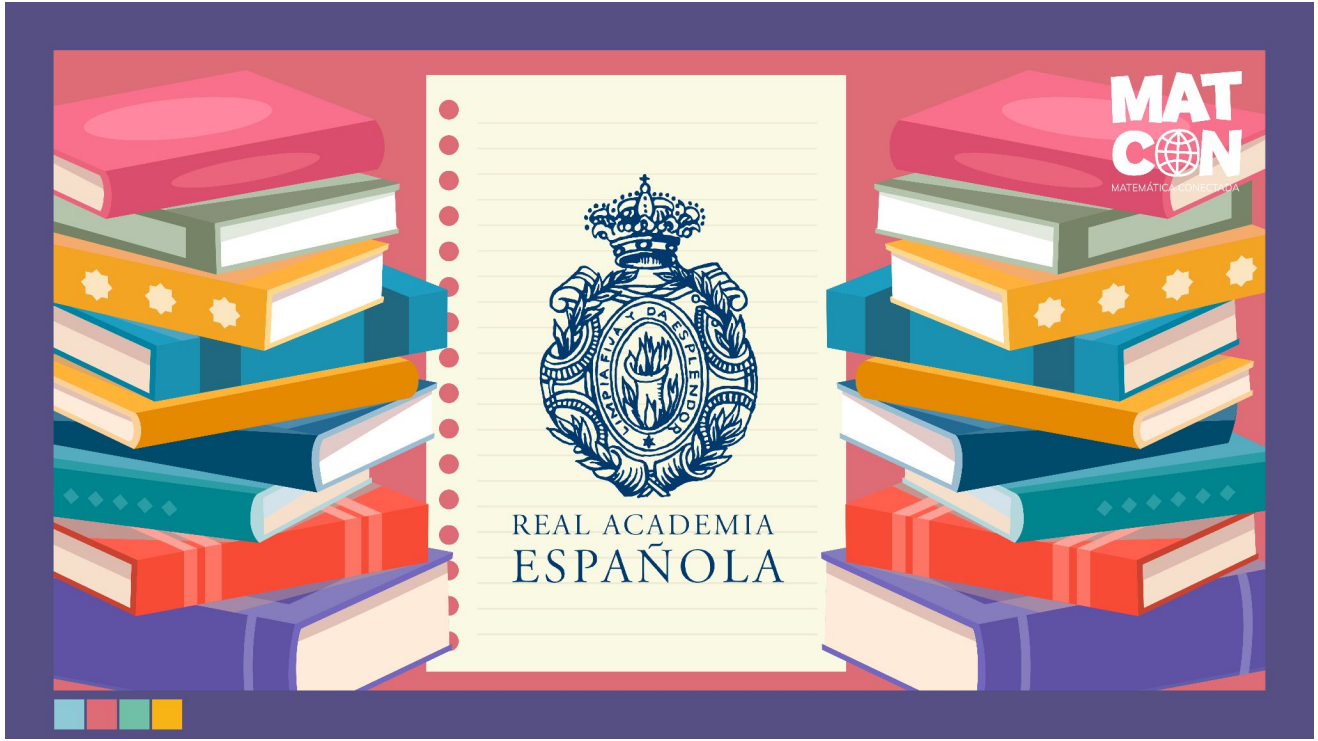
MAT CON
MATEMÁTICA CONECTADA

Matemática y lingüística: La ley de Zipf

fcfm
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
SEAMOS COMUNIDAD

El misterio de la ley de Zipf



El misterio de la ley de Zipf

1. ¿Qué significa la sigla CREA?
2. ¿Cómo podrías explicar con tus palabras la ley de Zipf?
3. ¿Cuántas veces aparecerá la décima palabra más usada en el CREA?

REAL ACADEMIA
ESPAÑOLA

Presentación del problema

La base de datos del Corpus de Referencia del Español Actual (CREA) es una herramienta importante, ya que permite acceder a grandes cantidades de datos lingüísticos en un formato estructurado y fácilmente consultable.

