaid indices and we analyze how the morphological typology or the phonetic-phonological characteristics of the languages affect issues such as the total number of words and phonemes, length of words or economy of language.

Keywords: Index of informative relativity, index of informative density, lexical informative efficiency index, phonic informative efficiency index, tokens.

INTRODUCCIÓN

El presente artículo analiza la correlación entre el número de palabras (tokens) que usa una lengua para expresar un contenido semántico dado (información) y el número de sonidos (unidades fónicas convencionales de token) totales empleados en expresar ese mismo contenido semántico. Especialmente, nos centramos en mostrar las lenguas que presentan valores extremos (máximos y mínimos) y en analizar y extraer conclusiones a partir de factores como el carácter tonal o no tonal de la lengua, la complejidad del sistema fonológico o la familia lingüística. Estos datos aportarán a la bibliografía ya existente sobre el tema un mayor conocimiento de principios matemáticos sobre la estructuración y funcionamiento de las lenguas en lo que a uso, cuantitativamente hablando, de unidades léxicas y fonéticas se refiere.

Actualmente, en la lingüística, especialmente en estudios relacionados con la carga de información, se suele hablar de 'token'[1], concepto que vamos a emplear nosotros y que en nuestro trabajo vendrá a equivaler a la 'palabra' según la división convencional que cada lengua hace de sus propias palabras[2]. Por otro lado, a la hora de analizar el número de componentes de un token en nuestro estudio, nos basamos estrictamente hablando en 'unidades fónicas[3] convencionales de token' (UFCT), que corresponden plenamente a unidades fonéticas en el caso de lenguas de las que disponemos transcripción fonético-fonológica[4] o a unidades de escritura fonémica en el caso de lenguas de las que no disponemos de transcripción pero que, en cualquier caso, poseen sistemas de escritura muy fonémica.

El coeficiente resultante de dividir el número de tokens entre el número de UFCT nos dará un valor de relación absoluto en el que el coeficiente máximo teóricamente posible sería 1 (si un idioma tuviera tantas palabras como número de UFCT), lo que es útil para comparar valores entre idiomas con un margen máximo establecido. A esto lo denominaremos 'índice de relatividad informativa', ya que estamos hablando de la relación entre el número de tokens y el número de UFCT empleadas en ellas.

El coeficiente, por su parte, resultante de dividir el número de UFCT entre el número de tokens, lo denominaremos 'índice de densidad informativa', ya que indica proporcionalmente la cantidad de información contenida en cada UFCT (evidentemente hablamos de una proporción matemática, de uso de recursos, no de que la información —la carga semántica— pueda ser fragmentada en fonemas). Ese coeficiente puede ser útil en el estudio de lenguas en cuanto no hay un valor máximo posible y se podría aplicar siempre a nuevas lenguas para ver si supera el coeficiente máximo encontrado hasta entonces.

Por último, tomada la información contenida en un texto y aplicando el valor matemático de 100, el coeficiente resultante de dividir 100 entre el número de tokens o entre el número de UFCT nos dará un coeficientre proporcional al número de unidades léxicas o al número de unidades fónicas que necesitan las distintas lenguas para expresar una misma información, por lo que a esto lo denominaremos 'índice de eficiencia informativa léxica' e 'índice de eficiencia informativa fónica'.

El índice de relatividad informativa y el índice de densidad informativa son valores absolutos resultantes del análisis intralingüístico de una lengua dada[5]. Los índices de eficiencia informativa léxica y fónica son valores relativos resultantes de comparar los coeficientes obtenidos en distintas lenguas sobre la base de un mismo contenido informativo (un mismo texto traducido a distintas lenguas). Somos conscientes de que no existe una única traducción posible exacta, pero, no obstante, creemos que el uso de ciertos textos minoriza este problema. Es por ello que en nuestro estudio nos valemos principalmente de la Declaración Universal de Derechos Humanos, ya que se trata de un texto con un contenido preciso que no da lugar a muchas interpretaciones ni valoraciones, además de ser un texto accesible en una gran cantidad de lenguas y de ser un texto de redacción reciente y con traducciones a las distintas lenguas en un periodo bastante reducido en términos históricos y de evolución de las lenguas.

Para el presente estudio hemos analizado un total de 459 lenguas pertenecientes a familias, macrofamilias y filos (o *phyla*) lingüísticos diferentes (incluyendo lenguas aisladas) y de características lingüísticas y tipología morfológica variada.

1. Marco teórico

Probablemente la primera alusión a un estudio sobre longitud de palabras sea una carta de Augustus de Morgan del 18 de agosto de 1851 dirigida a su amigo el reverendo W. Heald. En ella propone analizar la longitud media de las palabras para comprobar la autoría de un texto, y menciona en concreto que lo habría hecho con Heródoto, Tucídides y con las cartas paulinas, aunque no se conserva dicho estudio (Lord, 1958).

La propuesta de Augustus de Morgan fue llevada a cabo y plasmada por primera vez por Thomas Corwin Mendenhall en dos estudios de 1887 y 1901. En dichos

estudios, Mendenhall comparaba la longitud media de las palabras en diferentes fragmentos de mil palabras de diversos autores y comprobaba la coincidencia casi exacta de la longitud media de palabras en fragmentos de un mismo autor. El mismo Mendenhall ya sugería que este método podría ser aplicado a análisis de longitudes de sílabas y oraciones (Mendenhall, 1887, 1901; Grzybek, 2007).

Posteriormente, han sido numerosos los estudios dedicados a la longitud de las palabras y de sus distintos componentes en las lenguas del mundo (Menzerath, 1928, 1954; Menzerath & De Oleza, 1928; Altmann, 1980; Fenk & Fenk-Oczlon, 1993; Grotjahn & Altmann, 1993; Wimmer, Köhler, Grotjahn & Altmann, 1994; Ferrer-i-Cancho & Moscoso del Prado Martín, 2011, entre otros). En estos estudios uno de los planteamientos que mayor fortuna ha obtenido es la denominada Ley de Menzerath-Altmann, planteada por primera vez por Paul Menzerath en 1928 y luego desarrollada por Gabriel Altmann. Así, por ejemplo, estudios como los de Menzerath (1954), Altmann (1980), Fenk y Fenk-Oczlon (1993), Milicka (2014) o Coloma (2015) demuestran empíricamente que: 1) a mayor longitud de una oración medida en número de palabras, menor es la longitud de las palabras medida en número de sílabas; 2) a mayor longitud de una palabra medida en sílabas, menor es la longitud de las sílabas medida en fonemas; 3) a mayor longitud de una oración medida en palabras, menor es la longitud de las palabras medida en fonemas; 4) a mayor longitud de una oración medida en sílabas, menor es la longitud de las sílabas medida en fonemas.

Otros estudios sobre la longitud y distribución de palabras han ido siguiendo modelos como el geométrico de Elderton (1949) o Merkytė (1972), el de distribución de Poisson de Čebanov (1947), Fucks (1955a, 1955b, 1956) o Vranić y Matković (Vranić, 1965; Vranić & Matković, 1965), el de distribución de Lognormal de Herdan (1958, 1966) o Moreau (1963), el de distribución binomial negativa de Grotjahn (1982) o el de distribución Poisson-uniforme de Kromer (2001a, 2001b, 2001c, 2002).

Asimismo, la bibliografía científica ha estudiado la cantidad de recursos (por ejemplo, sílabas) necesarios en cada lengua para transmitir una misma información (Fenk-Oczlon, 1983) y la relación entre el uso de recursos lingüísticos (número y longitud de fonemas, sílabas y palabras) y la economía del lenguaje (la llamada, en este caso, 'economía cognitiva'; vid., por ejemplo, Fenk & Fenk-Oczlon, 1993).

