

APLICACIÓN DE LAS LEYES DE ZIPF EN LOS TEXTOS ESPAÑOLES

Daniel Petřík

Abstract: Zipf's Laws are rules that describe the distribution of words in the text by their frequency, the number of different meanings and other features. Its author, Harvard professor George Kingsley Zipf (1902-1950), while studying natural languages mentioned the so-called Principle of least effort, which means that the language is focused on the economy and clarity. The principle has an impact on how the speaker unconsciously choose the terms by which he expresses his ideas (for example, the mostly used words are usually the shortest and with multiple meanings, etc.). Zipf's Laws are valid in any natural language and even in other natural systems; however, we must take into account the individual features of each language. In Spanish texts, for example, is necessary to deal with a higher frequency of determinants and prepositions. These and other peculiarities of its implementation are discussed in this paper.

Key words: lexeme; word; distribution of words; meaning; clusters of words.

Resumen: Las Leyes de Zipf son reglas que describen la distribución de palabras en el texto mediante su frecuencia, el número de significados diferentes u otras características. Su autor, profesor de Harvard George Kingsley Zipf (1902-1950), al estudiar las lenguas naturales menciona el llamado Principio del mínimo esfuerzo, o sea que el lenguaje se centra en la economía y en la claridad. El principio tiene cierto impacto en cómo el hablante elige inconscientemente los términos mediante los cuales expresa sus ideas (por ejemplo, las palabras mayormente utilizadas suelen ser las más cortas y con significados múltiples, etc.). Las Leyes de Zipf están en vigor en cualquier lenguaje natural, e incluso en otros sistemas naturales; sin embargo, hay que tener en cuenta algunas características individuales de cada lengua. En los textos españoles, por ejemplo, es necesario hacer frente a una mayor frecuencia de determinantes y preposiciones. Estas y otras peculiaridades de su aplicación trataremos en este artículo.

Palabras claves: lexema; palabra; distribución de palabras; significado; agrupaciones de palabras.

1. Motivación

Un elemento esencial de la comunicación humana es la transmisión de determinados contenidos entre los comunicantes. Si el hablante quiere expresar una idea, en su conciencia se llevan a cabo ciertos procesos a los que pertenece también la selección de recursos léxicos pertinentes, lo cual lleva a la siguiente pregunta: ¿Qué influye al hablante en la elección de lexemas durante la comunicación?

La selección de palabras efectuada por el hablante durante la comunicación es una cuestión científica en la frontera entre la lingüística, la psicología cognitiva y la estadística, la cual abarca el tema de las Leyes de Zipf. En este texto siempre supondremos que el hablante es una persona adulta que tiene el español como su lengua materna. El proceso

de la selección de recursos léxicos se puede caracterizar por dos aspectos que aparecen a la vez¹:

1º condicional (cualquier factor que se asocia con las características de comunicantes o la situación comunicativa)

2º incondicional (como resultado de la naturaleza del lenguaje)

2. El aspecto condicional de la selección de lexemas

El modo de cómo el hablante elige las palabras para expresar sus ideas en la comunicación se ve afectado (condicionado) por varios factores extralingüísticos. Por ejemplo, las mujeres jóvenes que viven en la capital con el sueldo superior a la media se expresarían por medio de otras palabras, es decir, seleccionarían en ciertos contextos otros lexemas que los hombres jubilados en un pueblo en otra región. En la tabla siguiente presentamos los factores extralingüísticos más comunes de la elección de lexemas:

Tabla 1: Los factores extralingüísticos de la elección de lexemas

el hablante (el emisor)	factores demográficos	el sexo, la edad, etc.
	factores socioeconómicos	la pertenencia a grupos o estratos sociales, la influencia de los medios de comunicación, etc.
	factores geográficos	las variantes regionales de la lengua (dialectos, etc.)
	características personales	cualidades personales, el estado psíquico, el actual estado emocional, la formación y educación, etc.
la situación comunicativa	características de la situación comunicativa	¿cuándo?, ¿dónde? y ¿cómo? se realiza la comunicación
el oyente (el receptor)	características del oyente	las características demográficas, socioeconómicas, geográficas y personales del oyente

