Revista Signos ISSN 0718-0934
© 2012 PUCV, Chile
DOI: 10.4067/S0718-09342012000100005

45(78) 70-82

Índice de Palabras de Contenido (IPC) y Distribución Porcentual de *Legomena* (DPL) en artículos de investigación en español*†

Index of Content Words (ICW) and Percent Distribution of Legomena (PDL) in Research Articles in Spanish

Ken Matsuda
kmatsuda@userena.cl
Universidad de La Serena
Chile

Scott Sadowsky
ssadowsky@gmail.com
Universidad de la Frontera
Chile

Omar Sabaj
omarsabaj@userena.cl
Universidad de La Serena
Chile

Recibido: 11-I-2011 / Aceptado: 8-VIII-2011

Resumen: A partir de una revisión de los índices clásicos en estadística léxica (Leyes de Estoup-Zipf-Mandelbrot), se proponen dos índices lingüísticos que buscan aportar nuevos datos en la descripción de textos especializados. Se presenta un caso de la aplicación de estos índices a un corpus representativo y multidisciplinar de artículos de investigación en español, que se contrasta con otros siete corpus a modo de control. Si bien desde un punto de vista general los índices tienen un comportamiento estable en los distintos registros, de forma específica, los índices permiten distinguir registros de alta y de baja especialización y dan cuenta de la variación disciplinar de los corpus analizados.

Palabras Clave: Leyes de potencia, lingüística cuantitativa, tipos de *Legomena*, artículos de investigación.

Abstract: Based on a review of classic indices in lexical statistics (Laws of Estoup-Zipf-Mandelbrot), two linguistic indices are proposed in order to contribute new data in the description of specialized texts. A case is presented in which these indices are applied to a representative, multidisciplinary corpus of research articles in Spanish that is contrasted with seven other *corpora* serving as a control group. Although, from a general point of view the proposed indices present stable behaviour in different registers, implemented specifically, the proposed indices allow the distinction between high and low specialized registers and reveal the disciplinary variation of the *corpora* analyzed.

Key Words: Power Laws, quantitative linguistics, types of *Legomena*, research articles.

INTRODUCCIÓN

> Mathematicians believe in [Zipf's law] because they think that linguists have established it to be a linguistic law, and linguists believe in it because they, on their part, think that mathematicians have established it to be a mathematical law.
>
> Gustav Herdan (1996:33)

En la primera mitad del siglo XX, el lingüista estadounidense George Kingsley Zipf constató la existencia de ciertos patrones matemáticos que se manifiestan en la relación entre el rango y la frecuencia de las palabras de un texto, con el propósito de sustentar el principio de mínimo esfuerzo, el cual explica la comunicación como un fenómeno matemático e informático (Shannon, 1948). Desde su formulación, la Ley de Zipf (1949), basada a su vez en las observaciones del secretario del Instituto Francés de Esteganografía, Jean Baptiste Estoup, y en los postulados del sociólogo y economista italiano Wilfredo Pareto, ha tenido una enorme influencia en campos variados, claramente insospechados en un comienzo: la estructura del código genético (Sastre, Cañibano, Boubeé, Rey, Suhurt & Scempio, 2010), el funcionamiento de internet (Adamic & Huberman, 2002), la teoría del caos (Larsen-Freeman, 1997), la teoría de los fractales y las finanzas (Mandelbrot & Hudson, 2006), el modelamiento de bases de datos textuales y la recuperación de la información (Baeza-Yates & Navarro, 2005), entre otros.

Dada la especial interacción que la Ley de Zipf (formalmente no es una ley, sino un modelo) supone entre los fenómenos estadísticos y lingüísticos (Wyllys, 1981), desde la matemática se han propuesto diversos ajustes a esta ley (Jiang, Shan, Jiang & Xu, 2002; Izsák, 2006) que suponen una divergencia de la constante que ella predice de forma general. Estos ajustes reflejan de mejor forma rangos específicos de la distribución de datos (Sun, Shaw & Davis, 1999; Evert, 2006), o bien datos con rangos de gran escala (Gunther, Levitin, Schapiro & Wagner, 1996; Izquierdo, 1998; Debowski, 2002).

Aunque estos ajustes a la Ley de Zipf dan luces sobre aspectos de la distribución de la probabilidad de aparición de las palabras en textos de distinta naturaleza, su abordaje ha sido casi exclusivamente matemático y estadístico, generando una falta de reflexión sobre las implicancias de estos fenómenos desde un punto de vista netamente lingüístico. Algunos autores (García, 2004; Sabaj, 2004; Evert, 2006) han señalado ya el riesgo que implica la utilización de los datos estadísticos en la investigación lingüística de forma ciega, esto es, sin contar con fundamentos lingüísticos que orienten su interpretación o permitan extraer de estos datos estadísticos conclusiones relevantes para la teoría lingüística.

Por otra parte, si bien existen estudios en los que, además de un modelamiento matemático avanzado, se pueden constatar descripciones o implicancias lingüísticas acuciosas (Johansson, 1981; Montemurro, 2001; Ferrer & Solé, 2002; Ferrer & Solé, 2003; Ferrer, 2005; Maslov & Maslova, 2006), los corpus considerados en estos estudios son mayoritariamente en idioma inglés. La modelización estadística avanzada de textos en español que incorporen reflexiones lingüísticas profundas se han investigado fundamentalmente desde un punto de vista comparativo (Ha, Stewart, Hanna & Smith, 2006), y solo escasamente teniendo al español como foco principal del análisis. Las excepciones a este punto lo constituyen el estudio de Rojo (2008) y los datos que se pueden obtener de Sadowsky y Martínez (2008).