Las características específicas del presente estudio y que suponen sus aportaciones con respecto a estudios precedentes son:

1. Estudio de un elevado número de lenguas (459 en total).
2. Uso de tokens y unidades fónicas como unidades objeto de estudio.
3. Análisis matemático y estadístico del número de palabras y fonemas y la correlación global entre ellas (no hemos trabajado la distribución de palabras por su longitud).

4. Análisis matemático y estadístico del empleo de recursos lingüísticos (palabras y fonemas) para expresar un mismo contenido informativo.
5. Análisis matemático y estadístico de la relevancia en el número y longitud de palabras y fonemas de factores como la tipología morfológica y genética, la complejidad del sistema fonológico o el carácter tonal o no tonal de las lenguas[6].

2. Marco metodológico

Según indicábamos, nosotros hemos analizado en cada una de las 459 lenguas consultadas su tipología morfológica, número de tokens y número de UFCT empleados en el texto fuente, índice de relatividad informativa, índice de densidad informativa, índice de eficiencia informativa léxica e índice de eficiencia informativa fónica. Nos sería imposible, por cuestión de espacio, incluir aquí los datos de las 459 lenguas, para lo cual remitimos a un trabajo nuestro previo (Vercher García & Bullejos Lorenzo, 2022) para ver la tabla total con todas las lenguas examinadas en la que indicábamos, para cada una de ellas, su tipo morfológico predominante, nº de tokens, nº de UFCT, índice de relatividad informativa, índice de densidad informativa, índice de eficiencia informativa léxica e índice de eficiencia informativa fónica.

Exponemos algunos de los datos más relevantes para las conclusiones extraídas y mostraremos unas tablas con las lenguas que presentan los valores máximos y mínimos en cada uno de los índices y unas tablas en las que se indican otros datos como la complejidad del sistema fonético-fonológico o el carácter tonal/no tonal de la lengua en cuestión.

En la actualidad, la lingüística no suele manejar tipos morfológicos puros, sino diversos índices graduales (índice de síntesis, de aglutinación, de flexión, composición, prefijación, sufijación, aislamiento, flexión pura y concordancia; Greenberg, 1954). No obstante, lo cierto es que en las descripciones de las lenguas se suele hablar de predominio de un tipo morfológico y, dado el carácter de estudio macroestadístico que aquí presentamos, nos valdremos de esta clasificación en tipos morfológicos predominantes, pues nos vale para los objetivos del presente estudio. Los códigos usados son los siguientes: A/I (analítica-aislante), SA (sintética aglutinante), SF (sintética fusionante) y SP (sintética polisintética).

Igualmente, somos conscientes de que la cuestión de la clasificación genética de las lenguas no está exenta de debate y carece de una visión única y homogénea entre todos los autores. Nosotros nos hemos basado en la filiación (*Family*) recogida por Phoible.org, el proyecto de Moran y McCloy (2019), aun siendo conscientes de la diversidad de criterios y clasificaciones que pueden encontrarse en distintos autores y que incluyen conceptos como los de subfamilia, familia, macrofamilia, filo, macrofilo, megafilo o gigafilo (Moreno Cabrera, 1997). Moran y McCloy (2019) distinguen en su proyecto 175 grupos genéticos distintos, además de las lenguas aisladas.

Con respecto al texto fuente, en la mayor parte de los casos se trata de la traducción completa y correcta de los diez primeros artículos de la Declaración Universal de los Derehos Humanos, como se ha señalado. En estos casos, se han usado las lenguas para el análisis de los cuatro índices estudiados. En algunos casos, no obstante, la traducción de esos diez primeros artículos está incompleta y, en otros casos, no nos hemos podido valer de ella y hemos usado otros textos fuente. En estos dos últimos casos, hemos usado las lenguas para el análisis del índice de relatividad informativa y del índice de densidad informativa por ser valores intralingüísticos, pero no los hemos usado para el análisis del índice de eficiencia informativa léxica ni del índice de eficiencia fónica, por ser valores relativos que se basan precisamente en la comparación entre lenguas. Para este análisis, nos valemos de tres conjuntos distintos de lenguas. Un primer grupo es el constituido por todas las lenguas consultadas indistintamente de su texto fuente (en total 459 lenguas). El segundo grupo es el conformado solo por aquellas lenguas que tienen como texto fuente los diez primeros artículos de la Declaración Universal de los Derechos Humanos (376 lenguas). El tercer grupo es el resultante de haber depurado el grupo de lenguas que usan los diez primeros artículos de la DUDH, eliminando las lenguas que presentan valores extremos y que matemáticamente distorsionan el estudio (296 lenguas).

El número de lenguas tomadas en consideración será, por tanto, de 459 en el caso de los índices de relatividad y densidad, mientras que en el caso de los índices de eficiencia analizamos, por un lado, el grupo de 376 lenguas y, por otro lado, el de 296 lenguas.

Puntualizamos, igualmente, que nos valemos del sistema lingüístico tal y como aparece en los textos fuente, y que normalmente será la lengua estandarizada, cuando exista, o una variante lingüística concreta en el caso de lenguas con un *continuum* de hablas (por ejemplo, el aymara). El sistema de análisis que proponemos permitiría, no obstante, hacer un estudio de una variedad lingüística dada, si así se quisiera[7].

3. Los índices de relatividad, densidad y eficiencia informativa. Datos cuantitativos, tipológicos, genéticos y fonético-fonológicos de lenguas con los valores máximos y mínimos (459 / 376 / 296 lenguas)

Además de los datos aportados por Vercher García y Bullejos Loreno (2022), debemos indicar aquí algunos de los datos y conclusiones más importantes en lo que a correlaciones generales entre tokens, UFCT y longitud de tokens se refiere, antes de pasar al objeto de estudio principal del presente artículo (las lenguas con valores máximos y mínimos de cada índice y las conclusiones que de ello pueden extraerse).

En primer lugar, mostramos la dispersión de la media de tokens (variable W) por tipos morfológicos. Los datos nos indican que en los idiomas de tipo SF los tokens están más agrupados en torno a la media mientras que los idiomas de tipo A/I están

más dispersos, tal y como se ve en la Figura 1. En proporción, los idiomas de tipo A/I y SP son los que más dispersos están sobre la media. Obsérvese también que estas son las poblaciones de menor tamaño.

Realizamos este mismo estudio para las unidades fonémicas convencionales de token (variable TCPU) y, de nuevo, como vemos en la Figura 2, los idiomas de tipo SF tienen las UFCT más agrupadas en torno a la media; además, la dispersión aparece a ambos lados, hay casi tantos idiomas con UFCT por encima de la mediana como por debajo. Esto ocurre también para A/I y SA. Sin embargo, para SP aparecen más por encima de la media y la dispersión es mayor. Vemos que también aparecen valores extremos todos por encima de la media, salvo para SF que aparecen a ambos lados. Observamos, igualmente, que las medias son muy parecidas para todos los idiomas, excepto para los de tipo SP, que es más elevada.

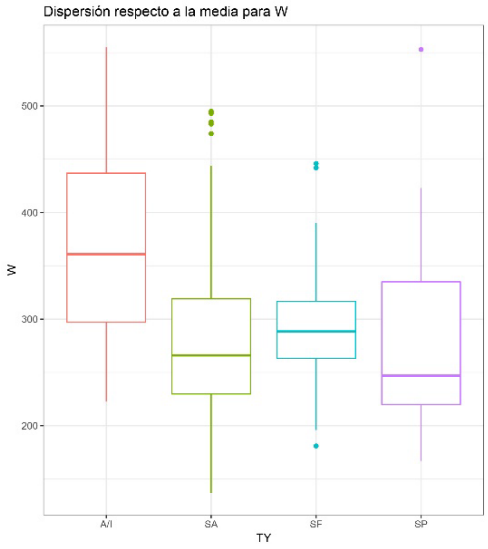

Figura 1. Diagrama para W.