Ranking	Palabra	Frecuencia absoluta (f)
1	de	9999518
2	la	6277560
3	que	4681839
4	el	4569652
5	en	4234281

¿Se comprueba la ley de Zipf para el CREA?

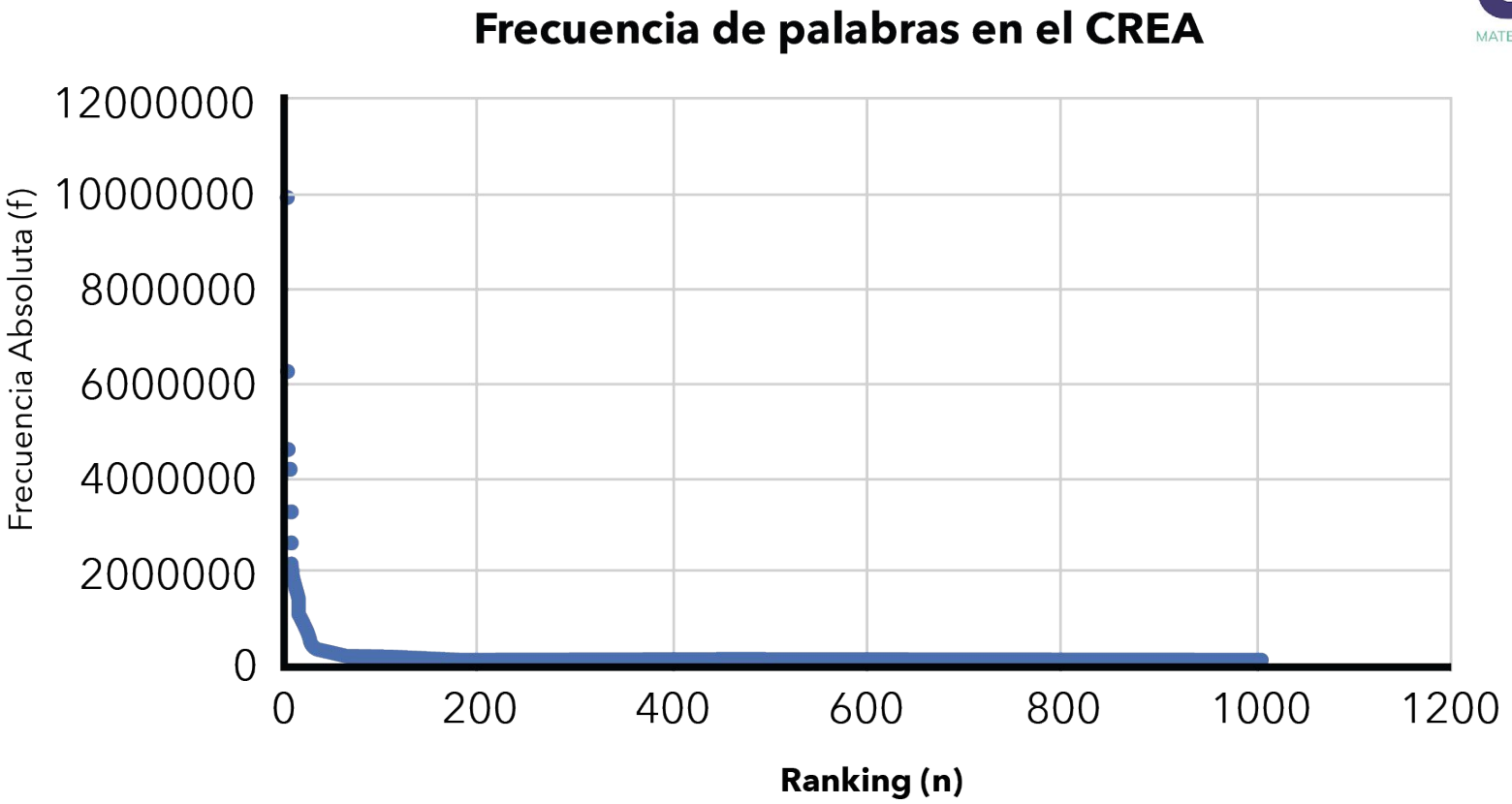
Analizando la base de datos

Ranking	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
13	por	1561904
14	con	1481607
15	no	1465503