3. El aspecto incondicional de la selección de lexemas

Si omitimos todos los factores que pertenecen al aspecto condicional, podría parecer que durante la comunicación el hablante elige las palabras arbitrariamente (entre todos los lexemas posibles para expresar una idea escoge cada uno de ellos aproximadamente con la misma frecuencia). Sorprendentemente en el proceso de la selección de palabras encontramos otros principios que no están directamente ligados a las características de los comunicantes, ni al carácter de la situación comunicativa o a los rasgos del lenguaje.² Estos principios se consideran las propiedades generales de todas las lenguas naturales. El comienzo de los estudios estadísticos de las generalidades lingüísticas se vincula con Jean-Baptiste Estoup (1868-1950), un taquígrafo en el parlamento francés que investigó reglas estadísticas aplicables a textos para facilitarse el trabajo. Luego Estoup recogió estas reglas en el libro *Les gammes sténographiques* (1916) y, además, las ideas de Estoup

¹ La clasificación fue libremente inspirada por la terminología utilizada por I. P. Pavlov en Biología para clasificar los tipos de reflejos.

² Añadamos que distintas lenguas difieren solamente en diferentes valores de parámetros que vamos a evaluar en relación con las Leyes de Zipf, sin embargo, desde el punto metodológico en todas las lenguas procedemos igual.

fueron desarrolladas por George Kingsley Zipf (1902-1950), quien realmente dio fama a la estadística lingüística.³

Zipf era un lingüista estadounidense, filólogo y profesor en Harvard. Sus estudios se basaban en la hipótesis de la posibilidad de matematizar todo el mundo que nos rodea, lo cual se considera una de las cuestiones filosóficas ya desde la Antigüedad. Hablaba sobre el llamado Principio del mínimo esfuerzo (Principle of least effort), que también se refleja en el comportamiento humano (incluyendo la selección de palabras durante la comunicación). Partiendo de las observaciones de Estoup describió matemáticamente las reglas estadísticas que rigen la distribución del léxico en textos (hoy en día conocidas como las Leyes de Zipf).

4. El principio del mínimo esfuerzo (Principle of least effort)

No sólo en el lenguaje, sino también en otros sistemas naturales se aplican ciertas reglas según las que se comportan inconscientemente los participantes con el objetivo de que ese comportamiento les cueste el menor esfuerzo posible. El principio del mínimo esfuerzo⁴ se basa en dos hechos conocidos, a saber, que el lenguaje se centra en la rentabilidad (como organismos en la teoría de la evolución de Darwin) y la claridad de la comunicación. El hablante trata de expresar sus pensamientos de modo que realice el menor esfuerzo posible, o sea la mayoría de las palabras que utiliza más frecuentemente pertenece a un grupo relativamente pequeño de lexemas. Por el contrario, el receptor intenta comprender lo máximo lo posible, por lo tanto, trata de mantener un vocabulario bastante amplio y de términos específicos para evitar la ambigüedad, lo cual contrarresta las intenciones del mínimo esfuerzo del hablante.

Algunas de las consecuencias conocidas del principio del mínimo esfuerzo las hallamos en las relaciones entre la longitud de la palabra, el número de sus significados y su frecuencia en el texto:

1º: palabras del uso más frecuente son las más cortas (mayormente vacías)

2º: palabras más cortas tienden a tener un mayor número de significados

3º: la frecuencia de las palabras en el texto o corpus crece con el número de sus significados

También las propias Leyes de Zipf se pueden considerar una consecuencia del principio del mínimo esfuerzo.

5. Las Leyes de Zipf

Formulemos ahora tres Leyes de Zipf (de distribución, de significado y de agrupaciones de palabras) y, a continuación, las aplicaremos al texto de la novela *El coronel no tiene quien le escriba*, de Gabriel García Márquez. En el siguiente texto trabajaremos a menudo con la tabla de frecuencia, el término que ahora vamos a explicar. Imaginémonos que en un texto largo calculamos (utilizando el ordenador) para cada palabra cuántas veces aparece en él y luego organizamos todas las palabras del texto en una tabla de la más frecuente a la menos frecuente. En este ordenamiento a cada término se le asigna un número r , que determina su posición en la tabla, y f , que indica el número de ocurrencias de la palabra en el texto. A la hora de crear una tabla de frecuencia de palabras, hay que recordar que bajo el término *palabra* entendemos la parte de texto entre dos espacios consecutivos.