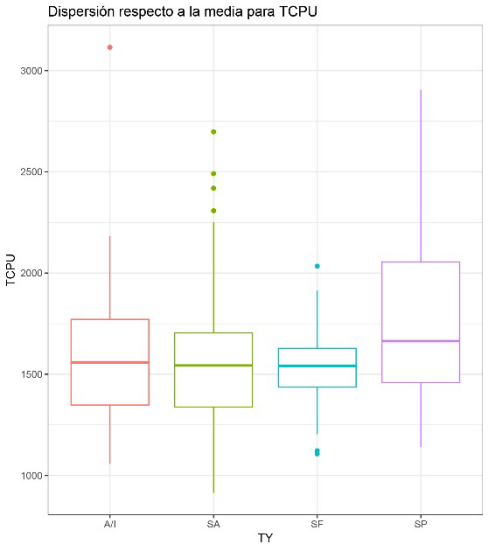

Figura 2. Diagrama para WLen.

Las medias, medianas y desviaciones típicas de la longitud de los tokens (WLen) por tipos morfológicos se muestran en la Tabla 1:

Tabla 1. Medias, medianas y desviaciones típicas para WLen.

Tipos morfológicos	**Medias para WLen**	**Medianas para WLen**	**Desviaciones típicas para WLen**
A/I	4.303524	4.079186	0.9750706
SA	5.776860	5.842068	1.4648717
SF	5.364402	5.258106	0.9093879
SP	6.803195	6.836879	1.8361209

Con respecto a la dispersión en la longitud media de palabras (token/UFCT) por tipo morfológico, vemos en la Figura 3 que los idiomas A/I y SF tienen una dispersión similar entre sí, al igual que los idiomas SA y SP. Los valores extremos aparecen principalmente por encima en todos los tipos morfológicos.

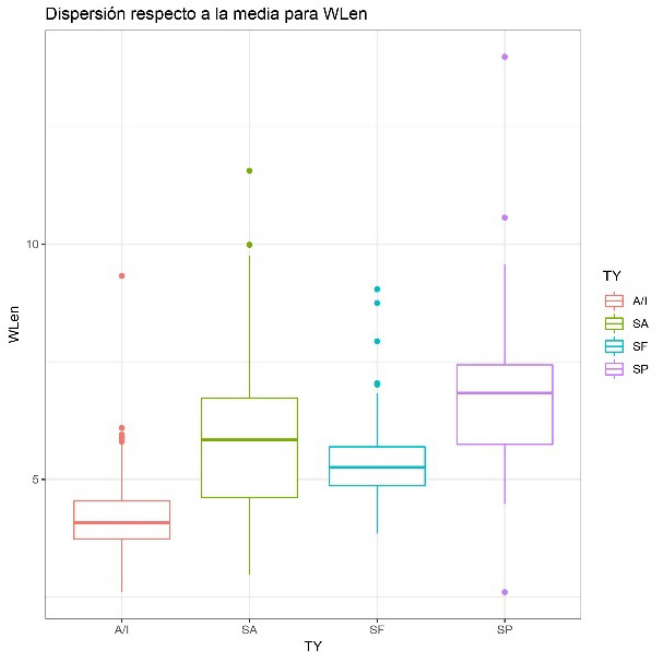

Figura 3. Diagrama para WLen.

Por su parte, en la Tabla 2, indicamos los coeficientes de correlación entre número de tokens (variable W) y número de UFCT (variable TCPU). Vemos que todas las correlaciones son positivas. Por tanto, confirmamos que a mayor número de tokens existe mayor número de UFCT, e indicamos el coeficiente de correlación exacto. Además, la correlación es mayor para los idiomas de tipo A/I y menor cuando se consideran todos en conjunto, sin atender al tipo morfológico.

Las correlaciones lineales no son muy fuertes, prácticamente, están todas por debajo de 0,5. La proporción es algo mayor en los idiomas de tipo A/I y SP, que es cercana a 1/2. En los idiomas de tipo SF y SA la proporción es cercana a 1/3 y, por tanto, es algo menor. La proporción de 1/3 también es la que obtenemos al tratar todos los idiomas en conjunto (sin tener en cuenta su tipo)[8].

Tabla 2. Coeficientes de correlación entre tokens y UFCT.

Tipos	**Correlaciones WTCPU**
A/I	0.5111750
SP	0.4420754
SF	0.3774098
SA	0.3537743
Todos	0.3340037

A continuación, pasamos a exponer los datos de aquellas lenguas que presentan los mayores valores por arriba y por abajo de cada uno de los índices analizados, indicando otras características lingüísticas como el carácter tonal o no tonal de la lengua, la complejidad del sistema fonológico o la familia lingüística.

La finalidad de tener en cuenta estos criterios es comprobar si los resultados arrojan la idea de que dichos criterios pueden influir en los valores de los índices. Por ejemplo, si ciertas familias lingüísticas tienden a destacar en algún índice (relatividad informativa, densidad informativa, etc.).

Tabla 3. Lenguas con mayor índice de relatividad informativa (459 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de relatividad informativa (tokens/UFCT)
Toba	SP	guaicuruanas	553	1440	0,384027778
Dangme	A/I	tonal; Atlántico-Congo	548	1431	0,382948987
Fon	A/I	tonal; Atlántico-Congo	409	1120	0,365178571
Nuosu/Yi	A/I	* tonal; sino-tibetanas	271	783	0,346104725
Chinanteco, Ajitlán	A/I	i tonal; otomangues	269	798	0,337092732
Moba	SA	tonal; Atlántico-Congo	407	1213	0,335531739
Vai	SA	tonal; mandé	474	1472	0,32201087
Tiv	SA	tonal; Atlántico-Congo	493	1571	0,313812858
Lamnso' (Lám nso')	SA	tonal; Atlántico-Congo	354	1139	0,310798946
Baoulé/Baule	A/I	tonal; Atlántico-Congo	490	1586	0,308953342

Tabla 4. Lenguas con menor índice de relatividad informativa (459 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de relatividad informativa (tokens/UFCT)
Inuktitut (Canadá)	SP	* esquimo-aleutianas	201	2874	0,06993737
Groenlandés	SP	esquimo-aleutianas	184	2.573	0,071511854
Malayalam	SA	dravídicas	137	1584	0,086489899
Karen (S'gaw)	A/I	* tonal; sino-tibetanas	145	1630	0,088957055
Tamil	SA	* dravídicas	21	224	0,09375
Caquinte	SP	arahuacas	268	2832	0,094632768
Ojibwa / Ojibway / Ojibwe	SP	* álgicas	18	182	0,098901099
Telugu	SA	dravídicas	186	1858	0,100107643
Kannada	SA	dravídicas	182	1777	0,102419809
Asháninca	SP	arahuacas	218	2088	0,10440613

Tabla 5. Lenguas con mayor índice de densidad informativa (459 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de densidad informativa (TCPU/tokens)
Inuktitut (Canadá)	SP	* esquimo-aleutianas	201	2874	14,29850746
Groenlandés	SP	esquimo-aleutianas	184	2.573	13,98369565
Malayalam	SA	dravídicas	137	1584	11,5620438
Karen (S'gaw)	A/I	* tonal; sino-tibetanas	145	1630	11,24137931
Tamil	SA	* dravídicas	21	224	10,66666667
Caquinte	SP	arahuacas	268	2832	10,56716418
Ojibwa / Ojibway / Ojibwe	SP	* álgicas	18	182	10,11111111
Telugu	SA	dravídicas	186	1858	9,989247312
Kannada	SA	dravídicas	182	1777	9,763736264
Asháninca	SP	arahuacas	218	2088	9,577981651

