MAT CON
MATEMÁTICA CONECTADA

Matemática y lingüística: La ley de Zipf

fcfm FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE
CMMEdu Laboratorio de Educación Centro de Modelamiento Matemático
UNESCO Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura
unitwin Cátedra UNESCO Formación de Docentes para Enseñar Matemática en el Siglo XXI Universidad de Chile
Ministerio de Educación Gobierno de Chile
PLAN DE REACTIVACIÓN EDUCATIVA

Revisemos el <u>video</u> de esta situación:

"¿Cuáles son las palabras que más utilizamos en el español?"

*Imagen referencial de la situación

A partir del video, respondamos:

- ¿Qué significa que la palabra “que” sea la tercera en el ranking?
- ¿Cómo se determinó la posición en el ranking de una palabra?

REAL ACADEMIA
ESPAÑOLA

Presentación del problema

La siguiente tabla muestra las doce primeras palabras de la Base de datos de la RAE:

¿Será posible estimar la frecuencia de una palabra dada su posición en el ranking?

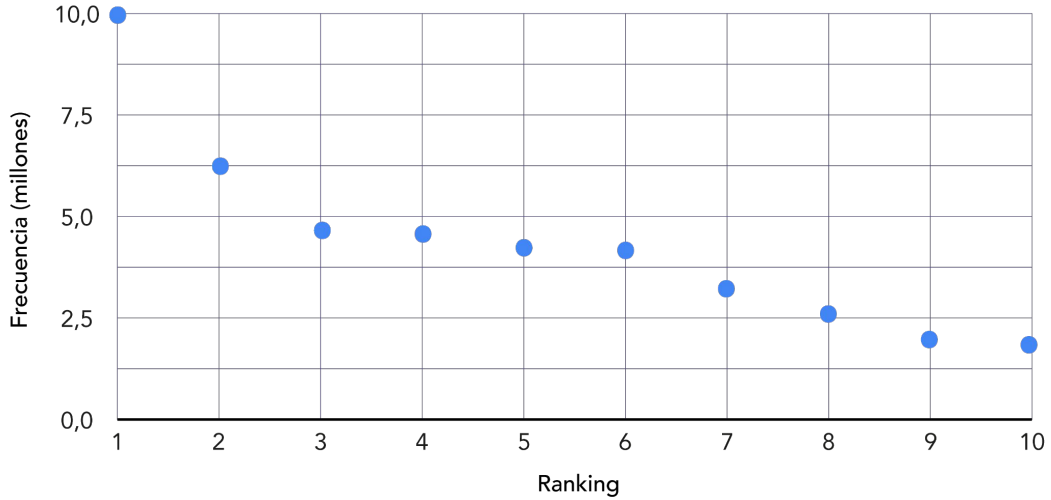
Base de datos de la Real academia Española		
Ranking (n)	**Palabra**	**Frecuencia absoluta (f)**
1	de	9999518
2	la	6277560
3	que	4681839
4	el	4569652
5	en	4234281
6	y	4180279
7	a	3260939
8	los	2618657
9	se	2022514
10	del	1857225
11	las	1686741
12	un	1659827

Actividad 1

1. **Grafica la frecuencia de las primeras 10 palabras y luego responde la siguiente pregunta: ¿Qué se puede observar respecto a la forma en que disminuye la frecuencia de las palabras a medida que se avanza en el ranking?**

Actividad 1

1. **Grafica la frecuencia de las primeras 10 palabras y luego responde la siguiente pregunta: ¿Qué se puede observar respecto a la forma en que disminuye la frecuencia de las palabras a medida que se avanza en el ranking?**



Actividad 1

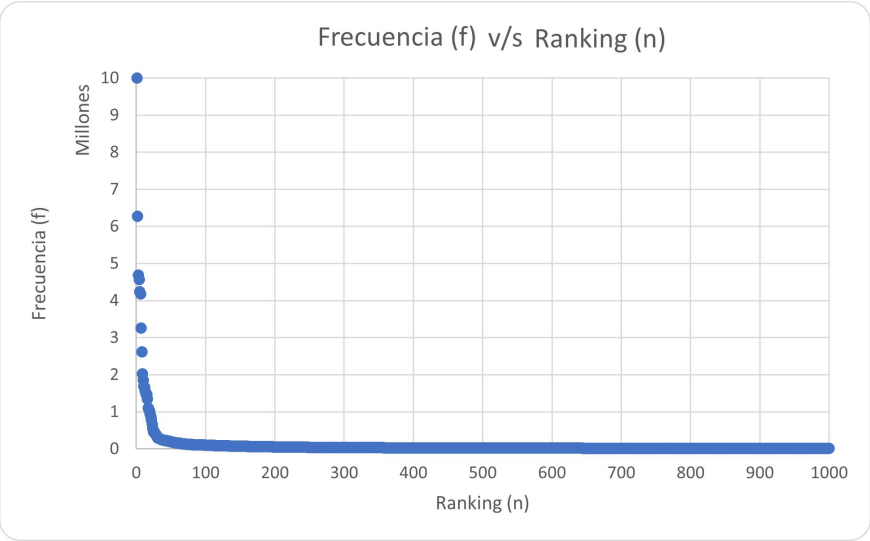
2. Responde las siguientes preguntas:

a) ¿Cómo varía la frecuencia a medida que se avanza en el ranking?