³ *Selected Studies of the Principle of Relative Frequency in Language* (1932).

⁴ *Human behavior and the principle of least effort* (1949), *The Psycho-Biology of Language* (1935).

5.1 La ley de distribución

La ley de distribución constata que el valor de f satisface aproximadamente la fórmula $f = K/r^r$, donde r es un parámetro de distribución (en torno a 1) que expresa el peso del rango en la fórmula y se determina experimentalmente, y K es una constante (también observada experimentalmente) dependiente del tamaño del texto. Mencionemos ahora un ejemplo canónico de la aplicación de la fórmula: Si la primera palabra en la tabla de frecuencia aparece 1.000 veces en el texto, la segunda se producirá aproximadamente 500 veces; la tercera, alrededor de 333 veces, etc. Es decir, el número de ocurrencias de cada palabra es inversamente proporcional a su rango de frecuencia en la tabla (nótese que en este ejemplo sigue siendo constante (aproximadamente 1.000) el producto de la frecuencia y el rango de cada palabra).

Usando el programa de ordenador *Mathematica* hemos obtenido la tabla de frecuencia para la novela *El coronel no tiene quien le escriba*. Citar entera la tabla de frecuencia es problemático, ya que es bastante extensa: la novela consta de 17.364 palabras, incluyendo cifras y el número de palabras diferentes (es decir, sin contar la aparición repetida de una palabra) es de 3.484 palabras, lo cual significa que esta tabla de frecuencia tiene 3.484 líneas. Por esta razón vamos a ilustrar sólo unas pocas líneas de su inicio, medio y fin.

Tabla 2a: Inicio

Palabra	f	r	$f.r$
el	939	1	939
de	842	2	1.664
la	827	3	2.481
en	436	4	1.744
coronel	358	5	1.790
a	352	6	2.212
que	322	7	2.554
se	284	8	2.272
y	275	9	2.475
los	265	10	2.650

Tabla 2b: Parte central

Palabra	f	r	$f.r$	r_{mod}	$f.r_{\text{mod}}^7$
entre	20	88	1.760	90	2.116
manos	20	89	1.780	90	2.116
oficina	20	90	1.800	90	2.116
abogado	20	91	1.820	90	2.116
todo	19	92	1.748	95	2.095
mi	19	93	1.767	95	2.095
dirigió	19	94	1.786	95	2.095
yo	19	95	1.805	95	2.095
salió	19	96	1.824	95	2.095
eso	19	97	1.843	95	2.095
Agustín	18	98	1.764	98	2.108