En este contexto, el objetivo del presente artículo es proponer dos índices estadístico-lingüísticos para la descripción de textos en español, específicamente artículos de investigación científica. Para ello, en la primera parte del trabajo presentamos una discusión teórica con evidencia empírica de los conceptos fundamentales en los que se sustenta nuestra propuesta. En la segunda sección, en el apartado de

la metodología, exponemos de forma detallada las características del corpus de análisis, los métodos usados para su descripción estadística general, los procedimientos de análisis llevados cabo, y el modo particular para el cálculo de los índices propuestos. En la última parte del trabajo, presentamos los principales resultados y sus implicancias, para finalizar con algunas de las conclusiones que se pueden derivar del estudio realizado.

1. Antecedentes teóricos

1.1. Ley de Zipf y los tipos de Legomena

Ya en 1932, George Zipf, a partir de las observaciones realizadas por Estoup (Petruszewycz, 1973), describió el comportamiento estadístico de la distribución de las palabras en los textos. Propuso que en un texto cualquiera, existe una relación matemática entre la frecuencia absoluta de cada palabra y el lugar que ocupa en el listado de las palabras usadas en el texto, ordenadas por su frecuencia decreciente. Esto se puede expresar mediante la siguiente fórmula:

$$(1) \quad f(r) = \frac{C}{r^{\beta}}$$

Siendo: f la frecuencia, r el rango, C una constante y β la pendiente.

Conceptualmente, el modelo propuesto por Zipf (1949) pertenece a una familia de modelos matemáticos conocidos como leyes de potencias (*power laws*), las cuales intentan dar cuenta de la relación que existe entre el rango y la frecuencia de un determinado fenómeno. En una distribución de ley de potencias, las frecuencias decrecen según un exponente cuando la variable aleatoria, en este caso el rango, aumenta.

La característica principal de las leyes de potencias es su invarianza de escala: si para la muestra lingüística definida por el texto de una novela de 500 páginas, la palabra más frecuente (de rango 1) aparece 10.000 veces, la de rango 10 aparecerá solo 100 veces (Simon, 1955). Como consecuencia, se puede establecer que en todo documento elaborado en lenguaje natural, existe un gran grupo de palabras de escasa utilización.

La propiedad de invarianza hace que una ley de potencias quede determinada por su exponente, formando las funciones con el mismo exponente una clase de equivalencia. Desde una perspectiva gráfica, la ley de potencias puede interpretarse como una línea aproximadamente recta en un gráfico doble-logarítmico, lo cual queda manifiesto en la siguiente reformulación de la ecuación de la fórmula 1:

$$(2) \quad \log(f(r)) = \log(C) - \beta \log(r)$$

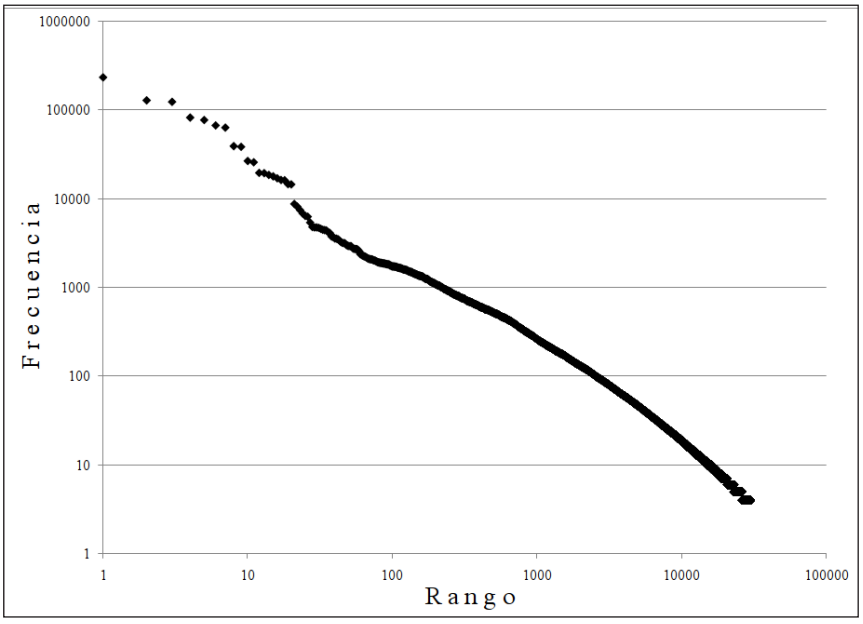

Gráfico 1. Aplicación de fórmula de Zipf al corpus académico-profesional de humanidades del Codicach.

El resultado de la aplicación de esta fórmula a los primeros 30.000 *tokens* del corpus académico-profesional de humanidades del Codicach se puede apreciar en el Gráfico 1.

Según Gelbukh y Sidorov (2001), la estimación de los parámetros de este modelo lineal –es decir, la constante *C* y la pendiente β– se puede calcular aplicando una regresión lineal basada en el Método de los Mínimos Cuadrados; en consecuencia, concluyen que los coeficientes exponenciales de la Ley de Zipf dependen sustancialmente del idioma del texto analizado. Arriban a esta conclusión a partir de la comparación de los coeficientes calculados para 39 diferentes textos de distintos géneros de ficción en ruso y en inglés, todos de un tamaño considerable y comparable para ambos idiomas. A una conclusión similar llegan Ha et al. (2006) al comparar los coeficientes en inglés, irlandés, latín y español.