Actividad 1

2. Responde las siguientes preguntas:

a) ¿Cómo varía la frecuencia a medida que se avanza en el ranking?



Las frecuencias disminuyen muy rápidamente en relación a las primeras del ranking.

Actividad 1

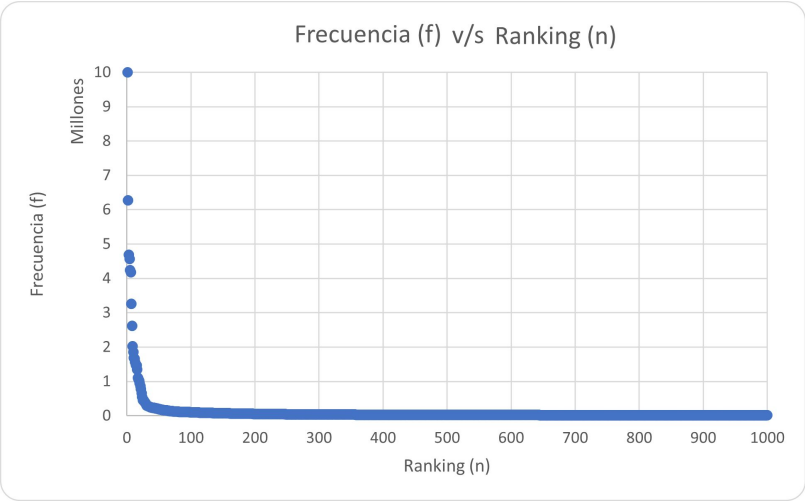
2. Responde las siguientes preguntas:

b) A partir del gráfico, ¿es posible apreciar la frecuencia de la palabra que está en la segunda posición del ranking? ¿Y en la posición 200?

Actividad 1

2. Responde las siguientes preguntas:

b) A partir del gráfico, ¿es posible apreciar la frecuencia de la palabra que está en la segunda posición del ranking? ¿Y en la posición 200?



En el gráfico se puede distinguir la frecuencia de las primeras palabras, pero no así palabras que están más adelante en el ranking.

Actividad 1

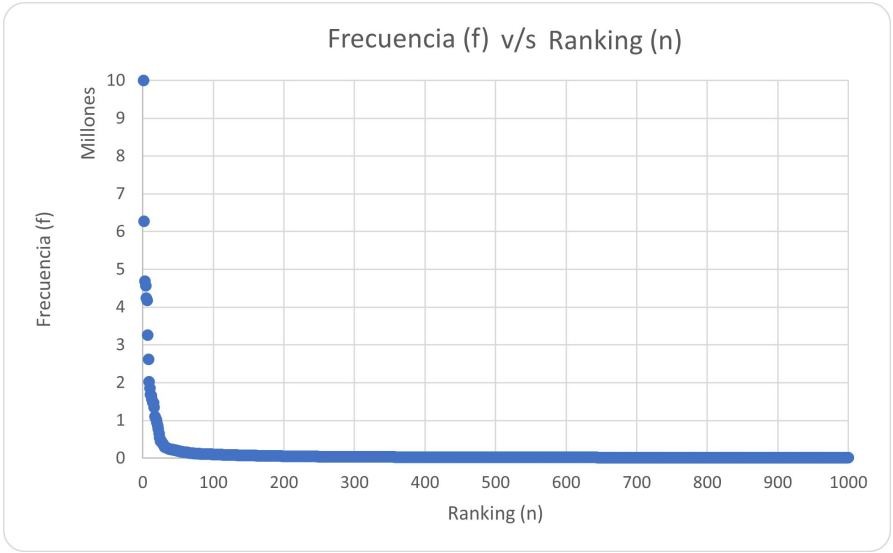
2. Responde las siguientes preguntas:

c) ¿Qué dificultad presenta este gráfico para visualizar los datos?

Actividad 1

2. Responde las siguientes preguntas:

c) ¿Qué dificultad presenta este gráfico para visualizar los datos?



La escala de la frecuencia está expresada en millones, lo que dificulta la visualización de las frecuencias cuando éstas son menores que una centena de mil, lo que ocurre, aproximadamente, para $n \geq 50$.

Actividad 1

3. Completa la tabla con el logaritmo de la frecuencia, aproximado los valores a dos cifras decimales, y luego responde:

¿Cómo cambia la variable frecuencia f al aplicar logaritmo?

Ranking (n)	Palabra	Frecuencia (f)	log (f)
3	que	4681839	
18	para	1062152	
142	trabajo	77478	
999	premio	13701	

Actividad 1

3. Completa la tabla con el logaritmo de la frecuencia, aproximado los valores a dos cifras decimales, y luego responde:

¿Cómo cambia la variable frecuencia f al aplicar logaritmo?

Ranking (n)	Palabra	Frecuencia (f)	log (f)
3	que	4681839	6,67
18	para	1062152	6,03
142	trabajo	77478	4,89
999	premio	13701	4,14

Actividad 1

3. Completa la tabla con el logaritmo de la frecuencia, aproximado los valores a dos cifras decimales, y luego responde:

¿Cómo cambia la variable frecuencia f al aplicar logaritmo?