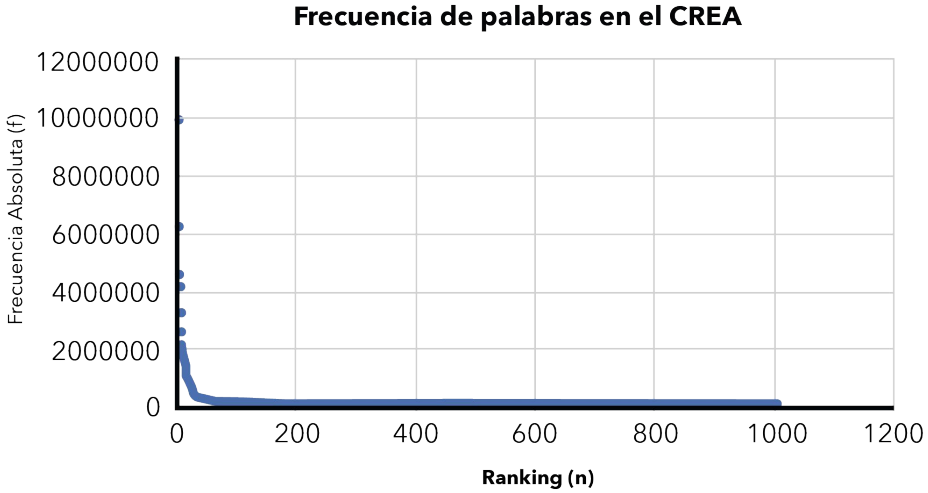
MAT CON
MATEMÁTICA CONECTADA

- ¿Qué representa la frecuencia absoluta?
- ¿Cuántas veces aparece la palabra “en”?
- ¿En qué posición del ranking está la palabra “que”?
- ¿Cuál es la palabra que está en la posición 999 del ranking?