Se puede inferir de estos estudios que en cualquier indagación de las frecuencias de un texto, se generan tres áreas: una pequeña zona de palabras de alta frecuencia, una zona de frecuencia media, y una gran zona de palabras de baja frecuencia. Sin embargo, los modelos lineales de estimación solo representan adecuadamente las palabras de la zona intermedia. Las posibles causas de esta desviación, según Gelbukh y Sidorov (2001), serían las diferencias gramaticales y la riqueza léxica que existen entre los diversos idiomas.

En síntesis, lo que establece la Ley de Zipf es una relación constante entre el rango y el tamaño o frecuencia que es útil en la descripción empírica de diferentes muestras de producción lingüística. Zipf (1949) atribuyó este fenómeno a la ley del mínimo esfuerzo, la que postula que –en el caso que nos concierne– siempre es más fácil usar una palabra conocida que una menos conocida.

Intentando hallar una explicación más razonable de la relación rango-frecuencia, Mandelbrot (1966) utiliza conceptos de la teoría de la información para proponer una reformulación de la Ley de Zipf. La esencia de su contribución consiste en considerar el costo de la comunicación de una palabra, en términos de la cantidad de letras y el espacio que las separa. Este costo se incrementa con el número de letras que tiene una palabra y con la extensión en un mensaje. Mandelbrot (1966) propuso la siguiente corrección al modelo de Zipf:

$$(3) \quad (r+m)^b f = C$$

Donde *f* es la frecuencia de una palabra; *r* es el rango de la palabra; y *b*, *C*, *m* son constantes que dependen del corpus. La mencionada corrección es una generalización, en tanto la ley Zipf es un caso particular del modelo propuesto por Mandelbrot (1966). En relación a los coeficientes, Gelbukh y Sidorov (2001) plantean que estos dependen del idioma. A partir del trabajo de Mandelbrot (1966), también se pueden establecer medidas para la complejidad de un corpus, como la entropía y la dimensión fractal de los textos, definida ésta, como el inverso del coeficiente *b*.

Estrechamente vinculados a la Ley de Zipf, se encuentran los denominados *Hapax Legomena*, a saber, palabras cuya frecuencia es igual a 1 (Rojo, 2008). En general, la proporción de estas palabras es estable en cualquier muestra lingüística: constituyen alrededor del 50% del total de las palabras distintas de un texto dado. Diversos autores han propuesto métodos para estimar las palabras que ocurren con una misma frecuencia (1, 2, 3, 4, etc.), entre los que destacan el valor promedio de los rangos de las palabras con misma frecuencia (Wyllys, 1981) y el método del valor máximo del orden (Sun et al., 1999).

1.2. Palabras de contenido y palabras vacías

En cualquier lengua se distinguen las palabras de contenido de las palabras vacías, denominadas en algunos casos palabras funcionales (Di Tullio, 2010). Se trata de tipos de palabras que, si bien pueden expresar contenidos semánticos y cumplen funciones sintácticas específicas, la mayoría de ellas no tiene en sí misma capacidad referencial. Algunos autores (Ha et al., 2006; Maslov & Maslova, 2006) han identificado los rangos en los que aparecen las palabras vacías: van del 1 al 20, es decir, son las palabras más frecuentes de cualquier corpus. Así también, debido a su alta frecuencia en cualquier texto, estas palabras a menudo son obviadas en los motores de búsqueda de Internet (Adamic & Huberman, 2002). En la mayoría de los trabajos en lingüística de corpus, estas palabras se incorporan, erróneamente en nuestra opinión, en lo que se denomina el tamaño del vocabulario, que se identifica con el número de palabras distintas de un corpus, denominadas *types*. Si bien las palabras vacías son parte de cualquier documento, no corresponden a elementos del vocabulario de la misma forma que

otras categorías gramaticales (las clases abiertas) como los sustantivos, los verbos y los adjetivos. El Índice de Palabras de Contenido busca dar cuenta de la proporción de las palabras vacías frente a las de contenido.

2. Metodología

En esta sección, presentamos las preguntas, las hipótesis y los objetivos de la investigación. Así también, describimos en esta sección las características del corpus, las técnicas de muestreo, así como los procedimientos específicos llevados a cabo para estimar los índices propuestos y mostrar su comportamiento en el corpus de análisis.

2.1. Preguntas de investigación, hipótesis y objetivos

La pregunta general que se busca responder con esta investigación se refiere al comportamiento de algunos índices léxico-estadísticos en un conjunto de registros diversificados. Específicamente, buscamos responder si estos índices presentan un comportamiento estable a través de estos registros. Como hipótesis, podríamos esperar que, dada la naturaleza diversificada de los registros incluidos en nuestro corpus, algunos de estos índices sean sensibles a las diferencias que se pueden establecer entre ellos, por ejemplo, el grado de especialización, las distintas áreas de la ciencia estudiadas, entre otros aspectos. El objetivo particular de este trabajo es proponer dos índices estadístico-léxicos y describir su comportamiento en artículos de investigación en español, contrastando este comportamiento con un conjunto de registros diversificados a modo de control.