Tabla 6. Lenguas con menor índice de densidad informativa (459 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de densidad informativa (TCPU/tokens)
Toba	SP	vocales largas/breves; guaicuruanas	553	1440	2,6039783
Dangme	A/I	tonal; Atlántico-Congo	548	1431	2,611313869
Fon	A/I	tonal; Atlántico-Congo	409	1120	2,738386308
Nuosu/Yi	A/I	* tonal; sino-tibetanas	271	783	2,889298893
Chinanteco, Ajitlán	A/I	i tonal; otomangues	269	798	2,966542751
Moba	SA	tonal; Atlántico-Congo	407	1213	2,98034398
Vai	SA	tonal; mandé	474	1472	3,105485232
Tiv	SA	tonal; Atlántico-Congo	493	1571	3,186612576
Lamnso' (Lám nso')	SA	tonal; Atlántico-Congo	354	1139	3,217514124
Baoulé/Baule	A/I	tonal; Atlántico-Congo	490	1586	3,236734694

Tabla 7. Lenguas con mayor índice de eficiencia informativa léxica: Datos del estudio en bruto (376 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa léxica (100/tokens)
Malayalam	SA	vocales largas/breves; 63 fonemas; dravídicas	137	1584	0,729927007
Runyankore-rukiga/Nkore-kiga	SA	tonal; 25 fonemas; Atlántico-Congo	149	1126	0,67114094
Aymara	SA	vocales largas/breves; 32 fonemas; aymara	158	1228	0,632911392
Diola (Jola-Fogny)	SA	vocales largas/breves; 29 fonemas; Atlántico-Congo	162	914	0,617283951
Ogiek	SA	tonal; vocales largas/breves; 27 fonemas; nilóticas	162	1142	0,617283951
Wayuu	SP	vocales largas/breves; 28 fonemas; arahuacas	167	1206	0,598802395
Ndebele (Northern)	SA	78 fonemas; Atlántico-Congo	167	1329	0,598802395
Yucaguiro	SA	vocales largas/breves; 32 fonemas; yucaguiras	168	1299	0,595238095
Kaonde	SA	tonal; vocales largas/breves; 25 fonemas; Atlántico-Congo	171	1027	0,584795322
Zulu	SA	tonal; vocales largas/breves; 61 fonemas; Atlántico-Congo	171	1406	0,584795322

Tabla 8. Lenguas con mayor índice de eficiencia informativa léxica: Datos del estudio depurado (296 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa léxica (100/tokens)
Runyankore-rukiga/Nkore-kiga	SA	tonal; 25 fonemas; Atlántico-Congo	149	1126	0,67114094
Aymara	SA	vocales largas/breves; 32 fonemas; aymaras	158	1228	0,632911392
Diola (Jola-Fogny)	SA	vocales largas/breves; 29 fonemas; Atlántico-Congo	162	914	0,617283951
Ogiek	SA	tonal; vocales largas/breves; 27 fonemas; nilóticas	162	1142	0,617283951
Wayuu	SP	vocales largas/breves; 28 fonemas; arahuacas	167	1206	0,598802395
Ndebele (Northern)	SA	78 fonemas; Atlántico-Congo	167	1329	0,598802395
Yucaguiro	SA	vocales largas/breves; 32 fonemas; yucaguira	168	1299	0,595238095
Kaonde	SA	tonal; vocales largas/breves; 25 fonemas; Atlántico-Congo	171	1027	0,584795322
Lunda/Chokwe-lunda	SA	tonal; vocales largas/breves; 31 fonemas; Atlántico-Congo	172	1227	0,581395349
Quichua/Kichwa	SA	31 fonemas; quechua	174	1253	0,574712644

Tabla 9. Lenguas con menor índice de eficiencia informativa léxica: Datos del estudio en bruto (376 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa léxica (100/tokens)
Maorí	A/I	vocales largas/breves; 20 fonemas; austronesias	555	2181	0,18018018
Toba	SP	vocales largas/breves; 28 fonemas; guaicuruanas	553	1440	0,180831826
Dangme	A/I	tonal; vocales orales/nasales; 34 fonemas; Atlántico-Congo	548	1431	0,182481752
Hmong (Miao) Njua	A/I	tonal; vocales orales/nasales; 61 fonemas; hmong-mien	515	1940	0,194174757
Haitian Creole (popular)	A/I	vocales orales/nasales; 32 fonemas; indoeuropeas	509	1932	0,196463654
Nigerian Pidgin English	A/I	tonal; 28 fonemas; indoeuropeas	505	1756	0,198019802
Hmong Daw (Miao) Northern Qiandong	A/I	tonal; vocales orales/nasales; 56 fonemas; hmong-mien	499	1851	0,200400802
Chin Hakha	SA	tonal; vocales largas/breves; 52 fonemas; sino-tibetanas	495	2169	0,202020202
Tibetano	SA	tonal; vocales largas/breves; vocales orales/nasales; 34 fonemas; sino-tibetanas	494	2053	0,20242915
Tiv	SA	tonal; vocales largas/breves; 48 fonemas; Atlántico-Congo	493	1571	0,202839757

Tabla 10. Lenguas con menor índice de eficiencia informativa léxica: Datos del estudio depurado (296 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa léxica (100/tokens)
Maorí	A/I	vocales largas/breves; 20 fonemas; austronesias	555	2181	0,18018018
Hmong (Miao) Njua	A/I	tonal; vocales orales/nasales; 61 fonemas; hmong-mien	515	1940	0,194174757
Haitian Creole (popular)	A/I	vocales orales/nasales; 32 fonemas; indoeuropeas	509	1932	0,196463654
Nigerian Pidgin English	A/I	tonal; 28 fonemas; indoeuropeas	505	1756	0,198019802
Hmong Daw (Miao) Northern Qiandong	A/I	tonal; vocales orales/nasales; 56 fonemas; hmong-mien	499	1851	0,200400802
Samoano	A/I	vocales largas/breves; 20 fonemas; austronesias	492	1865	0,203252033
Baoulé/Baule	A/I	tonal; 33 fonemas; Atlántico-Congo	490	1586	0,204081633
Tahitiano	A/I	vocales largas/breves; 19 fonemas; austronesias	488	2043	0,204918033
Dioula	A/I	vocales largas/breves; vocales orales/nasales; 42 fonemas; mandé	482	1620	0,20746888
Maorí (Cook Islands) (Rarotongano)	A/I	vocales largas/breves; 20 fonemas; austronesias	477	1929	0,209643606

Tabla 11. Lenguas con mayor índice de eficiencia informativa fónica: Datos del estudio en bruto (376 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa fónica (100/tcpu)
Diola (Jola-Fogny)	SA	vocales largas/breves; 29 fonemas; Atlántico-Congo	162	914	0,10940919
Kabyè	SA	tonal; vocales largas/breves; 47 fonemas; Atlántico-Congo	223	996	0,100401606
Kaonde	SA	tonal; vocales largas/breves; 25 fonemas; Atlántico-Congo	171	1027	0,097370983
Kinyamwezi (Nyamwezi)	SA	tonal; vocales largas/breves; 53 fonemas; Atlántico-Congo	193	1037	0,096432015
Ditammari	SA	tonal; vocales largas/breves; vocales orales/nasales; 32 fonemas; Atlántico-Congo	285	1049	0,095328885
Batonu (Bariba)	SA	tonal; vocales largas/breves; vocales orales/nasales; 43 fonemas; Atlántico-Congo	310	1052	0,095057034
Twi (Akan Kasa) [Asante, ashanti]	A/I	tonal; ATR[9]; 33 fonemas; Atlántico-Congo	296	1057	0,094607379
Maninka	A/I	tonal; vocales largas/breves; vocales orales/nasales; 38 fonemas; mandé	285	1064	0,093984962
Sirionó	SA	vocales orales/nasales; 30 fonemas; tupíes	246	1089	0,091827365
Soninké (Soninkanxaane)	A/I	tonal; vocales largas/breves; 39 fonemas; mandé	297	1106	0,090415913

Tabla 12. Lenguas con mayor índice de eficiencia informativa fónica: Datos del estudio depurado (296 lenguas).