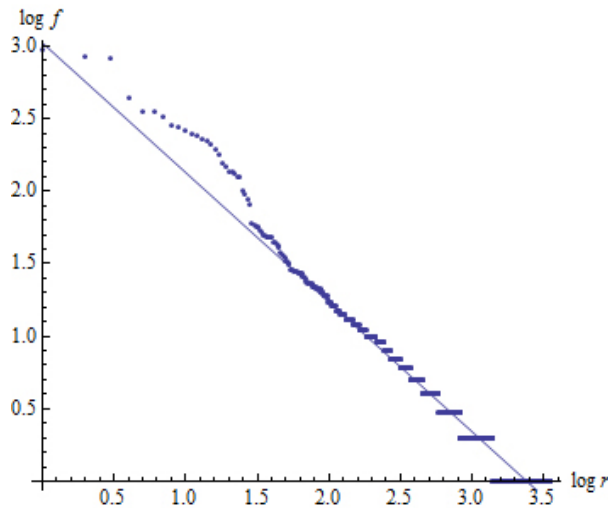
Tabla 2c: Final

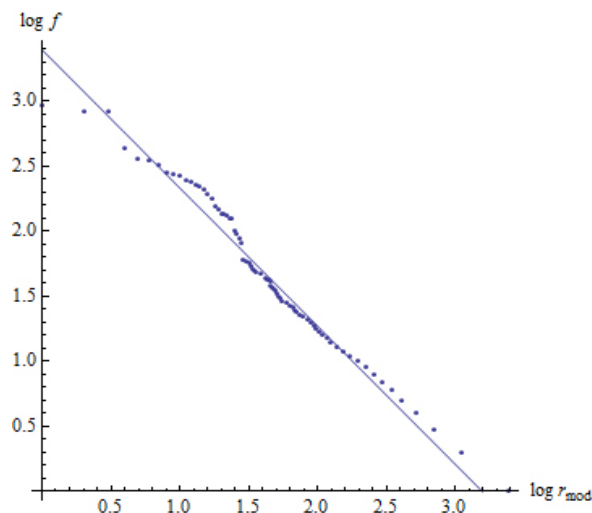
Palabra	f	r	$f \cdot r$	r_{mod}	$f \cdot r_{\text{mod}}^{\gamma}$
puro	1	3.482	3.482	2.442	2.442
explícito	1	3.483	3.483	2.442	2.442
invencible	1	3.484	3.484	2.442	2.442

La parte inicial de la tabla es básica y todos sus elementos ya han sido explicados. A diferencia de ella en la parte central y final se añadieron dos columnas que contribuyen a una forma más racional de tablas de frecuencia. Por ejemplo, las expresiones como *entre*, *manos*, *oficina* y *abogado* se encuentran todas igualmente 20 veces en el texto y el orden en que se citan en la tabla no es importante, sin embargo, se ha generado al azar por el programa. En tal caso no tendría ninguna justificación que el rango de estas palabras fuera diferente, y, por lo tanto, se modificó el orden de palabras para que a todas las palabras con la misma frecuencia se les asignara el mismo rango, a saber, el medio del intervalo de 88 a 92. Esta modificación se aplicó por analogía a todos los demás grupos de palabras de igual frecuencia. A continuación, usando el programa *Mathematica* se estimó el parámetro γ (matemáticamente la estima se realizó mediante el método de los mínimos cuadrados) y así se hizo el cálculo del producto $f \cdot r_{\text{mod}}^{\gamma}$ más preciso, lo cual puede proporcionarnos una mejor idea sobre el valor de la constante K .

Dada la extensión de la tabla de frecuencia es habitual y útil pasar a una representación gráfica, en la que se pueden observar todas las palabras del texto con facilidad. Si, por motivos de conseguir la representación gráfica más corriente, logaritmamos los dos lados de la fórmula $f = K/r^{\gamma}$, obtendremos $\log f = \log K - \gamma \log r$, o sea el logaritmo de la frecuencia depende de forma lineal del logaritmo del rango, lo cual se representa gráficamente más o menos como una línea recta hacia abajo. La ley de Zipf describe cuánto regularmente se colocan los puntos que representan las palabras del texto alrededor de la línea indicada por la fórmula. Si conocemos la ecuación del ajuste, cualquier palabra puede ser utilizada para determinar con qué frecuencia aproximadamente aparece en el texto sabiendo sólo su rango en la tabla de frecuencia (o viceversa).

Gráfico 1: La representación gráfica de la ley de distribución





Cada punto en los gráficos anteriores representa una sola palabra o grupo de palabras con la misma frecuencia de aparición en el texto. Además, también hemos marcado la línea de tendencia lineal⁵ con parámetros $K = 1052$, $\gamma = 0,8935$ en la imagen izquierda y $K = 2219$, $\gamma = 1,224$ en la derecha. La imagen izquierda muestra la situación sin la modificación del orden, la derecha, con las modificaciones. El mejor ajuste de los datos (cuando la mayoría de los puntos está en torno a la línea marcada) se da en la parte central de la escala de frecuencia. En el extremo izquierdo, por otro lado, se muestran una pocas palabras, sobre todo, gramaticales que son muy frecuentes en el texto y causan mayores desviaciones de la línea recta. En el extremo derecho se encuentra un gran número de términos a menudo muy especializados como topónimos y otros, que aparecen en el texto una sola vez (el llamado *hapax legomenon*), dos veces, etc. En esta parte de la escala de frecuencia, hemos mejorado el ajuste mediante la modificación del rango de palabras.