2.2. El corpus

El corpus de análisis de esta investigación fue recolectado en el marco del Proyecto FONDECYT 11080097. Se trata del Corpus de Artículos de Investigación en Español (CaiE), que es representativo de la biblioteca electrónica Scielo Chile. Para la delimitación del número de casos incluidos y su nivel de representatividad, se utilizó una técnica de muestreo aleatorio estratificado con afijación óptima proporcional, de forma que estuvieran representadas todas las disciplinas y las revistas contenidas en la base. Para ello, primero se hizo un estudio exploratorio que tuvo como objetivo conocer qué número de artículos de investigación en español se habían publicado entre los años 2000 y 2008 en la Base Scielo Chile. En ese trabajo (Sabaj, Matsuda & Fuentes, 2010), se generaron criterios funcionales para determinar qué era considerado un artículo de investigación, junto con otros criterios de inclusión y exclusión (idioma y año de publicación), y, de esta forma, se determinó el número de casos totales por cada uno de los estratos (revistas, disciplinas y áreas de la ciencia). Una vez conocido el número de unidades muestrales (en nuestro caso, los artículos) de cada estrato (revistas, disciplinas y áreas de las ciencias), se utilizó la siguiente fórmula para determinar el número de artículos a considerar para contar con un corpus representativo:

$$(4) \quad n = \frac{NZ_{\frac{\alpha}{2}} s^2}{Nd^2 + Z_{\frac{\alpha}{2}} s^2}$$

Donde *n* representa el tamaño de la muestra a estimar, *N* es el tamaño de la población, con un nivel de significación *alfa* del 5%. Esto supone que nuestra afirmación sobre el tamaño de la muestra tiene un 95% de probabilidad de ser verdadera. A esto, se denomina zeta de alfa medios, $Z_{\frac{\alpha}{2}}$, y su valor de probabilidad es de 1,96 con un *alfa* del 5%. s^2 representa la estimación de la varianza de la población. *d es* la precisión o el error del muestreo, en nuestro caso es aproximadamente 0,07688, el que representa un 7%. Para contar con un corpus representativo, el tamaño *n* era de 161,6 número que se aproximó a 162. Luego se utilizó un muestreo aleatorio simple, proporcional a cada estrato, para seleccionar los artículos que conformarían el corpus. Para la clasificación y agrupación de los artículos en disciplinas y áreas de la ciencia, se utilizaron los criterios de la clasificación de las ciencias de la UNESCO. Para más detalles de los procedimientos de recolección, la representatividad y el tipo de muestreo utilizado en el CaiE, véase Sabaj y Matsuda (2010). En la Tabla 1, se muestra la conformación del corpus:

Tabla 1. Conformación del CaiE.

Área de la ciencia	N° artículos
Ciencias de la salud	75
Ciencias de la tierra	11
Ciencias de la vida	9
Ciencias exactas	9
Ciencias sociales	45
Humanidades	13
TOTAL	**162**

Tal como se muestra en la Tabla 1, el CaiE consta de 162 artículos, agrupados en 6 áreas de la ciencia. Representa a 58 revistas y a 22 disciplinas que publicaron números entre los años 2000 y 2008 en la Base Scielo Chile.

Como una forma de contrastar los datos obtenidos del CaiE, se analizaron otros siete Corpus:

- El Diccionario de Frecuencias del Castellano Moderno, Difcam (Sadowsky & Martínez, 2011), un diccionario de frecuencias léxicas que contempla 637 millones de *tokens*.
- Cuatro subcorpus académico-profesionales del Corpus Dinámico del Castellano de Chile, Codicach (Sadowsky, 2006), compuestos por un total de 7,8 millones de *tokens*.
- La versión de la Reina Valera de la Biblia completa con un total de 990.835 *tokens*.
- Un extracto de 355.622 *tokens* del Corpus Oral de Referencia de la Lengua Española Contemporánea (Marcos Marín, 1992), correspondiente a interacciones verbales de servicios.

2.3. Procedimientos de análisis y forma de obtención de los índices

Los documentos del CaiE y los extractos de la Biblia y del Corpus Oral de Referencia de la Lengua Española fueron traspasados a texto plano y procesados con el programa de concordancias Antconc (Anthony, 2010). Luego, se calcularon las estadísticas relevantes con hojas de cálculo. Las estadísticas del Difcam y de los subcorpus del Codicach se calcularon con el programa Frequency List Wizard (Sadowsky, 2010). Debe señalarse que se eliminaron los números, los signos de puntación y todos los elementos gráficos (tablas, gráficos, fotos, etc.) de todos los corpus analizados en la presente investigación.

2.3.1. Índice de Palabras de Contenido (IPC)

El Índice de Palabras de Contenido es un índice porcentual que determina cuántas palabras de contenido tiene un texto. Conceptualmente equivale a la noción de densidad léxica propuesta por Williamson (2009). Para su cálculo, se utilizó la siguiente fórmula:

$$(5) \quad IPC = \frac{(Tk - PV)}{Tk} 100$$

Donde *Tk* corresponde al número total de *tokens* del corpus, y *PV* a la frecuencia total de las palabras vacías en el mismo corpus. Para su conteo, se utilizó el *software* de concordancias Antcon (Anthony, 2010) y se consideraron como palabras vacías las siguientes categorías gramaticales: los artículos definidos e indefinidos, las preposiciones, las conjunciones, las disyunciones y los nexos (el listado de las formas consideradas como palabras vacías en esta investigación se presenta en el Anexo).

