Estadística en *Don Quijote*

(4) Ley de Pareto

Hiroto Ueda (Universidad de Tokio)
Antonio Moreno (Universidad Autónoma de Madrid)

Al dar un paseo por el campo, nos encontramos con distintos aspectos de la naturaleza. Observamos que todas las flores (Foto 1: crisantemos silvestres, que florecen a principios de otoño) y también todas las bellotas (Foto 2: frutos de los árboles de la familia del haya) poseen un tamaño similar, lo que probablemente sigue un patrón biológico estable. ¿A qué se debe este parecido en el tamaño de las flores y en el de los frutos de las plantas? Según las inteligencias artificiales (usamos Google Gemini), en esencia, la planta ha evolucionado para producir una flor "lo suficientemente grande" como para ser polinizada. Posteriormente, esa misma estructura floral inicial se transforma en un fruto que es "lo suficientemente grande" como para dispersar las semillas de manera eficaz, todo ello bajo las mismas limitaciones de energía y recursos de la planta. Esto da lugar a la percepción de tamaños similares o coherentes entre las flores y los frutos maduros.


Foto-1: Crisantemos silvestres, Foto-2: Bellotas.

Al observar los objetos inorgánicos, como las piedras que hay en un río (Foto-1, Foto-2), esta vez se observa una gran diferencia de tamaño. Es más. Las piedras grandes pueden contarse con los dedos, mientras que el número de piedras pequeñas es inabarcable. ¿Por qué es así? La respuesta

que dan las inteligencias artificiales son las siguientes:

> La gran disparidad entre la escasez de piedras grandes y el inmenso número de piedras pequeñas se debe a los procesos geológicos fundamentales de meteorización y erosión. Es un principio básico de la naturaleza: las rocas grandes se rompen en muchas rocas menores. (...) La escasez de piedras grandes y la abundancia de fragmentos pequeños son el resultado inevitable de un ciclo continuo.


Foto-3: Piedras en un río, Foto-4: Piedras pequeñas en un sendero.

Esta desigualdad de frecuencias se presenta no solo en la naturaleza, sino también en la sociedad: por ejemplo, en la renta per cápita (con una pequeña cantidad de personas de altos ingresos), en las poblaciones de las ciudades o en la venta de libros, entre otros. También en el mundo de los seres vivos, existe una relación inversa bien establecida en ecología entre el tamaño del organismo y su abundancia poblacional (el número de individuos). En estos casos, se observa una gran diferencia: grandes números en pocos miembros (en el extremo superior de la curva) (Koch 1997) y pequeños números en muchos miembros (en el extremo inferior muy largo) (Anderson 2006).

La misma situación que ocurre con las piedras (grandes y pequeñas) en un río y en los senderos, se presenta en las palabras dentro del gran río de la literatura universal: el *Quijote*. Nuestro objetivo es determinar la correlación entre la extensión física de las palabras (medida por el número de fonemas) y su frecuencia de uso en la obra.

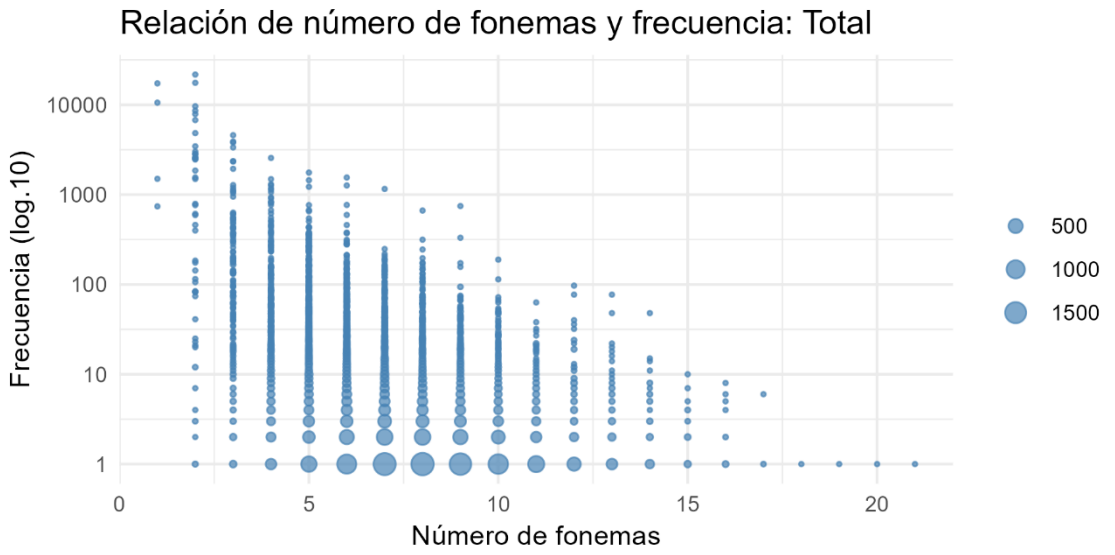
Nuestra pregunta inicial de estudio es: ¿Son frecuentes las palabras pequeñas (cortas, de pocos fonemas) y, a la inversa, se utilizan en pocas