Ranking (n)	Palabra	Frecuencia (f)	log (f)
3	que	4681839	6,67
18	para	1062152	6,03
142	trabajo	77478	4,89
999	premio	13701	4,14

- Los valores de log(f) son menores de los que la frecuencia f.
- La variación entre los valores de log(f) es mucho más acotada que la de la frecuencia.

Los valores log(f) en esta escala logarítmica indican entre qué potencias de 10 se encuentra la frecuencia f.

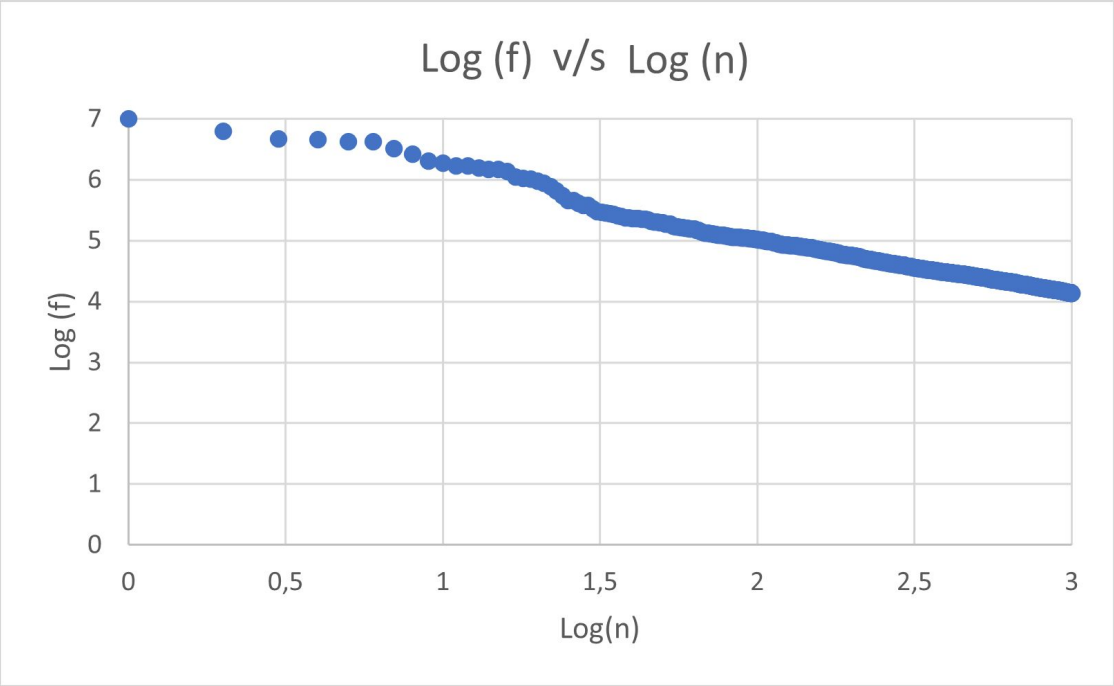
Palabra	Frecuencia (f)	log(f)
para	1 062 152	**6,03**
trabajo	77 478	**4,89**

Frecuencia de la palabra **para**

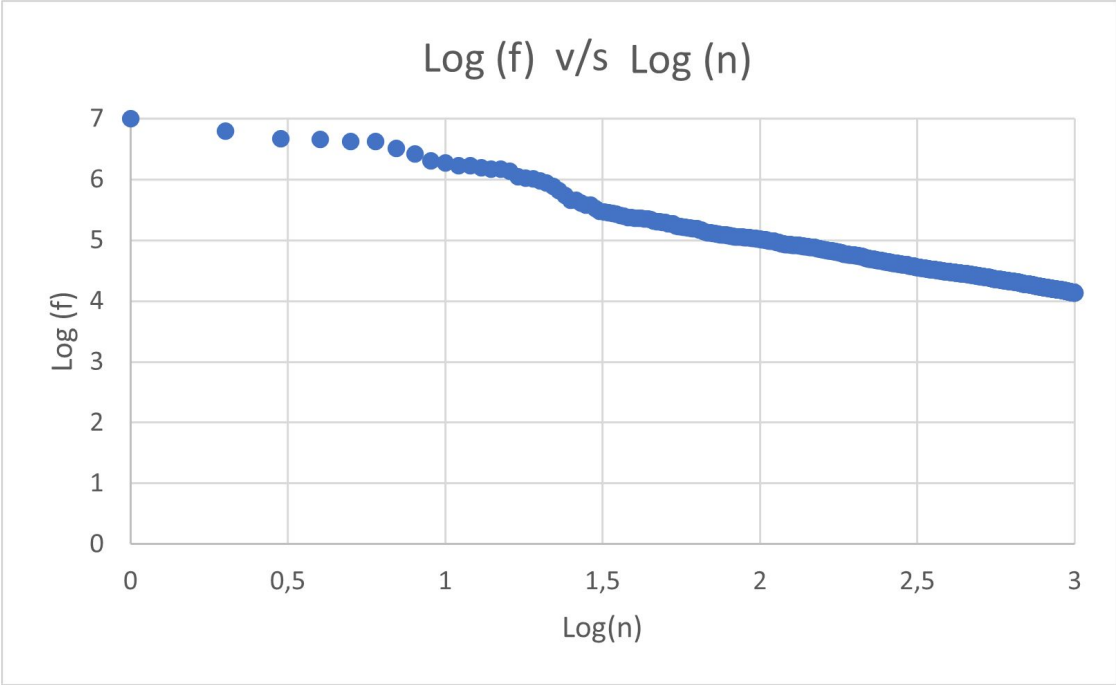
Aproximadamente $10^6 \approx 1\,000\,000$

Frecuencia de la palabra **trabajo**

Entre $10^4 = 10\,000$ y $10^5 = 100\,000$



A partir de este nuevo gráfico, ¿es posible estimar la frecuencia de la palabra número 200?