*2.3.2. Distribución Porcentual de Legomena (*DPL*)*

A diferencia de otros trabajos consultados, en esta investigación no solo calculamos la frecuencia y porcentaje de los *Hapax Legomena* (1-*Legomena*), sino que extendimos esta noción también a las palabras que tienen una frecuencia igual a 2 (*2-Legomena*), a 3 (*3-Legomena*) y a 4 (*4-Legomena*). Para la contabilización de los tipos de *Legomena* (1, 2, 3 y 4), se calculó la frecuencia absoluta de los cuatro tipos de *Legomena*, junto con su porcentaje de ocurrencia respecto del total de *types* del corpus correspondiente, utilizando la siguiente fórmula:

$$(6) \quad DPL = \frac{Legomena_i}{Tp} 100$$

Donde *i* corresponde al tipo de *Legomena* ($Legomena_{1, 2, 3, 4}$), y *Tp* al número de *types* (número total de palabras distintas) de cada corpus.

2.3.3. Variabilidad o Type-Token Ratio (TTR)

Para proporcionar otra métrica general para el análisis de los corpus, se calculó además la variabilidad de los textos en términos de las palabras distintas y las palabras totales que contienen, utilizando la clásica fórmula del *Type-Token Ratio* (TTR):

$$(7) \quad TTR = \frac{Tp}{Tk}$$

Donde *Tp* corresponde al número de *types*, y *Tk* al número de *tokens*.

3. Resultados y discusión

En la Tabla 2, se presentan los datos generales del análisis realizado: el tamaño de los corpus en términos *types* y *tokens*, y su tasa de variabilidad (TTR):

Tabla 2. *Types, tokens* y tasa de variabilidad (TTR).

Área de la ciencia en el CaiE	*Types*	*Tokens*	TTR
Ciencias de la salud	30.447	255.530	0,119
Ciencias de la tierra	10.135	60.276	0,168
Ciencias de la vida	8.777	40.819	0,215
Ciencias exactas	7.144	37.247	0,192
Ciencias sociales	42.424	464.562	0,091
Humanidades	12.062	77.795	0,155
Corpus de control	***Types***	***Tokens***	**TTR**
Difcam (Totalidad)	1.162.224	637.495.334	0,002
Codicach (Cs. aplicadas)	24.441	440.690	0,055
Codicach (Cs. naturales)	119.880	3.639.846	0,033
Codicach (Cs. sociales)	53.482	1.086.081	0,049
Codicach (Humanidades)	91.311	2.651.755	0,034
La Biblia	39.958	990.835	0,040
Oralidad	24.623	355.622	0,069

Tal como señala Richards (1987), si la TTR se acerca a 0, el corpus es menos variable, mientras que si se acerca a 1, es más variable. Según Richards (1987), la TTR está altamente determinada por el tamaño de los registros, de forma que los textos más grandes tienen siempre una TTR más baja. Esta idea se ve constatada en los datos que arrojó la presente investigación. En efecto, el corpus de mayor tamaño (el Difcam) tiene una TTR extremadamente baja (0,002); los corpus de gran tamaño (humanidades, ciencias naturales y ciencias sociales del Codicach, más la Biblia) cuentan con una TTR que varía entre 0,033 y 0,049; los corpus medianos (ciencias sociales y oralidad del CaiE; ciencias aplicadas del Codicach) tienen una TTR aún mayor (0,055 a 0,091); y los corpus más pequeños (humanidades y las ciencias del CaiE, a excepción de las Ciencias las sociales) tienen las TTR más altas de todos (0,119 a 0,215).

A continuación, en la Tabla 3, se expone el Índice de Palabras de Contenido para cada uno de los corpus analizados.

Tal como se expone en la Tabla 3, el índice de palabras de contenido (IPC) tiene un comportamiento estable en todos los corpus, con un promedio aproximado de un 70%. Desde un punto de vista lingüístico, esto supone que para construir una oración de 10 palabras, se requieren en promedio 3 palabras carentes de contenido léxico, las cuales se emplean solo para articular sintácticamente la

Tabla 3. Índice de Palabras de Contenido.

Área de la Ciencia en el CaiE	IPC
Ciencias de la salud	72.64
Ciencias de la tierra	72.63
Ciencias de la vida	76.16
Ciencias exactas	70.17
Ciencias sociales	68.24
Humanidades	68.66
Corpus de control	**IPC**
Difcam (Totalidad)	67.83
Codicach (Cs. aplicadas)	71.05
Codicach (Cs. naturales)	71.10
Codicach (Cs. sociales)	66.58
Codicach (Humanidades)	66.73
La Biblia	67.30
Oralidad	70.83

expresión de dicho contenido. Del mismo modo, se puede establecer que un mayor índice de palabras de contenido supone un mayor grado de especialización de los registros, es decir, una alta presencia de términos. Al contrario, un menor IPC supone una presencia mayor de palabras de uso cotidiano. Suponemos también que un mayor IPC puede estar fuertemente influenciado por las restricciones impuestas en relación a la extensión de los documentos que se presentan en revistas especializadas. Así, los autores de los textos altamente especializados como los de las ciencias de la salud, las ciencias de la tierra y las ciencias de la vida, se ven en la obligación de transmitir, en el menor espacio posible, una mayor cantidad de contenido. Respecto de este índice, es interesante apuntar que es sensible a las disciplinas. Así por ejemplo, tanto en los corpus de estudio como en los de control, las humanidades y las ciencias sociales se encuentran por debajo del índice de las otras áreas de las ciencias (exactas, aplicadas, de la salud, de la tierra).