Analizando el gráfico

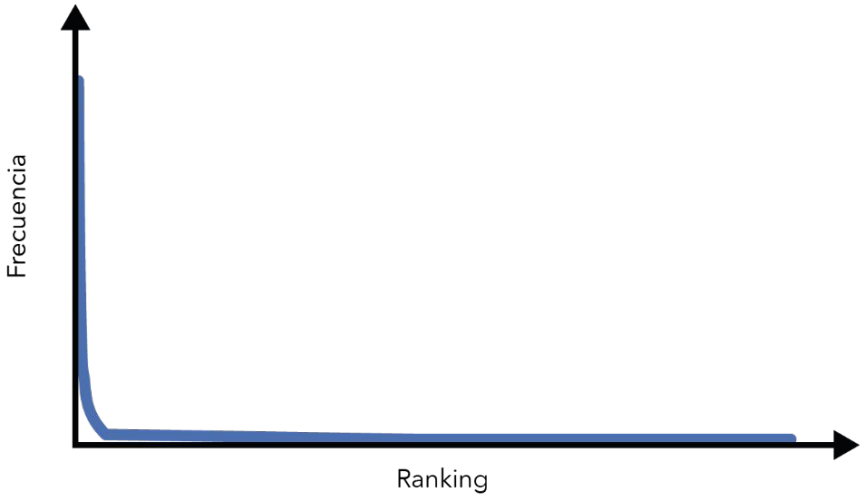


Analizando el gráfico

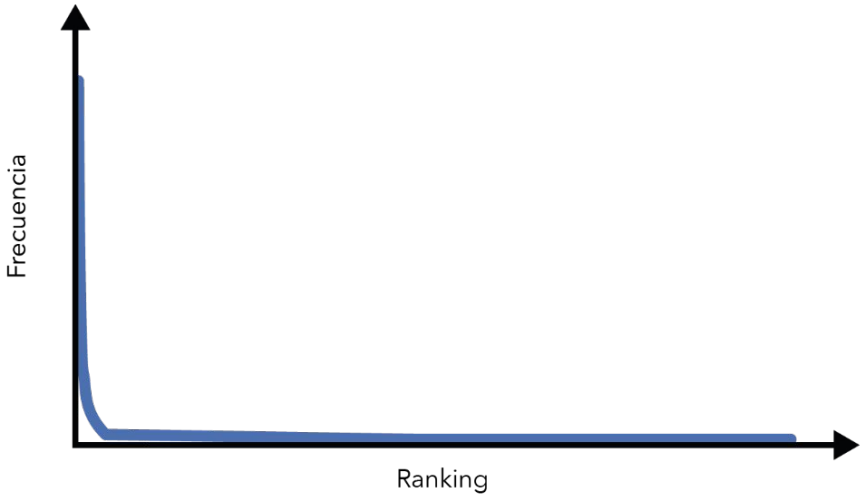


- ¿Por qué el gráfico se observa decreciente?
- A partir del gráfico, ¿es posible apreciar la frecuencia de la palabra número 2? ¿Y de la número 200?
- ¿Cómo es el decrecimiento a medida que se avanza en el ranking?
- ¿A qué tipo de función podría corresponder el gráfico?

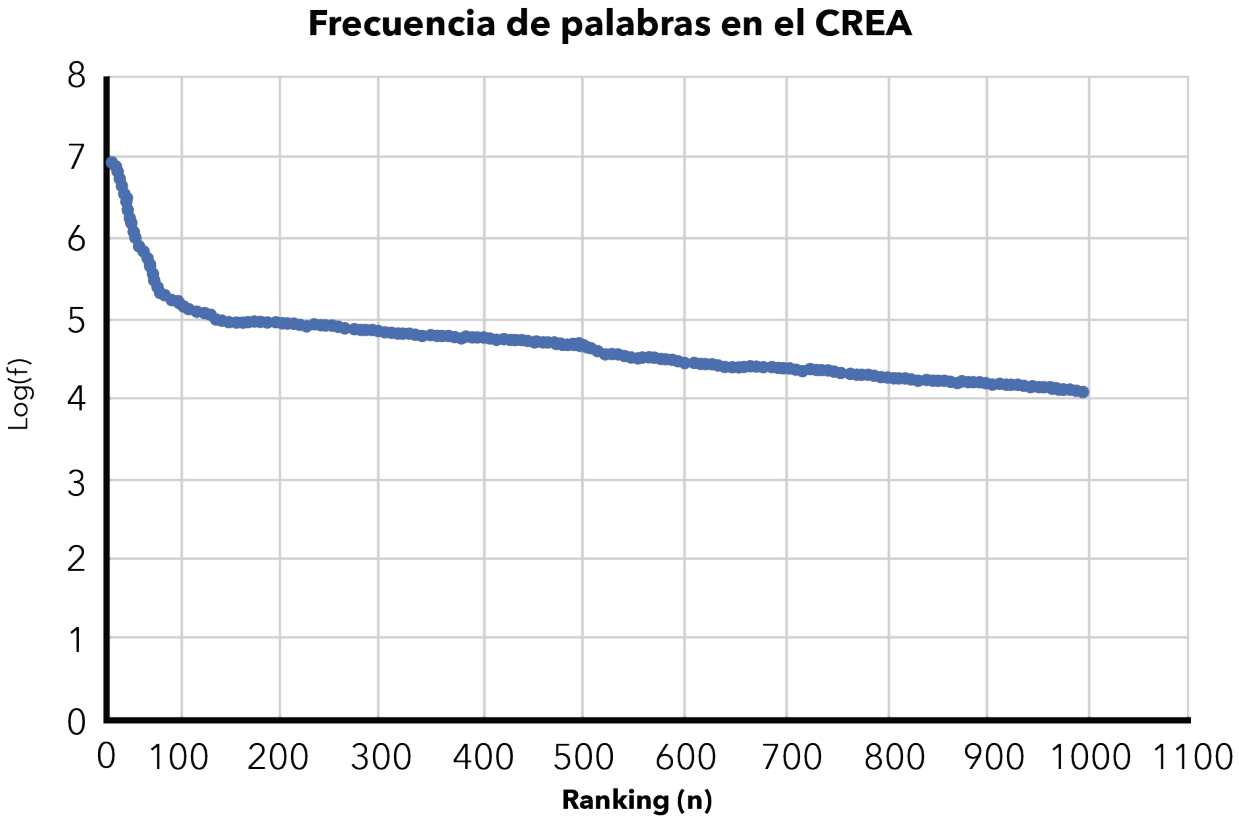
¿Cómo estudiar este gráfico de forma que podamos observar más detalladamente el decrecimiento?