ocasiones las palabras grandes (extensas, de numerosos fonemas)? Las inteligencias artificiales nos dan la respuesta inmediatamente:

> Sí, la observación de que las palabras cortas (con pocos fonemas o sílabas) son más frecuentes y las palabras largas (con numerosos fonemas) se usan menos frecuentemente es un fenómeno lingüístico bien documentado y universal en muchos idiomas. El hablante busca la máxima eficiencia. Es más fácil y rápido pronunciar palabras cortas que largas. Por lo tanto, las ideas o conceptos que se necesitan comunicar con más frecuencia tienden a evolucionar hacia formas más cortas para minimizar el esfuerzo físico y temporal. En este sentido, la Ley de Abreviación es importante: es decir, existe una relación inversa entre la frecuencia de uso de una palabra y su longitud. Las palabras de alta frecuencia tienden a ser cortas, mientras que las de baja frecuencia pueden permitirse ser más largas. (Zipf. 1936: 30-39).

Sin embargo, a partir de nuestros estudios anteriores, nos inclinamos a cuestionar la validez de la Ley de Abreviación. A continuación, presentamos los resultados de nuestros cálculos basados en esta obra maestra.

En primer lugar, analicemos la relación entre la longitud de las palabras (expresada como el número de fonemas) y sus frecuencias de uso de la totalidad de la obra:



Según este gráfico, es cierto que las frecuencias de las formas largas

de más de 18 fonemas son mínimas. Concretamente, se tratan de: *desenbarazadamente* (fonemas: 18), *estraordinariamente* (19), *extraordinariamente* (20) y *bienintencionadamente* (fonemas: 21). Son palabras largas y, por ello, aparentemente serían poco frecuentes.

Sin embargo, las palabras humanas no se desgastan como los cantos rodados de un río. En general, las palabras aumentan su extensión por la derivación, por ejemplo: *orden* > *ordinario* > *extraordinario* > *extraordinariamente*. Casi nunca disminuyen para formar nuevas palabras más cortas, a excepción de la formación abreviada: *fotografía* > *foto, motocicleta* > *moto,* etc. que son relativamente pocas. Por ello, las palabras no se desgastan con el uso frecuente, como si fueran zapatos viejos. Se crean con prefijos y sufijos según la necesidad de expresión y de comunicación.

En este sentido, el hecho de que las palabras frecuentes sean cortas no se debería al "acortamiento" de estas, sino, por el contrario, al "aumento" de la longitud de las palabras formadas por necesidades lingüísticas y expresivas. Dado que las palabras nuevas están marcadas por la adición de elementos especiales, su uso es limitado. Lo especial implica lo raro, lo poco frecuente.

Al volver a observar el gráfico anterior, vemos una distribución de frecuencias casi triangular, muy peculiar. Si se presentara una concentración alrededor de la línea recta descendente, podríamos confirmar la correlación inversa entre el número de fonemas (longitud de las palabras) y su frecuencia de uso, pero la realidad es muy distinta. Efectivamente, según el gráfico anterior, hay numerosos casos de palabras cortas con una frecuencia reducida. Por lo tanto, en el texto del *Quijote*, debemos abordar la Ley de Abreviación con cautela.