A partir de este nuevo gráfico, ¿es posible estimar la frecuencia de la palabra número 200?

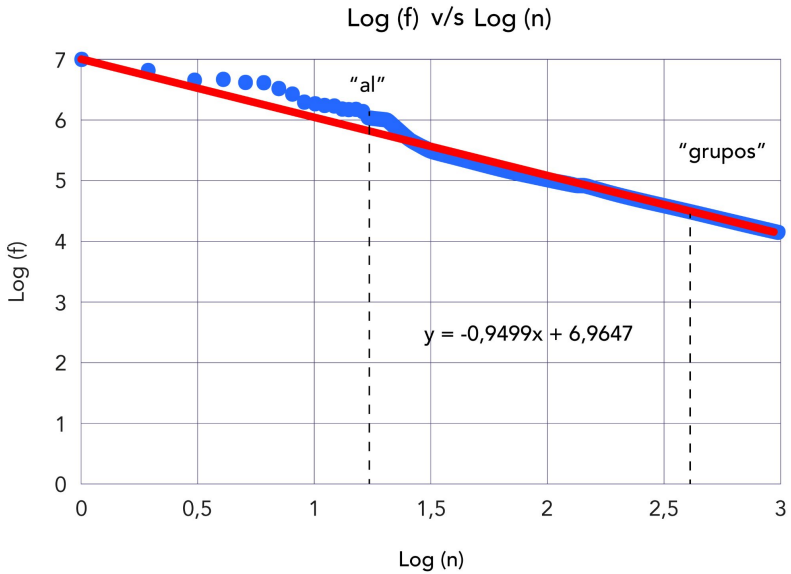
Se puede estimar que la palabra 200 está entre 10^4 = 10 000 y 10^5 = 100 000.

Actividad 1

4. ¿Qué tipo de función podría modelar el gráfico log(f) versus log(n)?

Actividad 1

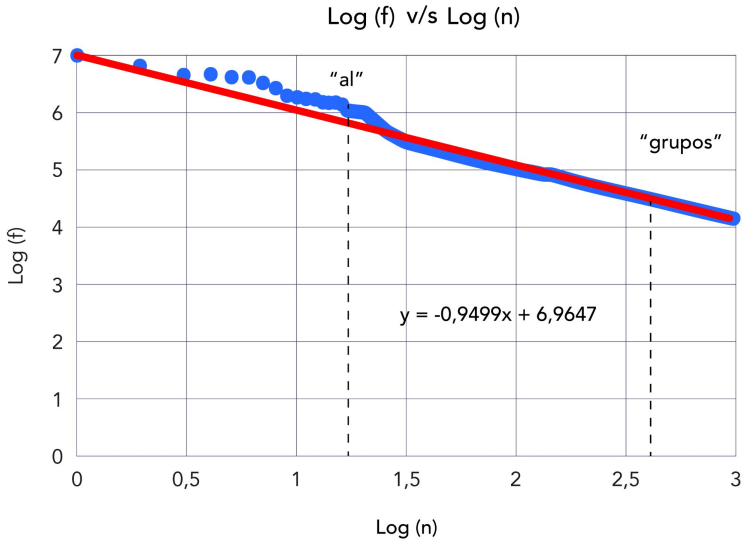
4. ¿Qué tipo de función podría modelar el gráfico log(f) versus log(n)?



Actividad 1

5. Responde las siguientes preguntas:

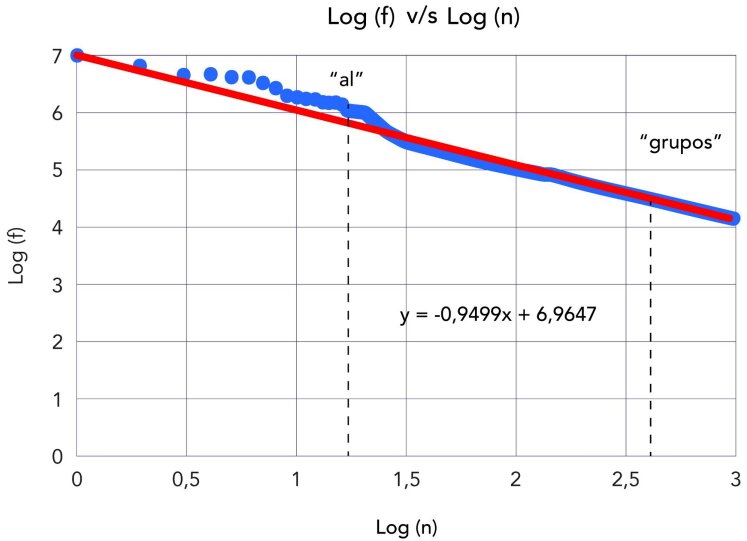
a) ¿La pendiente del modelo es positiva o negativa? ¿Dónde cruza el eje Y?