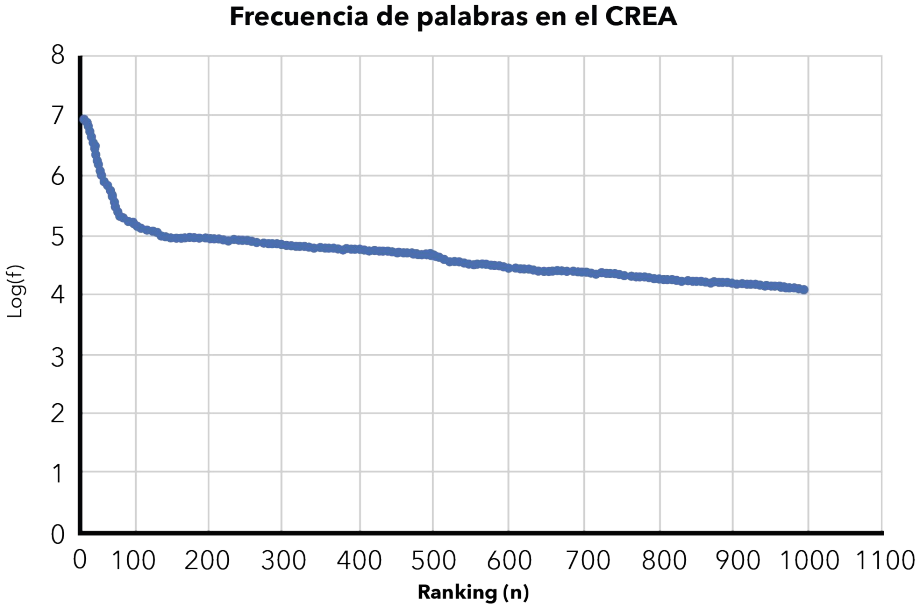
¿Cómo estudiar este gráfico de forma que podamos observar más detalladamente el decrecimiento?