Si queremos mejorar el ajuste de los datos en global, será necesario excluir del análisis de frecuencia algunas expresiones del inicio de la tabla (de la parte izquierda de la escala de frecuencia), donde la validez de la ley de distribución se ve lo más negativamente afectada. De la parte izquierda de la gama de frecuencia se pueden excluir o palabras específicas del texto como los nombres de los protagonistas de la novela (por ejemplo, *Coronel*, *mujer*, etc.) que aparecen en el texto notablemente más que en otros textos similares, o palabras gramaticales como los determinantes, preposiciones, conjunciones, etc. Vamos a ocuparnos del efecto adicional de la exclusión de ciertas expresiones gramaticales. Para el español, e incluso otros idiomas no romances es típico el uso de determinantes en combinación con otras palabras variables. Los determinantes por la definición de la palabra entran en nuestro análisis, pero el hablante no los elige por querer, sino porque los debe utilizar por razones gramaticales con nombres, etc. Por lo tanto, los determinantes no se someterán al proceso deliberado de la selección de lexemas y como su uso está condicionado por otros términos, podría resultar racional excluirlos del análisis en el marco del aspecto incondicional de la comunicación. En nuestro caso excluimos del

⁵ Se trata de la línea que presenta la menor suma posible de los cuadrados de distancias de todos los puntos (el llamado método de los mínimos cuadrados).

análisis los determinantes *el, la, los, las, un y una* y cambiamos las tablas de frecuencia correspondientes. La representación gráfica sería similar a las que ya hemos mencionado (es decir, que esta vez la saltamos), pero se han cambiado los parámetros: $K = 868$, $\gamma = 0,8682$ y con la modificación del orden, $K = 1345$, $\gamma = 0,9391$.

La precisión del ajuste de los puntos (y por lo tanto el grado en que cualquier palabra del texto satisface la ley de distribución) la vamos a determinar mediante el error cuadrado medio (MSE).⁶ Si todas las palabras cumplieran plenamente con la ley de distribución, el error cuadrado medio tendría que ser cero y todos los puntos se encontrarían en la línea recta, lo cual prácticamente nunca ocurre. Nuestro objetivo, sin embargo, fue que el error cuadrado medio se redujera al mínimo y por la posible exclusión de ciertos términos disminuyera aún más, por si era posible. Observemos el efecto en dos situaciones y sus combinaciones: en primer lugar, la exclusión de los determinantes y en segundo lugar, la modificación del orden:

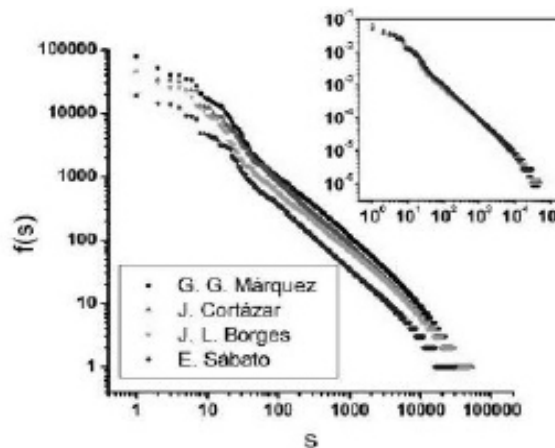
Tabla 3: El efecto de la modificación de orden y la exclusión de determinantes

MSE	sin la modificación del orden	con la modificación del orden
sin la exclusión de determinantes	0,0074	0,0092
con la exclusión de determinantes	0,0069	0,0057

Podemos concluir que la exclusión de determinantes significó un mejoramiento del ajuste de los datos que en combinación con la modificación del orden se hizo aún más importante. Por otra parte, la modificación (del orden) misma mejoró el ajuste en la parte derecha, sin embargo, en global sin la exclusión de determinantes el ajuste se empeoró.

Hemos aquí, por curiosidad, las curvas de distribución normalizadas en cuanto al tamaño del texto de cuatro escritores de América Latina tomado de (Calderón, F.; Curilef, S.; Ladrón de Guevara, M. L. 2009).

Gráfico 2: La ley de distribución en algunos escritores latinoamericanos



⁶ Se trata de la media de cuadrados de distancias de todos los puntos de la línea del ajuste.

5.2 La ley de significado

La ley de significado establece que el número de significados de la palabra (aquí sustituido por m) satisface aproximadamente la fórmula $m = A\sqrt{f} = B/\sqrt{r}$, donde A y B son constantes determinadas experimentalmente en función del tamaño del texto. Así que si en algún texto $A = 0.1$, $B = 10$ y la primera palabra más frecuente ($r = 1$) se produce 10.000 veces, entonces esta palabra conforme con la fórmula anterior debe tener 10 significados.