Actividad 1

5. Responde las siguientes preguntas:

a) ¿La pendiente del modelo es positiva o negativa? ¿Dónde cruza el eje Y?

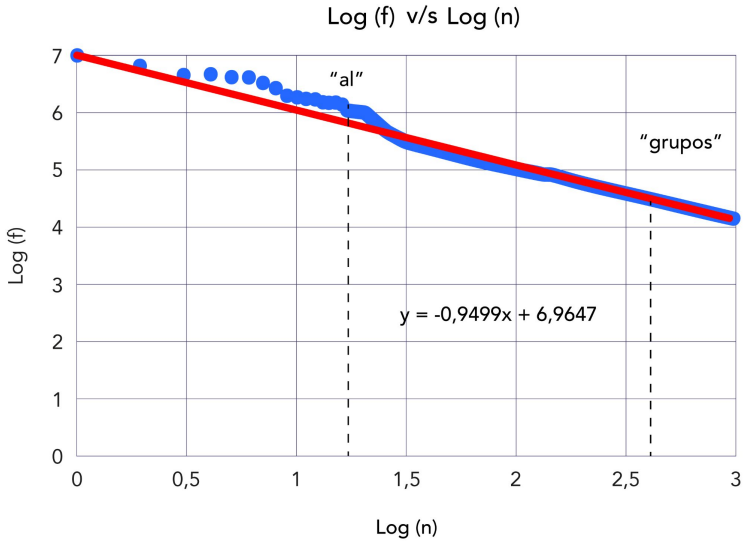


- Como la recta es decreciente tiene pendiente **negativa**, es decir, **a < 0**.
- La recta cruza el eje Y en **b = 7** aproximadamente.

Actividad 1

5. Responde las siguientes preguntas:

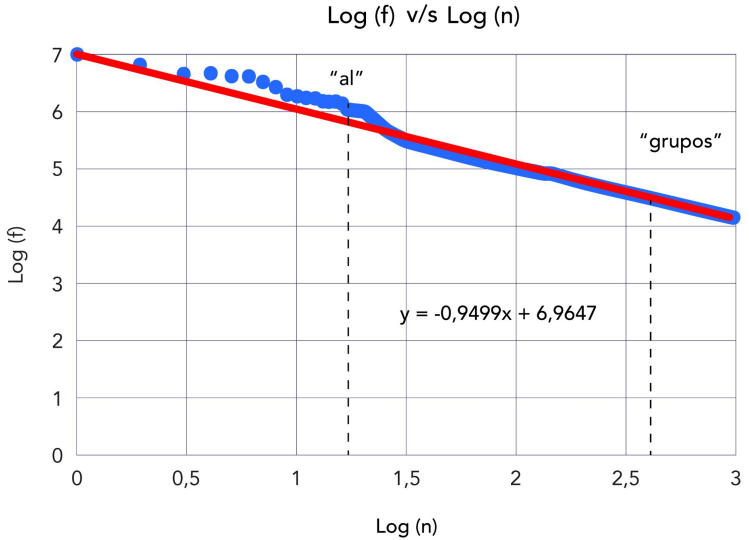
b) Observando los datos y la recta, ¿para qué rango de valores se ajusta mejor la recta?



Actividad 1

5. Responde las siguientes preguntas:

b) Observando los datos y la recta, ¿para qué rango de valores se ajusta mejor la recta?



Los datos se ajustan mejor a la recta desde log(n) = 1,5 aproximadamente.

Actividad 2

1. **Completa la siguiente tabla:**

Ranking (n)	Palabra	f real	f con el modelo	Diferencia positiva
20	al	951054		
453	grupos	27863		