Analizando el nuevo gráfico



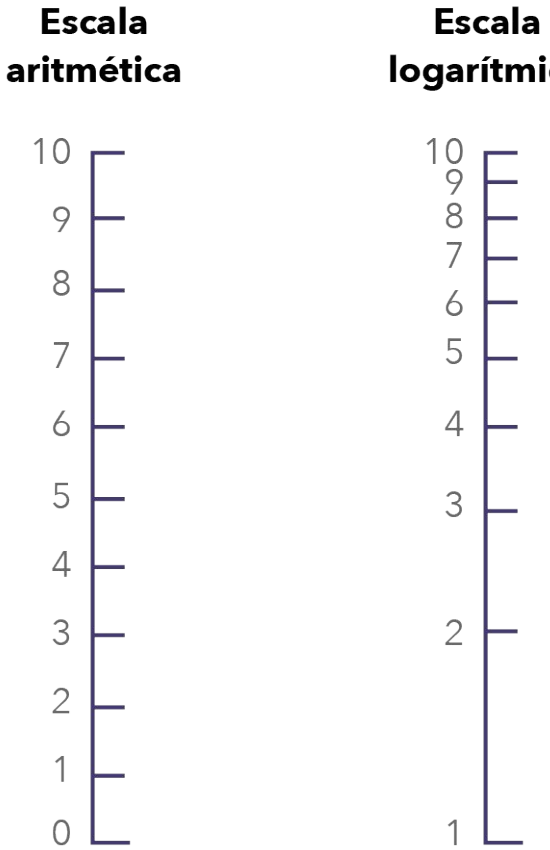
Analizando el nuevo gráfico



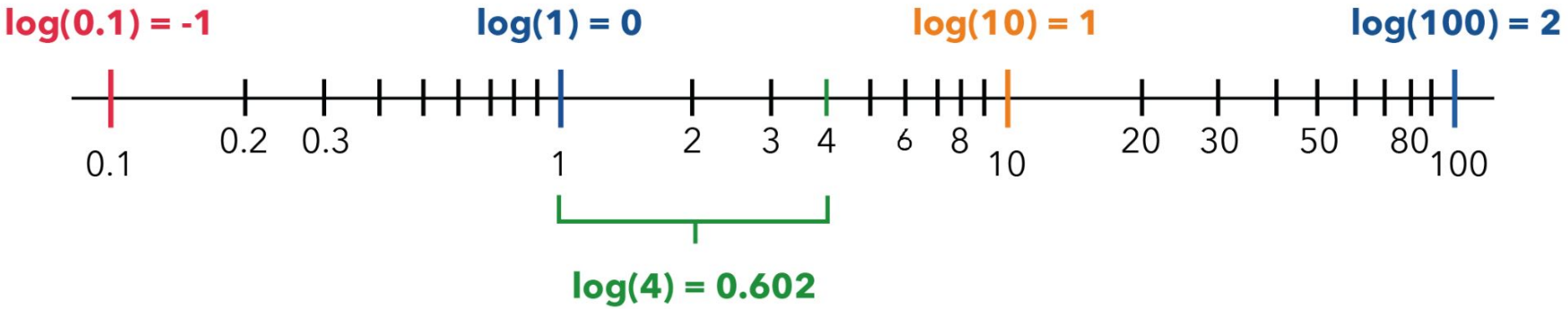
- ¿Qué significa que la palabra con mayor frecuencia esté aproximadamente en la coordenada (1, 7)?
- A partir de este nuevo gráfico, ¿es posible apreciar la frecuencia de la palabra número 200?

Escala logarítmica

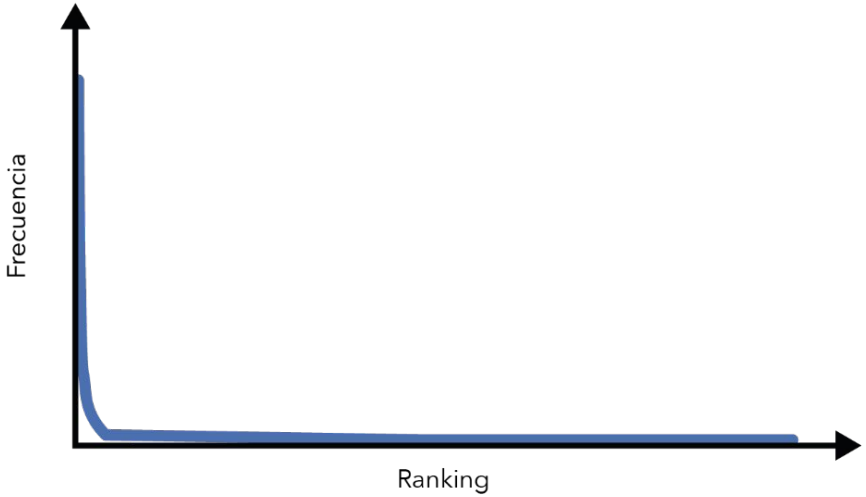
- La **escala logarítmica** es una forma de representación gráfica basada en potencias de un número fijo llamado **base logarítmica**.
- En la escala logarítmica, las distancias entre los valores aumentan **exponencialmente**, en lugar de linealmente.