Para aplicar la ley al texto de la novela de Márquez, primero debemos determinar el número de significados m (valor entrante) para un número suficientemente alto de palabras, por ejemplo, del diccionario. Como para estas palabras conocemos ambos valores f y r , siempre podemos calcular el valor de A y B para cada palabra elegida del texto, a continuación vamos a considerar un sólo valor de A y B (el promedio de todos los valores calculados). Sabiendo las constantes A , B , somos capaces de calcular m para cualquier palabra de la tabla de frecuencia utilizando sólo sus valores de f o r .

En el proceso que acabamos de describir se plantea la cuestión de cómo determinar objetivamente los valores entrantes de m . Es decir, la misma palabra puede tener un número diferente de significados en distintos diccionarios especializados, e incluso en el mismo diccionario las acepciones no tienen que mencionarse de manera consistente, lo cual lleva a imprecisiones. Una posible solución (aunque no la ideal) es elegir un diccionario de referencia (en nuestro caso el DRAE) y tener en cuenta las inexactitudes en la interpretación de los resultados.

La tabla siguiente se compiló a partir de las 15 palabras seleccionadas con diferentes valores de m . Su función es sólo para ver una vía posible de la aplicación de la ley de significado, es decir que para hacer verdaderas y generalmente válidas conclusiones, las palabras elegidas tendrían que seleccionarse por métodos estadísticos y no ad hoc como en nuestro ejemplo. La conclusión que podemos hacer a base de esta tabla es que las palabras de la novela deberían satisfacer aproximadamente $m = 109\sqrt{f} = 4652/\sqrt{r}$.

Tabla 4: La ley de significado

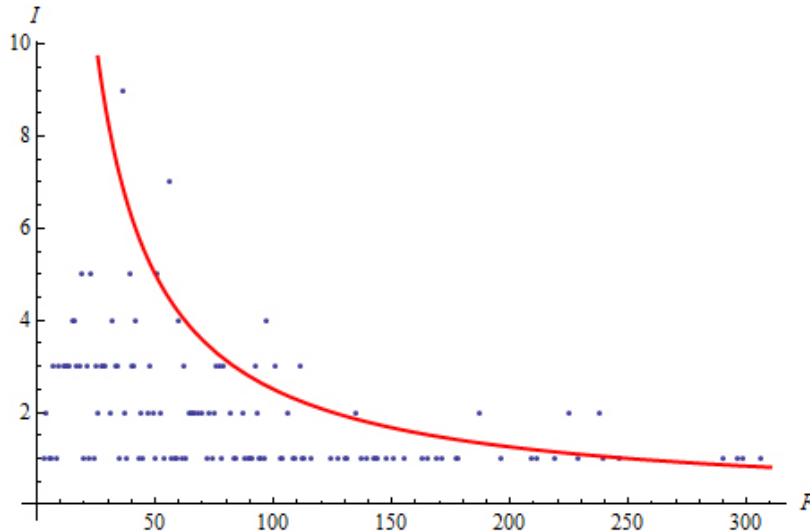
Palabra	m	f	r	A	B
de	27	842	2	0,93	38,18
que	25	322	7	1,39	66,14
para	13	136	20	1,11	58,14
cuando	7	81	28	0,78	37,04
puerta	8	29	54	1,49	58,79
tarde	4	28	57	0,76	30,20
ahora	6	23	71	1,25	50,56
pueblo	5	22	77	1,07	43,78
oficina	5	20	88	1,12	46,90
mañana	6	17	99	1,46	59,70
ventana	4	12	148	1,15	48,66
maíz	2	10	183	0,63	27,06
hambre	3	8	243	1,06	46,77
envejecido	2	3	585	1,15	48,37
pelea	1	1	1.401	1,00	37,43
Media				1,09	46,52

5.3 La ley de agrupaciones de palabras

Llamemos al intervalo todas las palabras que están entre dos apariciones consecutivas de una palabra conceptual que elegimos antes de hacer el análisis. La longitud del intervalo se define como el número de palabras que contiene el intervalo y la designamos como I . Sea f el número de intervalos de la misma longitud I , después aproximadamente $F = C/I$, donde C es la constante de proporcionalidad en función del texto. Se puede constatar que cuánto más corto el intervalo es, tanto más frecuente en el texto o corpus es. Esto prácticamente significa que las palabras conceptuales tienden a agruparse (es decir que en el texto unos se encuentran relativamente cerca de otros).