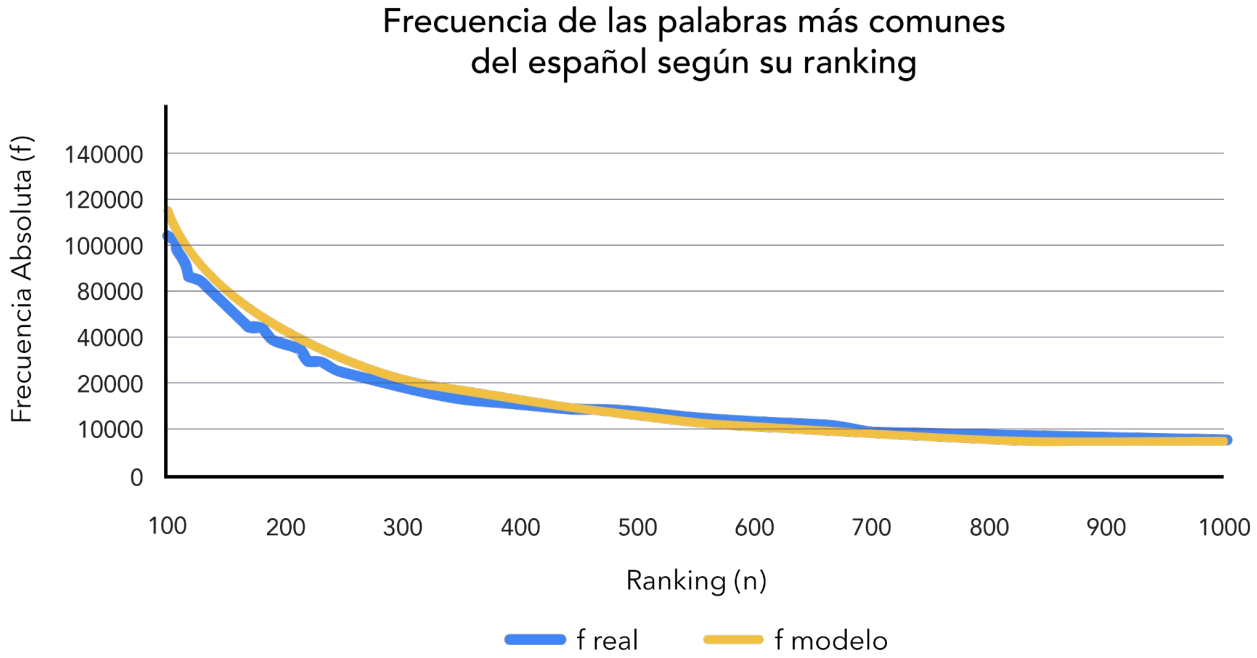
Actividad 2

MAT CON
MATEMÁTICA CONECTADA

1. **Completa la siguiente tabla:**

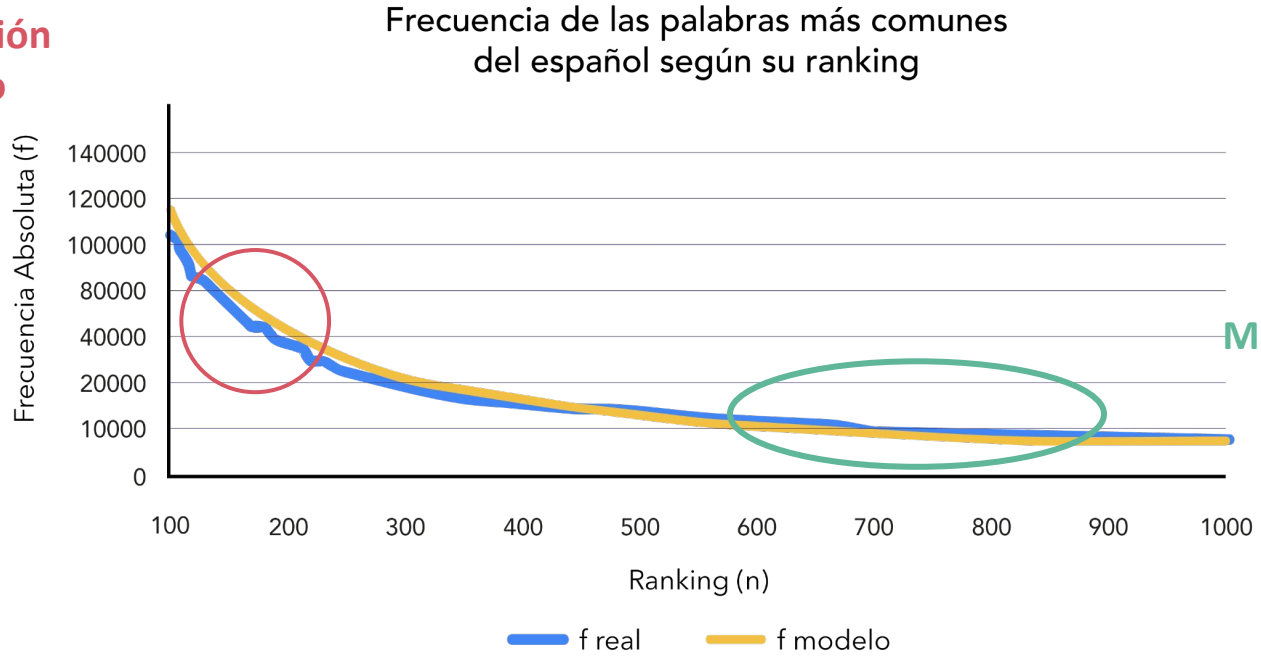
Ranking (n)	Palabra	f real	f con el modelo	Diferencia positiva
20	al	951054	535614	415440
453	grupos	27863	27649	214

- ¿En qué rangos del ranking el modelo predice de forma más precisa las frecuencias reales?
- ¿Qué tan bien predice las frecuencias el modelo?