Nota: La tabla de lenguas en el estudio depurado coincidiría exactamente con la del estudio en bruto sin depurar, ya que ninguna de las lenguas que aparecen en la tabla ha sido eliminada en el estudio depurado.

Tabla 13. Lenguas con menor índice de eficiencia informativa fónica: Datos del estudio en bruto (376 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa fónica (100/tcpu)
Birmano/Myanmar	A/I	tonal; 46 fonemas; sino-tibetanas	334	3115	0,032102729
Pintupi-Luritja	SP	vocales largas/breves[10]; 23 fonemas; pama-ñunganas	339	2906	0,034411562
Amuesha-Yanesha	SP	vocales breves/largas/glotalizadas; 33 fonemas; arahuacas	420	2903	0,034447124
Shipibo-Conibo	SP	vocales orales/nasales; 26 fonemas; pano-tacanas	423	2892	0,034578147
Caquinte	SP	vocales largas/breves; 27 fonemas; arahuacas	268	2832	0,035310734
Matsés	SA	21 fonemas; Pano-Tacanan	364	2698	0,037064492
Groenlandés	SP	22 fonemas; esquimo-aleutianas	184	2.573	0,038865138
Quechua	SA	tonal; vocales largas/breves[11]; 17-37 fonemas[12]; quechuas	306	2491	0,04014452
siSwati (Swazi)	SA	tonal; 47 fonemas; Atlántico-Congo	292	2419	0,041339396
Rukonzo (Konjo)	SA	33 fonemas; Atlántico-Congo	278	2308	0,043327556

Tabla 14. Lenguas con menor índice de eficiencia informativa fónica: Datos del estudio depurado (296 lenguas).

Lengua	Tipología morfológica	Observaciones tipológicas y fonético-fonológicas	Nº de tokens	Nº de UFCT	Índice de eficiencia informativa fónica (100/tcpu)
Arabela	SP	16 fonemas; zaparoanas	335	2299	0,043497173
Amahuaca	SP	tonal; vocales largas/breves; vocales orales/nasales; 29 fonemas; pano-tacanas	259	2284	0,043782837
Aguaruna	SA	vocales orales/nasales; 20 fonemas; jívaras	338	2250	0,044444444
Maorí	A/I	vocales largas/breves; 20 fonemas; austronesias	555	2181	0,045850527
Náhuatl	SP	vocales largas/breves; 23 fonemas; uto-aztecas	320	2105	0,047505938
Amarakaeri	SP	vocales largas/breves; vocales orales/nasales; 33 fonemas; harákmbut–katukinas	292	2062	0,048496605
Achuar-Shiwiar	SA	vocales largas/breves; vocales orales/nasales; 27 fonemas; jívaras	289	2060	0,048543689
Weenhayek	SP	28 fonemas[13]; matacoanas	312	2054	0,048685492
Tahitiano	A/I	vocales largas/breves; 19 fonemas; austronesias	488	2043	0,048947626
Totonaco	SP	vocales largas/breves; vocales laringalizadas/no laringalizadas; 30 fonemas; totonacas	266	2026	0,049358342

4. Análisis de datos

Pasamos, en un primer lugar, a analizar los datos que arroja el estudio del número de tokens y UFCT, así como de los índices de relatividad, densidad y eficiencia informativa, aplicando como base de comparación la adscripción de cada lengua a un tipo morfológico predominante para todas las lenguas. En un segundo lugar, analizamos las lenguas que presentan los mayores valores por arriba y por abajo de cada uno de los índices, aplicando como base de comparación otros rasgos como el carácter tonal o no tonal de la lengua, la complejidad del sistema fonológico o la familia lingüística.

En primer lugar, podemos resumir a grandes rasgos que el número de tokens, el índice de relatividad informativa y el índice de densidad informativa (longitud media de las palabras) dependen del tipo morfológico. Dicho de otra manera, las lenguas

analíticas-aislantes van a emplear un mayor número de palabras para expresar el mismo contenido (en nuestro caso, rara vez bajarán de las 300 palabras) y las lenguas polisintéticas un menor número de palabras (rara vez subirán de las 260 palabras). A su vez, las lenguas analíticas-aislantes presentan las palabras más breves (rara vez por encima de los 4,4 UFCT por token de media), mientras que las lenguas polisintéticas presentan las palabras más largas (rara vez por debajo de las 6,5 UFCT por token de media). Estos datos se consiguen analizando todas las 376 lenguas que tienen los diez primeros artículos de la DUDH como texto fuente, y también depurando el análisis a 296 lenguas, eliminando los valores extremos. Estas ideas, en principio ya conocidas *a priori*, quedan confirmadas y expuestas en datos exactos en un estudio macro como el presente.

En segundo lugar, podemos concluir que el número total de UFCT empleado para expresar un mismo contenido de información no depende del tipo morfológico.

Pasamos a comentar estas conclusiones con más detalle; primero, a partir del estudio general de todas las lenguas que tienen como texto fuente los diez primeros artículos de la DUDH (tanto sin depurar -376 lenguas, como depurándolo -296 lenguas), usando como base de comparación el tipo morfológico a partir del análisis de las lenguas con los mayores y menores valores de los índices estudiados, teniendo como base de comparación otras características lingüísticas (filiación lingüística, sistema fonológico, tono).

De los datos expuestos podemos concluir que las lenguas tonales tienden a tener un mayor índice de relatividad informativa. De hecho, las diez lenguas con mayor índice de relatividad informativa son todas tonales, con excepción del toba (una lengua polisintética, a la que, además, habría que poner en cuarentena por presentar datos extremos). Esto, que *a priori* podría parecer ya lógico, viene a ser demostrado estadísticamente con los datos obtenidos durante el presente estudio. Las lenguas tonales necesitan un menor número de fonemas en cada palabra, ya que el tono sirve como rasgo fonológicamente distintivo (una misma vocal puede tener valor fonológico y, por tanto, semántico, diferente en función del tono). Esta característica les confiere mayor capacidad de distinción fonológica y semántica con menor cantidad de unidades fonémicas. Por su parte, y para corroborar esto, ninguna de las diez lenguas con menor índice de relatividad informativa es tonal, con excepción del sgaw karen, que es además una lengua tonal, pero de tonos nivelados[14], no modulados[15]. Lo contrario ocurriría con el índice de densidad informativa.

De los datos obtenidos también podemos afirmar que la complejidad del sistema fonológico (mayor o menor número de fonemas) no es un rasgo relevante que condicione a ninguno de los índices. Así, por ejemplo, entre las lenguas con mayor índice de eficiencia informativa léxica, encontramos desde las que tienen un sistema fonológico muy simple (como el runyankore, con solo 25 fonemas) hasta las que

tienen sistemas fonológicos muy complejos (el malayalam, con 63 fonemas, o el zulú, con 61). Lo mismo ocurre con las lenguas con menor índice de eficiencia informativa léxica (el maorí, 20 fonemas, frente al hmong (miao) njua, con 61). Otro tanto podría decirse del índice de eficiencia informativa fónica: entre las que presentan un mayor valor encontramos desde el kaondé (25 fonemas) hasta el kinyamwezi (53 fonemas), y entre las de menor índice de eficiencia informativa fónica encontramos desde el matsés (21 fonemas) hasta el suazi (47 fonemas).