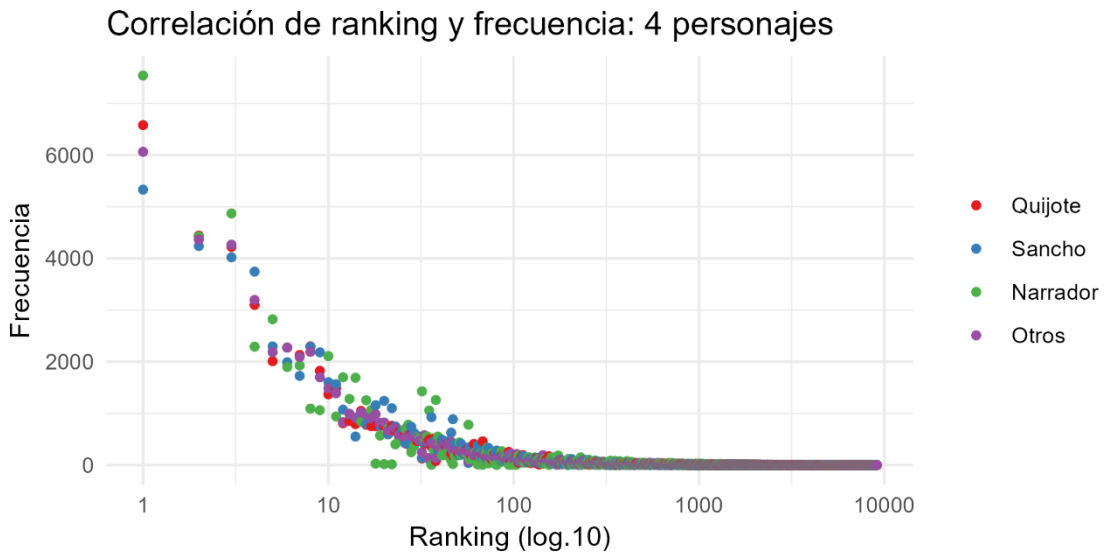
La ausencia de correlación entre el número de fonemas y la frecuencia de uso se comprueba en el siguiente gráfico de índices de correlación [−1, 1]:

Correlación de frecuencias de palabras entre personaje, suma y fonema

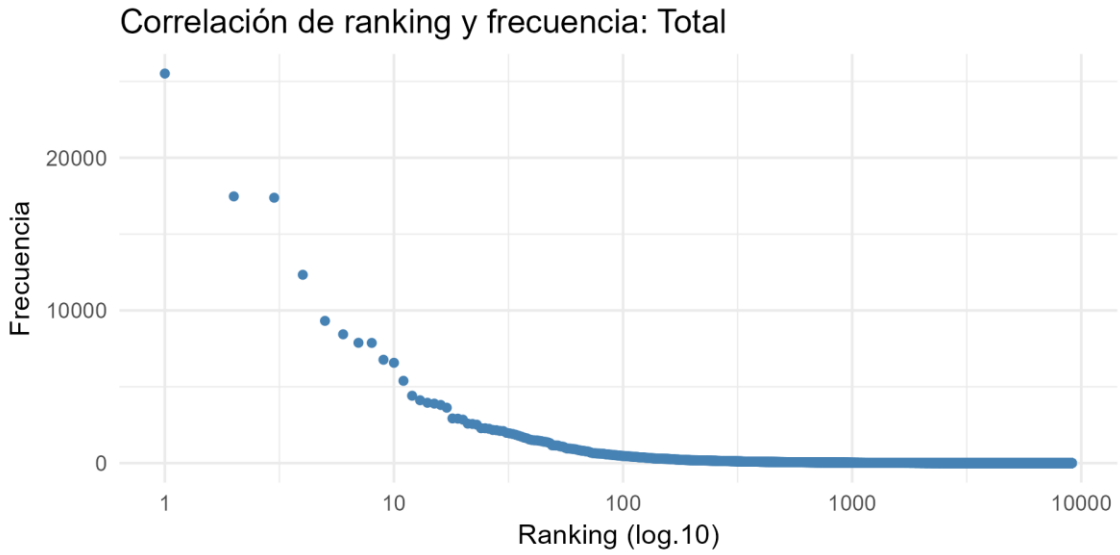
	Quijote	Sancho	Narrador	Otros	Suma	Fonema
Quijote	1.000	0.983	0.928	0.995	0.994	-0.099
Sancho	0.983	1.000	0.895	0.987	0.983	-0.101
Narrador	0.928	0.895	1.000	0.936	0.957	-0.093
Otros	0.995	0.987	0.936	1.000	0.997	-0.099
Suma	0.994	0.983	0.957	0.997	1.000	-0.100
Fonema	-0.099	-0.101	-0.093	-0.099	-0.100	1.000

En general, la frecuencia de las palabras se caracteriza por la ley de Pareto y, en su versión lingüística, por la ley de Zipf: pocas palabras de alta frecuencia ocupan la mayor parte de la frecuencia total. Esta distribución desigual se conoce como el Efecto Mateo. Según nuestra consulta a la inteligencia artificial, se trata de un fenómeno sociológico y psicológico que describe un patrón de ventaja acumulativa para quienes ya poseen riqueza, prestigio o éxito, y una desventaja acumulativa para quienes no los tienen. Se resume en la frase: "Al que tiene se le dará más, y tendrá en abundancia; pero al que no tiene, aun lo que tiene le será quitado" (*Evangelio de Mateo*, 13:12 y 25:29).