En la novela de Márquez observamos las ocurrencias del término *dijo* y todos los intervalos que éstas apariciones delimitan en la novela. En el texto la palabra *dijo* aparece 250 veces, el intervalo más corto tiene 3 palabras de largo y se produce sólo una vez en el texto, mientras que el intervalo más frecuente es de longitud 36 y aparece 9 veces en el texto. Para poder representar todas las palabras procedemos, como anteriormente, a los gráficos. Todos los intervalos se representan en forma de los puntos redondos en el gráfico siguiente, luego también indicamos la curva de tendencia (una rama de hipérbola) resultante de la ley de agrupaciones de palabras. En esta imagen determinamos $C = 250$.

Gráfico 3: La ley de agrupaciones de palabras



La validez de esta afirmación no es muy visible en el texto analizado por su insuficiente longitud que causa más variación estadística. Las agrupaciones recién descritas se reflejarían mejor en un corpus más coherente del lenguaje.

6. Conclusión

Las Leyes de Zipf son reglas que describen la distribución estadística de palabras en textos y algunas propiedades relacionadas (por ejemplo, el número de significados de la palabra, las agrupaciones de palabras, etc.). Están demostradas en todas las lenguas

naturales, pero también se aplican a otros sistemas naturales. En el análisis de la novela *El coronel no tiene quien le escriba* averiguamos que el texto cumple muy bien la ley de distribución, mientras que la ley de agrupaciones de palabras muestra mayores errores estadísticos. Como se trata de leyes estadísticas, su validez mejoraría gradualmente con la longitud creciente del texto. Sin embargo, la ley de distribución es la más frecuentemente mencionada, porque su cumplimiento ya es evidente en textos más cortos.

Si consideramos la importancia práctica de las Leyes de Zipf, el análisis de frecuencia en textos y la evaluación de los parámetros utilizados en las leyes pueden resultar como una herramienta muy útil para clasificar el texto o compararlo con otros según diversos aspectos como el autor (las curvas de distribución de varios autores son diferentes), el estilo, el contenido, la claridad o el público al que está destinado (por ejemplo, en los textos corrientes la relación entre el número de palabras diferentes y la extensión del texto suele ser 1÷6, por el contrario, en los libros infantiles esa relación es menor).

Bibliografía

- ABNEY, Steve (1996), «Statistical methods and linguistics», in: KLAVANS, J. – RESNIK, P. (editors), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, The MIT Press.
- CALDERÓN, F. – CURILEF, S. – LADRÓN DE GUEVARA, M. L. (2009), «Probability distribution in a quantitative linguistic problem», *Brazilian Journal of Physics*, vol. 39, no. 2, São Paulo. Departamento de Física, Universidad Católica del Norte, Antofagasta, Chile. [online] [cit. 2011-01-20]. Disponible en: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-97332009000400028
- PETŘÍK, Daniel (2010), *Sémanticko-stylistická perspektiva při volbě lexikálních prostředků ve španělštině*, Diplomová práce ÚRS FF UK.
- ROBERTS, Carl (1989), «Other Than Counting Words: A Linguistic Approach to Content Analysis», *Social Forces*, Vol. 68, No. 1, s. 147-177.
- ZIPF, George Kingsley (1949), *Human behavior and the Principle of least effort*, Addison-Wesley Press.
- ZIPF, George Kingsley (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Harward University Press.
- ZIPF, George Kingsley (1941), «Semantic Frequency List», *American Speech*, Vol. 16, No. 1, s. 43-45.
- ZIPF, George Kingsley (1937). «Statistical Methods and Dynamic Philology», *Language*, Vol. 13, No. 1, s. 60-70.
- ZIPF, George Kingsley (1965), *The Psycho-Biology of Language*, The MIT Press.

Daniel Petřík
Ústav románských studií
Filozofická fakulta
Univerzita Karlova v Praze
Náměstí Jana Palacha 2
110 00 Praha 1
República Checa
dapet@email.cz