La filiación genética de las lenguas sí parece ser pertinente y cabe un estudio más detallado de los datos. En primer lugar, debemos remitir a lo dicho anteriormente sobre la cuestión de las clasificaciones genéticas de las lenguas y los distintos niveles genéticos existentes (subfamilia, familia, macrofamilia, filo, macrofilo, megafilo o gigafilo) y a lo indicado con respecto a que nosotros nos valdremos de la clasificación realizada por Moran y McCloy en 175 grupos genéticos.

Con respecto a la relatividad informativa, las lenguas de la familia Atlántico-Congo claramente muestran una tendencia a tener valores elevados (dicho de otra forma, presentan un elevado número total de tokens con respecto al número total de UFCT), ya que seis de las diez lenguas con mayor índice de relatividad informativa pertenecen a esta familia, mientras que ninguna de las diez lenguas con menor índice de relatividad informativa es Atlántico-Congo[16]. Entre las lenguas con mayor índice de relatividad informativa se encuentran también representantes de las familias guaicurú, otomangue, mandé y sino-tibetana (nuosu-yi). De esta última familia también hay una lengua, sgaw karen, entre las que tienen menor índice de relatividad informativa, a pesar de que esta lengua pertenece a las lenguas karénicas, un grupo aparte dentro de las sino-tibetanas y diferente al grupo tibeto-birmano al que pertenecería el nuosu-yi. Incluso si ampliamos el análisis más allá del décimo puesto, varias de las siguientes lenguas pertenecen a la familia Atlántico-Congo (belanda viri, more, kasem, etc.), además de aparecer también otra lengua otomangue (el otomí mezquital –hñähñú–).

Por su parte, de estos datos se desprende también que las lenguas dravídicas tienden a presentar valores bajos de índice de relatividad informativa (cuatro de las diez lenguas con menor índice de relatividad informativa son dravídicas). Lo mismo podría decirse de las lenguas esquimo-aleutianas, arahuacas y álgicas, además del caso excepcional ya mencionado del sgaw karen. No obstante, también es cierto que entre las siguientes lenguas con menor índice de relatividad informativa aparecen lenguas de familias lingüísticas muy diversas, incluidas algunas de las Atlántico-Congo, aunque todavía encontramos un representante de la familia álgica (*swampy cree*) y otro de la sino-tibetana (birmano). Las lenguas que muestran, por su parte, mayores y menores valores del índice de densidad informativa vienen a ser las mismas, pero, lógicamente, con los valores invertidos.

Con respecto al índice de eficiencia informativa léxica, podemos constatar que las familias que tienden a presentar mayores valores son la Atlántico-Congo (cinco de las diez lenguas con mayor eficiencia informativa léxica), aymara, nilótica, arahuaca, yucaguira y dravídica, esta última con la salvedad de ser una de las lenguas eliminadas en el análisis depurado, donde sí aparece una lengua quechua. Especialmente, es relevante el caso de la familia yucaguira, ya que esta familia solo cuenta con dos lenguas (de hecho, el texto de la Declaración de los Derechos Humanos que recoge la ONU es en realidad yucaguiro septentrional, aunque se menciona simplemente como *yucaguiro*). Más allá de esas lenguas indicadas con mayor índice de eficiencia informativa, encontramos representantes de una diversidad de familias lingüísticas.

Entre las familias de las diez lenguas que recogen valores de índice de eficiencia informativa léxica más bajos estarían la austronesia, guaicurú, Atlántico-Congo, hmong-mien, indoeuropea y sino-tibetana, si bien habría que poner en cuarentena a los dos representantes de la familia indoeuropea, se trata en realidad de un pidgin y una lengua criolla. Más allá de esas diez primeras lenguas, encontramos representantes de la familia austronesia (samoano, tahitiano, rarotongano, tongano, marshalés), Atlántico-Congo (baoulé, lobiri, khisa), sino-tibetana (chin falam) o mandé (yulá, vai). Es decir, frente a una relativa mayor diversidad de familias lingüísticas en los índices anteriores, parece claro que familias como la austronesia, la Atlántico-Congo o la sino-tibetana dominan entre las que tienen menores valores de índice de eficiencia informativa léxica. En el análisis depurado, por su parte, encontramos que cuatro de las diez lenguas con valores más bajos son austronesias, dos son hmong-mien, dos indoeuropeas, una Atlántico-Congo y una mandé, es decir, los resultados vienen a ser similares a los arrojados por el análisis no depurado de lenguas, con lo que podemos concluir que estas familias claramente tienden a mostrar índices de eficiencia informativa léxica muy bajos (son las lenguas que necesitan emplear mayor cantidad de palabras para expresar un mismo contenido informativo).

El grupo de lenguas con mayor índice de eficiencia informativa fónica (es decir, empleo de menor número de fonemas para expresar una misma información) está claramente dominado por las lenguas Atlántico-Congo, con siete representantes. También hay dos lenguas mandé en este grupo y una tupí. Si ampliamos el análisis a las veinte primeras lenguas, encontraremos representantes de la familia Atlántico-Congo (*wolof, serer, gonja, fon, runyankore, belanda viri*), mandé (mendé), además de una lengua indoeuropea (el criollo santotomense –que de todos modos tiene sustrato de lenguas Atlántico-Congo–), una barbacoana (tsafiki) y una mayense (huasteco). Es decir, las lenguas Atlántico-Congo destacan claramente entre las que presentan un mayor índice de eficiencia informativa fónica, seguidas por las mandé[17]. Es muy significativo que en el estudio depurado la lista de lenguas con mayor índice de eficiencia informativa fónica coincide exactamente con la del estudio en bruto, es

decir, ninguna de ellas muestra valores extremos que nos hagan dudar de la validez de los resultados en bruto.

Finalmente, las familias que presentan tendencia a un menor índice de eficiencia informativa fónica en el análisis no depurado de lenguas son la sino-tibetana, pamañungana, arahuaca, pano-tacana, esquimo-aleutiana y quechua, aunque también hay dos representantes de la Atlántico-Congo, la lengua suazi y la lengua rukonzo. Ampliando el análisis hasta las veinte lenguas con menor índice de eficiencia informativa fónica, encontramos también representantes de la familia zaparoana (arabela), pano-tacana (amahuaco), chicham o jívara (aguaruna), austronesia (maorí), sino-tibetana (chin hakha), mayense (kaqchiquel), yuto-nahua o yuto-azteca (náhuatl), arahuaca (asháninca), austroasiática (jemer) y harákmbet (amarakaeri). Como vemos, entre las lenguas con menor índice de eficiencia informativa fónica descubrimos una amplia variedad de familias lingüísticas distintas, si bien es cierto que muchas de ellas pertenecerían a la gigafamilia de lenguas amerindias (frente a las lenguas con mayor índice de eficiencia informativa fónica, muchas de estas pertenecían a la macrofamilia de lenguas Niger-Congo).

El análisis depurado de lenguas arroja, no obstante, unos resultados muy diferentes y, al contrario de lo que ocurría con las lenguas con mayor índice de eficiencia informativa fónica, entre las diez lenguas con menor índice de eficiencia informativa del análisis depurado no se encuentra ninguna de las que aparecieron en el análisis en bruto. En el análisis depurado, en el que se han eliminado lenguas con valores extremos distorsionantes, hay que remontarse a la lengua arabela (posición 11 en el análisis bruto) para encontrar la lengua con el menor valor en este índice, y llegar a la lengua totonaca (posición 26 en el análisis en bruto) para encontrar la lengua que ocupa la décima posición de lenguas con menor índice de eficiencia informativa fónica. En el análisis depurado de lenguas con menores valores de este índice tampoco observamos ninguna familia que predomine, ya que encontramos representantes de una gran variedad de familias lingüísticas: zaparoana, pano-tacana, chicham (dos lenguas), austronesia (tres lenguas), yuto-azteca, harákmbet, mataca y totonaca.