Si contrastamos estos datos con los de la Tabla 2, podemos observar además que los resultados relativos al IPC son independientes del tamaño de los registros. La regularidad en este índice de palabras de contenido estaría basada en la universalidad de las restricciones sintácticas y gramaticales que cualquier lengua impone a la expresión de los significados.

En la Tabla 4, se muestra la distribución porcentual de los tipos de *Legomena* en los registros analizados. En la segunda columna de la Tabla 4, se puede observar que los 1-*Legomena*, constituyen un promedio del 50% de los *types*. En términos generales, estos datos son coincidentes con la mayoría de los estudios que describen la presencia de los *Legomena* en registros diversos en otras lenguas (Johansson, 1981; Montemurro, 2001; Ferrer & Solé, 2002; Ferrer & Solé, 2003; Ferrer, 2005; Ha et al., 2006; Maslov & Maslova, 2006) y confirman el cumplimiento de la Ley Zipf y la generalización propuesta por Mandelbrot (1966), al menos en el promedio. Si analizamos los datos de forma individual, se aprecian, aunque menores, ciertas variaciones: a) existen algunos registros que pasan el promedio (en orden decreciente: ciencias de la vida, ciencias exactas, humanidades, ciencias de la tierra, ciencias de la salud); b) un registro se ajusta de forma perfecta a Zipf (ciencias sociales); y c) todos los registros de control están bajo el promedio. Como sabemos, el principio de mínimo esfuerzo, defendido por Zipf, (1949) supone que si un sujeto tiene la posibilidad de elegir una palabra de uso cotidiano, elegirá esa palabra antes que una palabra poco frecuente. Una mayor cantidad de *Hapax* en un registro supone un mayor número de palabras poco frecuentes lo que, oponiéndonos a Zipf, implica seguir un principio de máximo esfuerzo. Debido a la naturaleza especializada de los registros acá considerados, los datos muestran que en todos los registros en estudio, se sobrepasa la predicción de Zipf; en otros términos, en este tipo de textos es común que existan más 'palabras raras' que en un texto no especializado. Aunque esta constatación no es novedad para los estudiosos de la terminología,

Tabla 4. Distribución porcentual de los tipos de *Legomena* por *types*.

Área de la Ciencia en el CaiE	**% *Legomena*$_1$**	**% *Legomena*$_2$**	**% *Legomena*$_3$**	**% *Legomena*$_4$**
Ciencias de la salud	51,60	15,36	7,62	4,50
Ciencias de la tierra	52,54	16,79	7,52	5,00
Ciencias de la vida	55,71	18,24	7,06	4,49
Ciencias exactas	54,84	17,39	7,43	4,37
Ciencias sociales	50,10	15,27	7,72	4,89
Humanidades	54,63	16,29	7,36	4,30
Corpus de control	**% *Legomena*$_1$**	**% *Legomena*$_2$**	**% *Legomena*$_3$**	**% *Legomena*$_4$**
Difcam (Totalidad)	41,61	13,15	7,01	4,94
Codicach (Cs. aplicadas)	38,64	16,32	8,62	5,64
Codicach (Cs. naturales)	43,16	14,32	7,58	4,84
Codicach (Cs. sociales)	44,18	15,36	7,91	5,06
Codicach (Humanidades)	44,18	14,88	7,46	4,86
La Biblia	41,56	15,29	7,94	5,40
Oralidad	47,96	15,60	7,85	4,98

desde un punto de vista psicolingüístico, específicamente, el de la producción escrita del discurso científico, esto implica que la selección del vocabulario es cuidada, que el conocimiento que subyace a la temática es altamente especializado, que la audiencia es restringida, entre otros aspectos. Esto, en términos zipfeanos, maximiza el esfuerzo en la comunicación. Por el contrario, en los registros que presentan un porcentaje bajo el promedio que la Ley de Zipf predice, se minimiza el esfuerzo, por lo que no importa si se utilizan muchas veces las mismas palabras. Desde un punto de vista discursivo, se trata de textos destinados a audiencias amplias con pocos términos especializados.

Al observar consecutivamente las columnas de los tipos de *Legomena*$_{1, 2, 3, 4}$ de la Tabla 4, se puede advertir que los datos presentan menor variación. La proporción de palabras que se utilizan 2, 3 y 4 veces es sucesivamente más estable y la diferencia esperada, considerando la distinta naturaleza de los registros, desaparece. En general, se pueden establecer las siguientes constataciones:

- En promedio, un 16, 2% de las palabras distintas de un registro se utilizan dos veces.
- En promedio, un 7,5% de las palabras distintas de un registro se utilizan tres veces.
- En promedio, un 4,7% de las palabras distintas de un registro se utilizan cuatro veces.

Desde un punto de vista estadístico, estos datos son siempre atrayentes, ya que suponen encontrar regularidades matemáticas en datos aparentemente caóticos. Asimismo, desde una perspectiva lingüística, estas regularidades son especialmente interesantes, ya que corresponden a patrones estocásticos que son independientes de la naturaleza diversificada de los registros analizados y, por ello, no son sensibles a los contextos de producción y consumo de estos textos. Otra inferencia que se puede obtener de los datos de la Tabla 4, es de los datos no considerados en el estudio, a saber, la proporción de palabras con frecuencia igual o mayor a 5. Si se suman las filas de la Tabla 4 y luego se obtiene un promedio da como resultado un 80%. Esto quiere decir que en promedio un 20% de las palabras de cualquier texto tienen una frecuencia mayor o igual a 5, dato coincidente con la distribución de Pareto (Petruszewycz, 1973), en este caso, pocas palabras tienen las más altas frecuencias y existen muchas palabras con baja frecuencia.