Veamos la realidad estadística de este efecto en nuestro corpus lingüístico del *Quijote:*

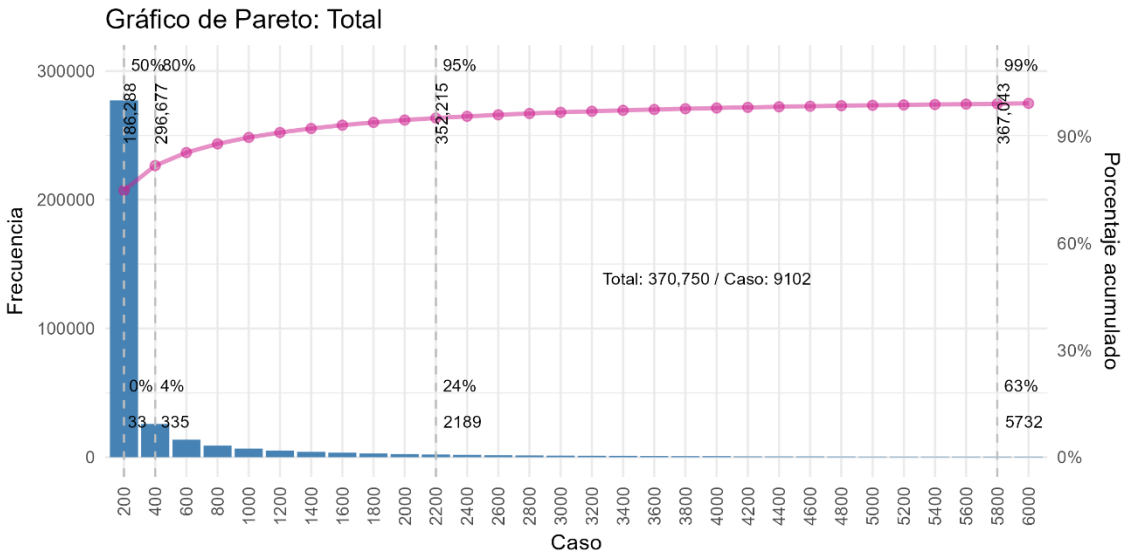


Efectivamente, se observa que en los cuatro personajes (Quijote, Sancho, Narrador, Otros) las primeras 100 palabras más frecuentes representan la mayor parte de la frecuencia total. Parece ser una ley general que se aplica sin distinción de personajes. Por ello, de aquí en adelante, nos enfocaremos únicamente en la totalidad de personajes, cuya distribución se presenta de la siguiente manera:



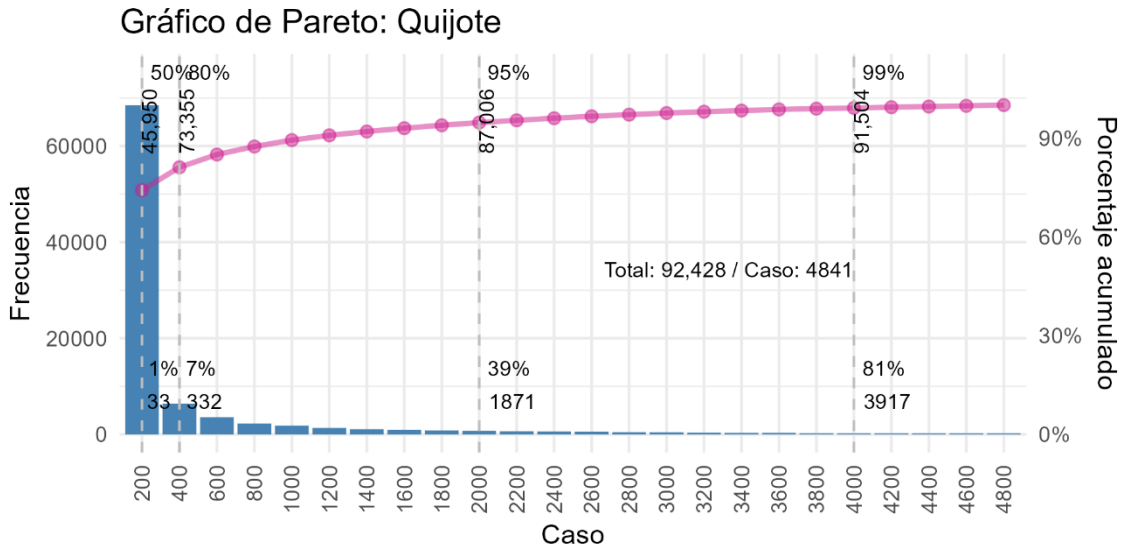
En este gráfico, de nuevo, se confirma que las primeras 100 palabras más frecuentes ocupan la mayor parte de la frecuencia total. Precisamente, se trata del "extremo o cabeza alta" (anterior al rango R=100), seguido de la "cola inferior muy larga" (a partir de R=100). No obstante, para conocer con mayor detalle las características de la concentración en la primera parte,