CONCLUSIONES

De los datos y análisis expuestos podemos extraer una serie de conclusiones. En primer lugar, la filiación genética parece ser relevante en los índices informativos lingüísticos, ya que las lenguas de la hipotética macrofamilia Niger-Congo presentan claramente los mayores índices de eficiencia informativa fónica (son las lenguas que emplean un menor número de fonemas para expresar una misma información), mientras que la supuesta gigafamilia de lenguas amerindias presentan los menores índices de eficiencia informativa fónica (emplean un mayor número de fonemas que otras para expresar una misma información), con todas las salvedades que se quieran hacer por la falta de aceptación unánime de la existencia de la gigafamilia amerindia y

por haber bastante diversidad de familias entre las lenguas con menor índice de eficiencia informativa fónica. Esto es lógico porque lenguas emparentadas entre sí van a tener similitudes en sus recursos léxico y morfológicos (mayores similitudes que las semejanzas que pueda haber entre lenguas del mismo tipo morfológico, pero no emparentadas genéticamente).

Por otro lado, el tipo morfológico afecta directamente a los índices de relatividad informativa y de densidad informativa, es decir, al número de palabras que necesita una lengua dada para expresar un mismo contenido informativo y a la longitud media de las palabras de cada lengua. Sin embargo, también podemos concluir que no existe una correlación directa entre el tipo morfológico de una lengua y el número total de UFCT que necesita para expresar un mismo contenido informativo.

De los datos y del análisis realizado, también podemos concluir que algunos aspectos fonético-fonológicos, como el tono, sí afectan al índice de relatividad informativa, mientras que otros, como la complejidad del sistema fonológico, no afectan a ninguno de los índices informativos.

Otra conclusión más es que el principio de economía del lenguaje, tal y como se entiende desde Sweet (1888), Passy (1890), Vendryes (1939), Zipf (1949) y Martinet (1955), entre otros, parece existir dentro de cada lengua, pero si comparamos lenguas distintas, se observa que el esfuerzo articulatorio (número de fonemas-sonidos) para expresar un mismo contenido semántico varía significativamente de unas a otras. El principio de economía del lenguaje no se manifiesta en las lenguas reales como algo matemático y perfecto que haya conseguido reducir el esfuerzo articulatorio de todas las lenguas a unos mismos valores, sino que *de facto* cada lengua ha evolucionado condicionada por sus propias características fonético-fonológicas y los valores en número de fonemas o sonidos necesarios para expresar una misma información, variando significativamente de unas lenguas a otras.

Finalmente, podemos añadir que los índices de relatividad informativa, densidad informativa, eficiencia informativa léxica y eficiencia informativa fónica suponen nuevos indicadores para el estudio y descripción de las lenguas y abren las puertas a su aplicación en los campos de la tipología lingüística, la lingüística matemática, la lingüística comparativa o la traductología.

REFERENCIAS BIBLIOGRÁFICAS

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, 2, 1-10.

Čebanov, S. G. [Чебанов, С. Г.] (1947). О подчинении речевых укладов «индоевропейской» группы закону Пуассона. *Доклады Академии наук СССР*, *55*(2), 103-106.

Coloma, G. (2015). The Menzerath-Altmann Law in a Cross-Linguistic Context. *SKY Journal of Linguistics*, 28, 139-159.

Elderton, W. P. (1949). A Few Statistics on the Length of English Words. *Journal of the Royal Statistical Society, Series A (general)*, 112, 436-445.

Fenk-Oczlon, G. (1983). *Bedeutungseinheiten und sprachliche Segmentierung. Eine sprachvergleichende Untersuchung über kognitive Determinanten der Kernsatzlänge*. Tübingen: Narr.

Fenk, A. & Fenk-Oczlon, G. (1993). Menzerath's Law and the Constant Flow of Linguistic Information. En R. Köhler & B. Rieger (Eds), *Contributions to Quantitative Linguistics* (pp. 11-31). Dordrecht: Kluwer. DOI: 10.1007/978-94-011-1769-2_2

Ferrer-i-Cancho, R. & Moscoso del Prado Martín, F. (2011). Information Content Versus Word Length in Random Typing. *Journal of Statistical Mechanics: Theory and Experiment*, 12, 1-8.

Fucks, W. (1955a). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen / Arbeitsgemeinschaft fur Forschung des Landes Nordrhein-Westfalen, Heft 34a*. Köln/Opladen: Vs Verlag Fur Sozialwissenschaften.

Fucks, W. (1955b). Theorie der Wortbildung. *Mathematisch-Physikalische Semesterberichte zur Pflege des Zusammenhangs von Schule und Universität*, 4, 195-212.

Fucks, W. (1956). Mathematische Analyse von Werken der Sprache und der Musik. *Physikalische Blätter*, 16, 452-459.

Greenberg, J. H. (1954). A Quantitative Approach to the Morphological Typology of Languages. En R. F. Spencer (ed.). *Method and Perspective in Anthopology: Paper in Honor of Wilson D. Wallis* (pp. 192-220). Minneapolis: University of Minnesota Press (publicado posteriormente en *International Journal of American Linguistics, 26*(3), 178-194).

Grotjahn, R. (1982). Ein statistisches Modell f ¨ur die Verteilung der Wortl¨ange. *Zeitschrift für Sprachwissenschaft*, 1, 44-75.

Grotjahn, R. & Altmann, G. (1993). Modelling the Distribution of Word Length. Some Methodological Problems. En R. Köhler & B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 141-153). Dordrecht: Kluwer.

Grzybek, P. (2007). History and Methodology of Word Length Studies. The State of the Art. En P. Grzybek (Ed.), *Contributions to the Science of Text and Language* (pp. 15-90). Dordrecht: Springer.

Herdan, G. (1958). The Relation between the Dictionary Distribution and the Occurrence Distribution of Word Length and its Importance for the Study of Quantitative Linguistics. *Biometrika*, 45, 222-228.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin: Springer.

Jitwiriyanont, S. (2019). Ban Pa La-U Sgaw Karen Tones: An Analysis of Semitones, Quadratic Trendlines and Coefficients. En *RGJ Seminar Series LXXXII on Southeast Asian Linguistics* (pp. 60-77). Bangkok: Research Institute for Languages and Cultures of Asia at Mahidol University.

Kromer, V. V. (2001a). Word Length Model Based on the One-Displaced Poisson-Uniform Distribution. *Glottometrics*, 1, 87-96.

Kromer, V. V. (2001b). Двухпараметрическая модель длины слова «язык-жанр». En *Electronic archive Computer Science, March 8, 2001* [en línea]. Disponible en: http://arxiv.org/abs/cs.CL/0103007

Kromer, V. V. (2001c). Математическая модель длины слова на основе распределения Чебанова-Фукса с равномерным распределением параметра. En *IIнформатика и проблемы телекоммуникаций. Международная научно-техническая конференция. Материалы конференции* (pp. 74-75). Новосибирск: Изд-во СибГУТИ.

Kromer, V. V. (2002). Об одной возможности обобщения математической модели длины слова. En *IIнформатика и проблемы телекоммуникаций. Международная научно-техническая конференция. Материалы конференции* (pp. 139-140). Новосибирск: Изд-во СибГУТИ.

Lord, R. D. (1958). Studies in the History of Probability and Statistics. VIII: De Morgan and the Statistical Study of Literary Style. *Biometrika*, 45, 282.

Martinet, A. (1955). *Économie des changements phonétiques. Traité de phonologie diachronique*. Berne: Éditions A. Francke.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science*, Supplement, *214*(9), 237-249.

Mendenhall, T. C. (1901). A Mechanical Solution of a Literary Problem. *Popular Science Monthly*, *60*(7), 97-105.

Menzerath, P. (1928). Über einige phonetische Probleme. En A. Meillet (Ed.), *Actes du premier congrès international de linguistes* (pp. 104-105). Leiden: Sijthhoff.