Dentro del 20% de palabras con alta frecuencia, sabemos, se encuentran las palabras vacías (las más frecuentes) y algunas palabras de contenido de alta frecuencia. Estas palabras de contenido de alta frecuencia, que son excelentes predictores de la temática de un texto, imponen un desafío para el cálculo de un punto de transición, modelado exclusivamente desde la estadística, para predecir el punto exacto donde se distinguen las palabras vacías y las de contenido.

Muchas de estas regularidades, sin embargo, están lejos de tener una buena explicación relacionada con la naturaleza de los textos y la voluntad humana y serían coincidentes con los patrones descritos por Zipf (1949).

CONCLUSIONES

En este trabajo hemos revisado críticamente las aplicaciones de la Ley de Zipf, y hemos propuesto dos índices en estadística léxica para la descripción de artículos de investigación en español: el índice de palabras de contenido y la distribución porcentual de los *Legomena*. Del trabajo realizado, se pueden extraer las siguientes conclusiones.

En primer término, se puede establecer que la tasa de variabilidad de los registros analizados es dependiente de los tamaños de las muestras y no de las características distintivas provenientes de la naturaleza diversificada de esos registros.

En segundo lugar, se puede concluir que el índice de palabras de contenido tiende a ser una constante de un 70% con una variación mínima entre los registros, es independiente del tamaño del corpus, pero analizado en detalle, sí es sensible al grado de especialización de los registros.

En tercer término, podemos determinar que el porcentaje de *Legomena*$_1$ también tiende a ser una constante, pero existen diferencias entre los registros: algunos que sobrepasan el promedio y los registros control, que están por debajo. Dentro de los artículos de investigación, también existen diferencias entre las disciplinas, en los que el mayor porcentaje de *Hapax*$_1$ se vincula a una alta presencia de terminología. Hemos sugerido con estos datos (IPC y *Hapax*$_1$), que en el caso de algunos de los artículos de investigación aquí considerados (los

más especializados), podríamos contradecir el principio de mínimo esfuerzo propuesto por Zipf (1949). En otras palabras, aquellos registros con un mayor índice de palabras de contenido y mayor porcentaje de *Hapax*$_{1,}$ suponen un esfuerzo máximo en la comunicación: son registros altamente especializados, cargados de terminología, cuyo emisor y destinatario son expertos y su audiencia reducida.

En cuarto lugar, la distribución de los otros tipos de *Legomena*$_{2,3,4}$ es estable, y no existen diferencias entre los artículos de investigación de las diferentes áreas, tampoco entre estos y los registros control, lo que nos permite concluir que en este caso, sí se cumplen las regularidades descritas por Zipf (1949). Este índice particular también es independiente del tamaño de los registros.

Por último, y respondiendo de esta forma nuestra pregunta de investigación podemos concluir que, si bien desde un punto de vista general (en los promedios) los índices propuestos tienden a tener un comportamiento estable, la micro-variación de estos índices permite distinguir algunos registros, en base a su grado de especialización, a la disciplina a la que pertenecen y a la audiencia a la que están destinados. En este sentido, estos índices o bien algunas de sus dimensiones específicas, pueden ser considerados como rasgos que merecen ser tenidos en cuenta al momento de realizar descripciones de los fenómenos lingüístico-estadísticos asociados a los textos científicos.

Los resultados de esta investigación pueden ser útiles para aquellos interesados en estadística léxica, específicamente, para replicar los procedimientos que aquí utilizamos en otros tipos de textos o en comparación con otras lenguas. Asimismo, esta indagación puede ser de provecho para los estudiosos de los fenómenos estadísticos y su vinculación con el uso del lenguaje en textos científicos.

REFERENCIAS BIBLIOGRÁFICAS

Adamic, L. & Huberman, B. (2002). Zipf's Law and the Internet. *Glottometrics,* 3, 143-150.

Anthony, L. (2010). *Antconc software* [en línea]. Disponible en: http://www.antlab.sci.waseda.ac.jp/software.html

Baeza-Yates, R. & Navarro, G. (2005). Modelling text databases. En R. Baeza-Yates, J. Glaz, H. Gzyl, J. Hüsler & J. Palacios (Eds.), *Recent advances in applied probability* (pp.1-25). Berlin/ Heidelberg: Springer.

Di Tullio, A. (2010). *Manual de gramática del Español*. Buenos Aires: Waldhuter.

Debowski, L. (2002). Zipf's Law against the text size: A half-rational model. *Glottometrics*, 4, 49-60.

Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), 177-190.

Ferrer, R. (2005). The variation of Zipf's Law in human language. *The European Physical Journal B*, 44, 249-257.

Ferrer, R. & Solé, R. (2002). Zipf's Law and random texts. *Advances in Complex Systems*, *5*(1), 1-6.

Ferrer, R. & Solé, R. (2003). Least effort and the origins of scaling in human language. *PNAS*, *100*(3), 788-791.

García, A. (2004). *Los procedimientos matemáticos en estudios e investigaciones lingüísticas: Utilidad y riesgo* [en línea]. Disponible en: http://usuarios.multimania.es/angarmegia/ProceMatS.pdf

Gunther, R., Levitin, L., Schapiro, B. & Wagner, P. (1996). Zipf's Law and the effect of ranking on probability distribution. *International Journal of Theoretical Physics, 35*(2), 395-417.