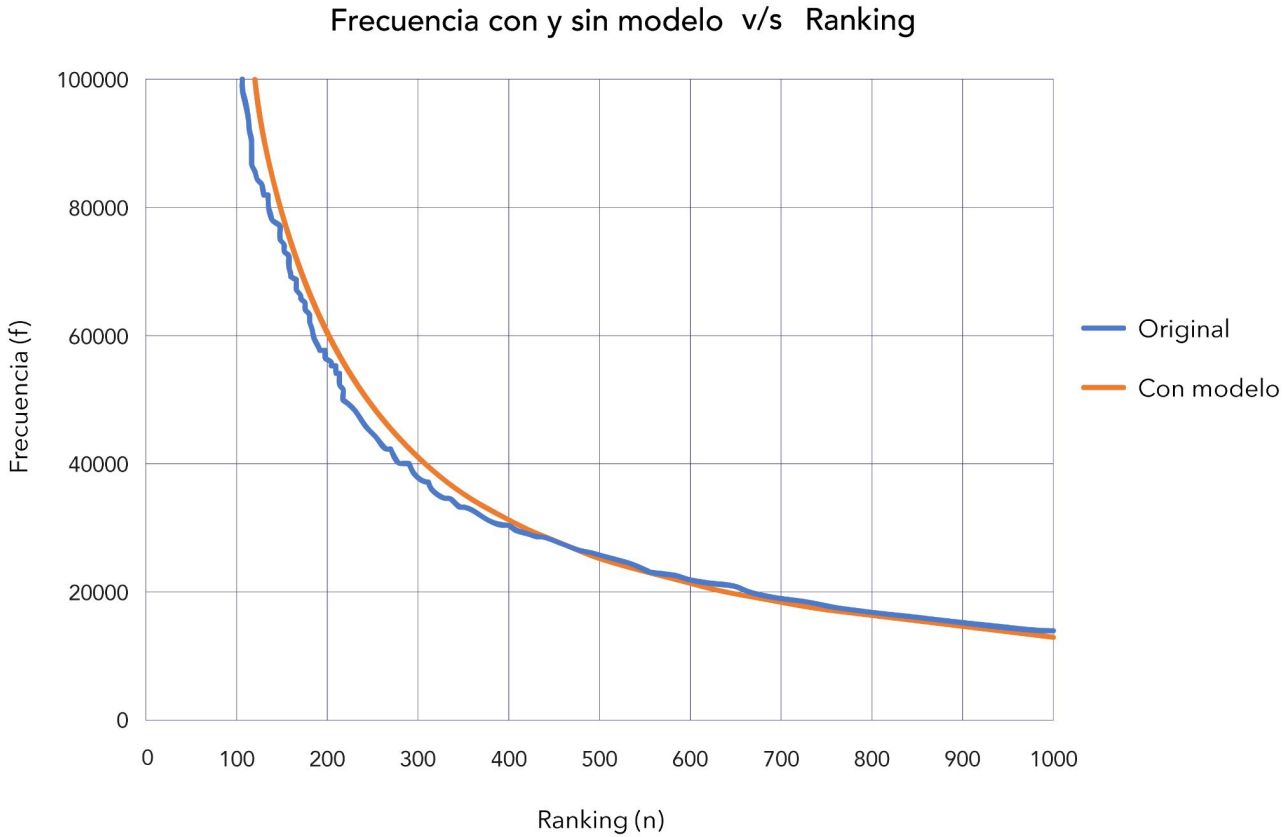


- ¿En qué rangos del ranking el modelo predice de forma más precisa las frecuencias reales?
- ¿Qué tan bien predice las frecuencias el modelo?

Menor precisión del modelo

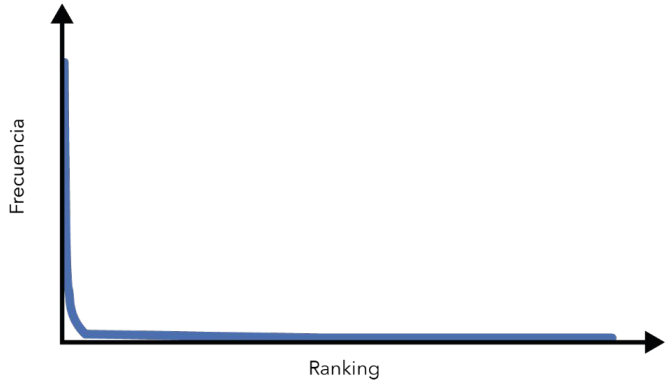


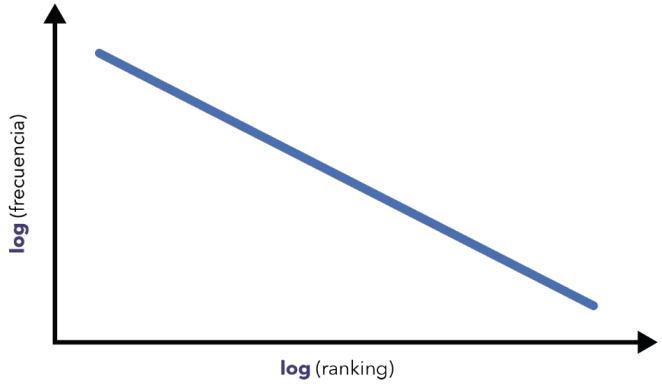
Mayor precisión del modelo


Frecuencia con y sin modelo v/s Ranking
100000
80000
60000
40000
20000
0
Frecuencia (f)
0
100
200
300
400
500
600
700
800
900
1000
Ranking (n)
Original
Con modelo

Sistematización

- Al trabajar con el cambio de escala logarítmica, se facilita la tarea de encontrar el modelo que se ajusta a la situación, en este caso lineal, simplificando así la tarea de análisis.