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

Menzerath, P. & De Oleza, J. M. (1928). *Spanische Lautdauer. Eine experimentelle Untersuchung*. Berlin / Leipzig: de Gruyter.

Merkytė, R. J. (1972). Закон, описывающий распределение слогов в словах словарей. *Lietuvos matematikos rinkinys*, *12*(4), 125-131.

Milicka, J. (2014). Menzerath's Law: The Whole is Greater than the Sum of its Parts. *Journal of Quantitative Linguistics*, *21*(2), 85-99. DOI: 10.1080/09296174.2014.882187

Moran, S. & McCloy, D. (Eds.) (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History [en línea]. Disponible en: http://phoible.org.

Moreau, R. (1963). Sur la distribution des formes verbales dans le français écrit. *Études de linguistique appliquée*, 2, 65-88.

Moreno Cabrera, J. C. (1997). *Introducción a la lingüística: Enfoque tipológico y universalista*. Madrid: Síntesis.

Passy, P. (1890). *Etude sur les changements phonétiques et leurs caractères généraux*. Paris: Firmin-Didot.

Sweet, H. (1888). *A History of English Sounds from the Earliest Period*. Oxford: Clarendon Press.

UNITED NATIONS, *Universal Declaration of Human Rights* [en línea]. Disponible en: https://readtiger.com/www.unicode.org/udhr/assemblies/full_all.html https://www.ohchr.org/SP/UDHR/Pages/SearchByLang.aspx.

Vendryes, J. (1939). Parler par Èconomie. En *Mélanges de linguistique offerts à Charles Bally* (pp. 49-62). Genève: Université de Genève-Georg & Cie.

Vercher García, E. J. & Bullejos Lorenzo, M. (2022). Los índices de relatividad, densidad y eficiencia informativa en las lenguas: Estudio de las correlaciones matemáticas entre palabras y fonemas. *ELUA,* 36, 23-66.

Vranić, V. (1965). Statističko istraživanje hrvatskosrpskog jezika. *Statistička revija*, *15*(2-3), 174-185.

Vranić, V. & Matković, V. (1965). Mathematic Theory of the Syllabic Structure of Croato-Serbian. *Rad JAZU (odjel za matemati˘cke, fizi˘cke i tehni˘cke nauke, 10*(331), 181-199.

Wetzel, L. (2018). Types and Tokens. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition) [en línea]. Disponible en: https://plato.stanford.edu/archives/fall2018/entries/types-tokens/

Wimmer, G., Köhler, R., Grotjahn, R. & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, *1*(1), 98-106.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge (Mass.): Addison-Wesley Press.

NOTAS

[1] Wetzel (2018).

[2] El concepto mismo de 'palabra' es complejo y ha sido ampliamente debatido en la bibliografía científica. Se sale de los límites del presente trabajo profundizar en ello. Autores que han abordado esta cuestión son Jespersen, Admoni, Brøndal, Šahmatov, Vinogradov, González Calvo, Ušakova, Mauro, Krivono-sov, Ščerba, Steblin-Kamenskij, Serebrennikov, Suprun, Sunik, Migirin, Reichenbach, Potebnja, Hockett, entre muchos otros.

[3] Cuantitativamente (el aspecto relevante en nuestro estudio) pueden equivaler indistintamente a fonemas o sonidos (adviértase que tanto en /ˈdweNde/ como en [ˈdwe̯n̪de̯] el número de UFCT es el mismo: seis). No obstante, debemos puntualizar al respecto algunas cuestiones. En primer lugar, el acento, la aspiración, el tono o el carácter largo de una vocal no se contabilizan cuantitativamente como una unidad más, sino como rasgos caracterizadores de un mismo sonido, fonema o sílaba. En segundo lugar, contabilizamos las africadas como una sola unidad fonética; así, por ejemplo, /ʧ/ se contabiliza como una sola UFCT. A esto podríamos añadir que hemos tenido que unificar criterios de distintos autores a la hora de contabilizar la pausa glotal como un fonema más o no, considerar en ciertos casos dos sonidos como alófonos de un mismo fonema o no, considerar dos moraes como un mismo fonema o no, etc.

[4] Es el caso del alemán, catalán, danés, español, francés, hebreo, inglés, inuktitut, irlandés, italiano, japonés, latín, luxemburgués, portugués, tamil, telugu y turco.

[5] Evidentemente, el valor de estos índices puede variar en función del tipo de texto (coloquial, técnico, poético, etc.). Más que de índices de relatividad, densidad y eficiencia de las 'lenguas' habría que hablar de índices de un texto concreto en una lengua concreta. No obstante, los datos macro nos dan una idea de qué grado general de relatividad, densidad y eficiencia lingüística tienen las distintas lenguas.

[6] La necesidad de tener en cuenta estos aspectos en los estudios sobre longitud de palabras ya fue señalada por Wimmer et al. (1994).

[7] Por ejemplo, en ciertas zonas de Andalucía la sílaba terminada en vocal + /s/ pierde el fonema /s/ abriendo la vocal, con lo que numéricamente sería un fonema en lugar de dos en esas variedades lingüísticas.

[8] El que la correlación esté cerca del 1/2 indica, por ejemplo, que si cuadruplicáramos el número de tokens, doblaríamos el número de UFCT. Si está cerca de 1/3 tendríamos que sextuplicar el número de tokens para duplicar el número de UFCT.

[9] ATR (*advanced tongue root*) es un rasgo fonético por el que un sonido se pronuncia con la raíz de la lengua avanzada, teniendo ello un valor fonológico distintivo.

[10] Pero solo tiene seis vocales en total, ya que cuenta con tres largas y tres breves.

[11] En lenguas quechua centrales (Áncash, Huánuco, Pasco, Junín y Lima).

[12] Varía entre los 17 fonemas del quechua de Ayacucho y los 37 del quechua del sur de Bolivia.

[13] El sistema vocálico es bastante simple (6 fonemas) pero tiene un sistema consonántico complejo en el que muchas consonantes cuentan con una serie simple, otra glotalizada y otra aspirada.

[14] En lingüística se distingue entre lenguas tonales con 'tonos nivelados' o 'de registro' (tonos sin variación en su realización, se distinguen con respecto al nivel de otros tonos, las palabras polisílabas en este tipo de lenguas suelen pronunciarse con un mismo tono –frente a diferenciar el tono de cada sílaba– y suelen tener un valor gramaticalmente –no léxicamente– distintivo) y lenguas tonales con 'tonos modulados' o 'de contorno' (el tono –contorno– del sonido varía durante su realización –ascendente, descendente–, en palabras polisílabas cada sílaba puede tener su propio tono –y muchas palabras monosílabas pueden distinguirse solo por el tono– y tienen normalmente un valor distintivo léxico y semántico –más que gramatical, como ocurre con los tonos nivelados–).

[15] Así lo consideran la mayoría de los trabajos sobre sgaw karen, aunque Jitwiriyanont (2019) hace constar que el sistema tonal del dialecto ban pa la-u es de tipo modulado.

[16] Por un lado, es cierto que este predominio de las lenguas Atlántico-Congos podría estar en consonancia con el predominio de esta familia lingüística (más del 19% de las lenguas recogidas por phoible.org pertenecen a esta familia, frente al resto repartidas entre otras 174 familias, además de las lenguas aisladas), pero esta conclusión debería quedar refutada por el hecho de que ninguna de las lenguas con menor índice de relatividad informativa (y ninguna de las lenguas con mayor índice de densidad informativa) pertenece a esta familia lingüística.

[17] A este respecto cabe destacar que hay autores que adscriben tanto las Atlántico-Congo como las mandé a la macrofamilia de las lenguas Níger-Congo.