Gelbukh, A. & Sidorov, G. (2001). *Zipf and Heaps Laws' coefficients depend on language.* Proceedings of the second CICLing Conference, Intelligent Text Processing and Computational Linguistics. Ciudad de México, México.

Ha, L., Stewart, D., Hanna P. & Smith, F. (2006). Zipf and Type-Token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics, 1*(8), 1-12.

Herdan, G. (1996). *The advanced Theory of Language as choice and chance*. Berlin: Springer-Verlag.

Izsák, J. (2006). Some practical aspects of fitting and testing the Zipf-Mandelbrot model. *Scientometrics, 67*(1), 107-120.

Izquierdo, J. (1998). El declive de los grandes números: Mandelbrot y la estadística social. *Empiria. Revista de Metodología de Ciencias Sociales,* 1, 51-84.

Jiang, G., Shan, S., Jiang, L. & Xu, X. (2002). A new rank-size distribution of Zipf's Law and its applications. *Scientometrics, 54*(1), 119-130.

Johansson, S. (1981). Word frequencies in different types of English texts. *ICAME NEWS,* 5, 1-13.

Larsen-Freeman, D. (1997). Chaos/complexity Science and second language acquisition. *Applied Linguistics,* 18, 141-165.

Mandelbrot, B. (1966). *Information Theory and Psycholinguistics: A theory of word frequencies*. Cambridge, MA: MIT Press.

Mandelbrot, B. & Hudson, R. (2006). *Fractales y finanzas*. Barcelona: Tusquets.

Marcos Marín, F. (1992). *Corpus oral de referencia de la lengua española contemporánea* [en línea]. Disponible en: http://www.lllf.uam.es/~fmarcos/informes/corpus/corpulee.html

Maslov, V. & Maslova, T. (2006). On Zipf's Law and rank distributions in linguistics and semiotics. *Mathematical Notes, 80*(5), 679-691.

Montemurro, M. (2001). Beyond the Zipf-Mandelbrot Law in quantitative linguistics. *Physica: A Statistical Mechanics and its Applications, 300*(4-5), 567-578.

Petruszewycz, M. (1973). L'histoire de la loi d'Estoup-Zipf: Documents. *Mathématiques et Sciences Humaines*, 44, 41-56.

Richards, B. (1987). Type-token ratios: What do they really tell us? *Journal of Child Language*, 14, 201-209.

Rojo, G. (2008). *Lingüística de corpus y lingüística del español*. Ponencia presentada en el XV Congreso de la Asociación de Lingüística y Filología de América Latina, Montevideo, Uruguay.

Sabaj, O. (2004). Especificidad, especialización y variabilidad verbal: Una aproximación computacional en estadística léxica. *Revista Signos. Estudios de Lingüística*, *37*(56), 75-89.

Sabaj, O. & Matsuda, K. (2010). *Informe CaiE* [en línea]. Disponible en: http://omarsabaj.wordpress.com/anexos-investigaciones/

Sabaj, O., Matsuda, K. & Fuentes, M. (2010). Un modelo para la homogeneización de las clases textuales de la biblioteca electrónica Scielo-Chile: La variabilidad del artículo de investigación en diversas disciplinas. *Información Tecnológica*, *21*(6), 133-148.

Sadowsky, S. (2006). *Corpus dinámico del castellano de Chile* [en línea]. Disponible en: http://ssadowsky.hostei.com/codicach.html

Sadowsky, S. (2010). *Frequency list Wizard* [en línea]. Disponible en http://ssadowsky.hostei.com/flw.html

Sadowsky, S. & Martínez, R. (2008). *Lista de frecuencias de palabras del castellano de Chile (Lifcach)* [en línea]. Disponible en: http://ssadowsky.hostei.com/lifcach.html

Sadowsky, S. & Martínez, R. (2011). *Diccionario de frecuencias del castellano moderno (Difcam*) [en línea]. Disponible en http://ssadowsky.hostei.com/*corpora*.html

Sastre, P., Cañibano, A., Boubeé, C., Rey, G., Suhurt, V. & Scempio, V. (2010). *Leyes de Estoup - Zipf– Mandelbrot y el lenguaje genético* [en línea]. Disponible en: http://www.unsa.edu.ar/domefa/documentos/VIII-reunion/04-Leyes%20de%20Estoup.pdf

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal,* 27, 379-423.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425-440.

Sun, Q., Shaw, D. & Davis, Ch. (1999). A model for estimating the occurrence of same-frequency words and the boundary between high- and low frequency words in texts. *Journal of the American Society for Information Science, 50*(3), 280-286.

Williamson, G. (2009). *Lexical density* [en línea]. Disponible en: http://www.speech-therapy-information-and-resources.com/lexical-density.html

Wyllys, R. (1981). Empirical and theoretical bases of Zipf's Law. *Library Trends, 30*(1), 53-64.

Zipf, G. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Cambridge University Press.

Zipf, G. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

ANEXO

Palabras vacías

Artículos
el|la|los|las|un|una|unos|unas

Preposiciones
a|ante|bajo|cabe|con|contra|de|desde|durante|en|entre|hacia|hasta|mediante|para|por|según|sin|sobre|tras|

Conjunciones
y|e|ni|pero|sino

Disyunciones
o|u|o bien

Nexo
que

* Este trabajo se enmarca en el desarrollo del Proyecto FONDECYT 11080097, 'El artículo de investigación a través de las disciplinas: El caso del indexador Scielo Chile'.

† In Memoriam Benoit Mandelbrot 1924-2010.