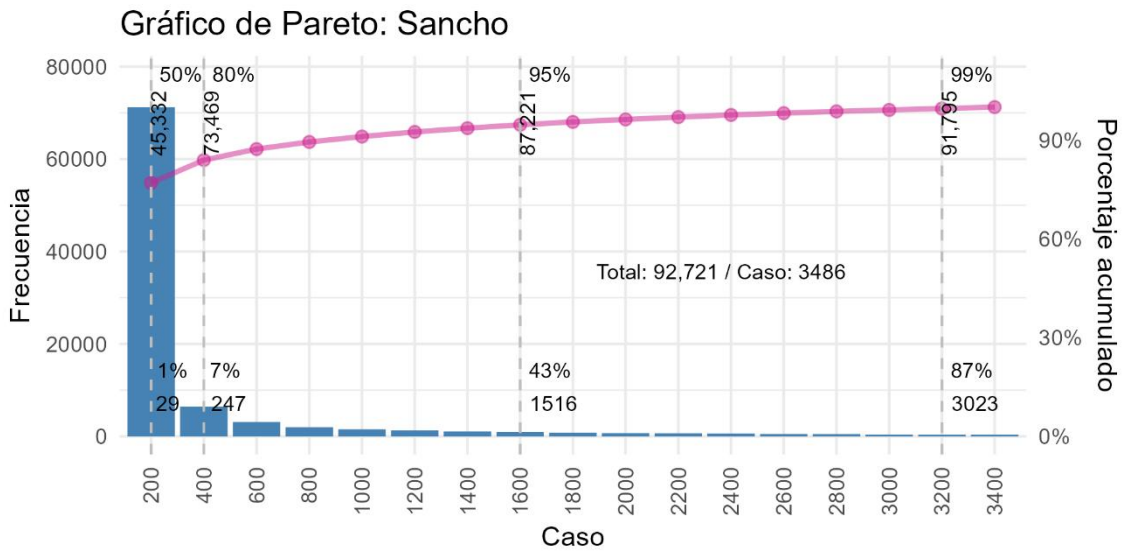
conviene recurrir al siguiente gráfico de Pareto.



En este gráfico, constituido por barras que representan las sumas de frecuencias de cada grupo de casos (palabras: 200, 400, 600...) y una curva que representa los porcentajes acumulados, se observa que solo las primeras 33 palabras más frecuentes, que no alcanzan el 1% del total de casos, ocupan la mitad de la frecuencia total de uso (50%: 186.288). Además, las primeras 335 palabras (4% del total) ocupan el 80% de la totalidad. La distribución es sumamente desigual.

Ahora bien, comparemos las palabras de los dos protagonistas:





Nos concentramos ahora en el umbral del 99% de vocabulario:

Quijote: 3.917 palabras (81% del vocabulario total de Quijote).
Sancho: 3.023 palabras (87% del vocabulario total de Sancho119900).

Esto significa que, para alcanzar el 99% del vocabulario del Quijote, es necesario conocer casi 4.000 palabras (3.917), mientras que en el caso de Sancho se requieren alrededor de 3.000 palabras (3.023). La diferencia es significativa. Parece que Sancho Panza repite las mismas palabras más que Don Quijote, quien, en contraste, varía más su vocabulario.

Referencia:

Anderson, Chris. 2006. *The Long Tail. Why the Future of Business is Selling Less of More*. New York. Hyperion.

Zipf, George Kingsley. 1936. *The Psycho-biology of language. An Introduction to Dynamic Philology*. London. Rouledge.

Koch, Richard. 1997. *The 80/20 Principle. The Secret of Achieving More with Less.* London. Nicholas Brealey Publishing.