Sistematización

- La función inversa del logaritmo se llama exponencial. Si se tiene una función logarítmica de la forma $f(x) = \log_a(x)$, su función inversa es $f^{-1}(x) = a^x$.

$$f(x) = \log_a(x)$$

Función logaritmo

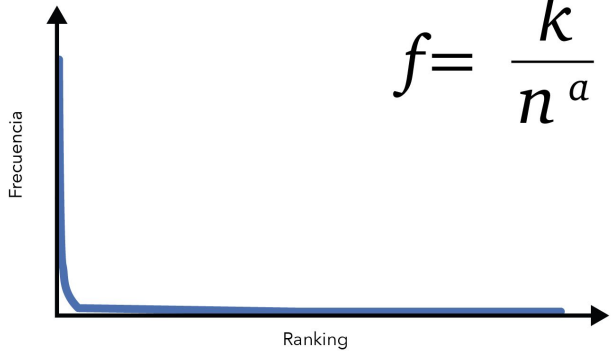
$$f^{-1}(x) = a^x$$

Función exponencial

Sistematización

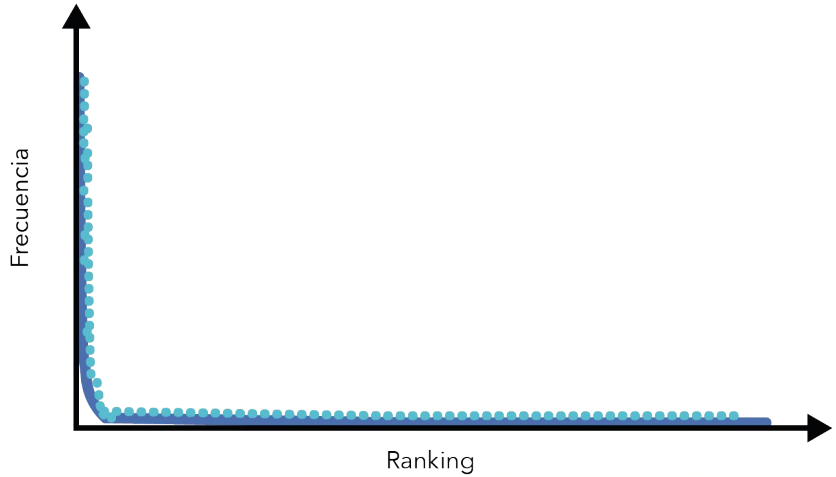
La **ley de Zipf** es una ley empírica que describe la distribución de las frecuencias de las palabras en corpus lingüísticos. Esta ley establece que la frecuencia de una palabra en un texto es inversamente proporcional a su posición en el ranking de frecuencias, es decir, la palabra más frecuente aparecerá aproximadamente el doble de veces que la segunda palabra más frecuente, tres veces más que la tercera palabra más frecuente, y así sucesivamente.

$$f = \frac{k}{n^{a}}$$



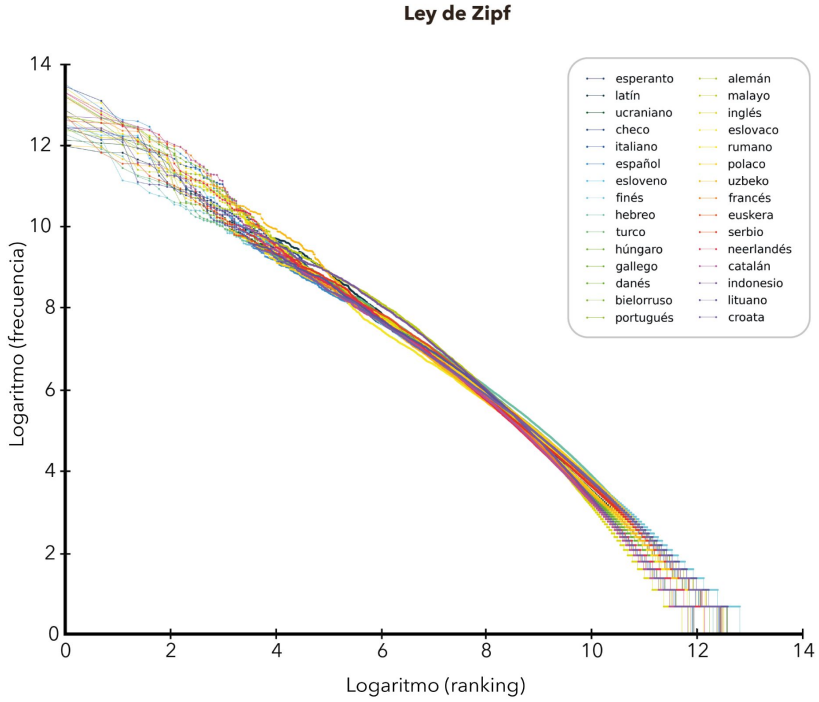
Sistematización

- La ley de Zipf se comprueba y se valida a través de análisis estadísticos y empíricos de grandes corpus lingüísticos. En esta clase, validamos la distribución de frecuencia de las palabras en un corpus, observando como los datos se comportan de acuerdo con los patrones descritos por la ley.



Sistematización

- Es interesante comprender cómo la matemática se aplica en campos que parecieran tan lejanos a ella como la lingüística. Se ha comprobado que la ley de Zipf se cumple en una gran variedad de idiomas.



MAT CON
MATEMÁTICA CONECTADA

Matemática y lingüística: La ley de Zipf

fcfm FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE
CMMEdu Laboratorio de Educación Centro de Modelamiento Matemático
UNESCO Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura
unitwin Cátedra UNESCO Formación de Docentes para Enseñar Matemática en el Siglo XXI Universidad de Chile
Ministerio de Educación Gobierno de Chile
PLAN DE REACTIVACIÓN EDUCATIVA