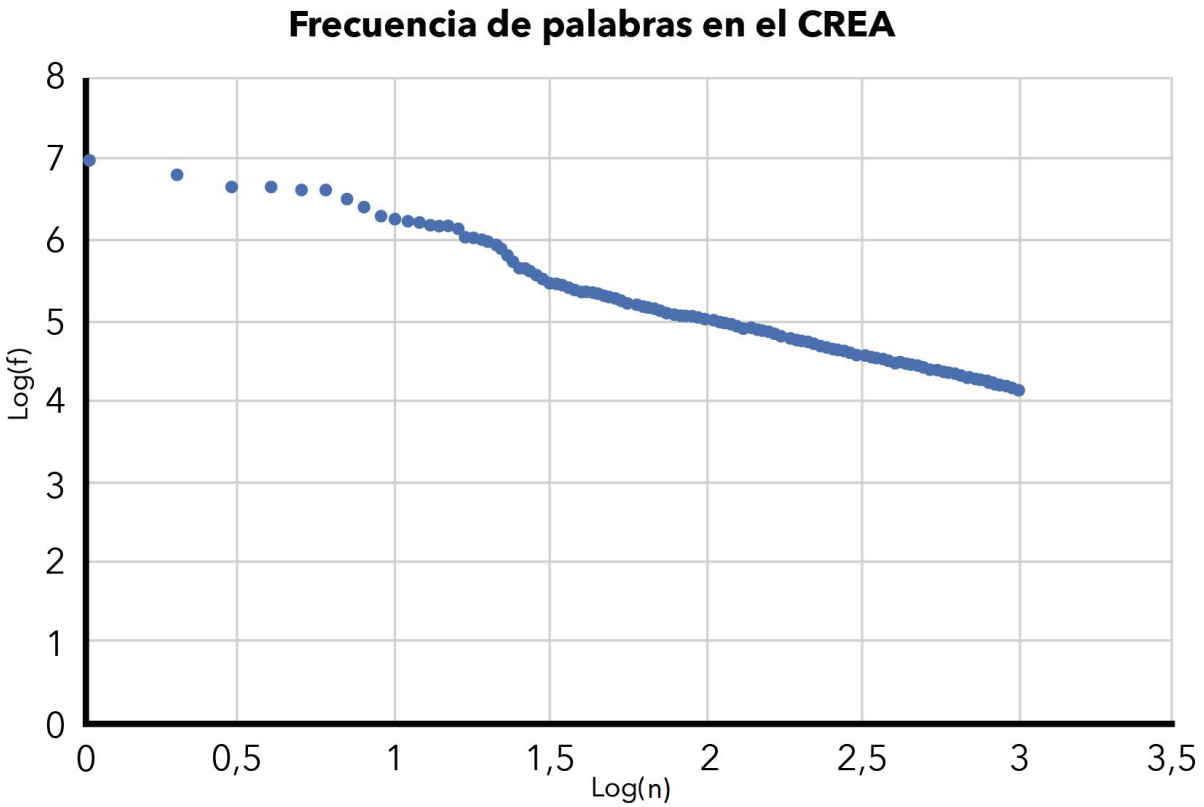
Escala logarítmica



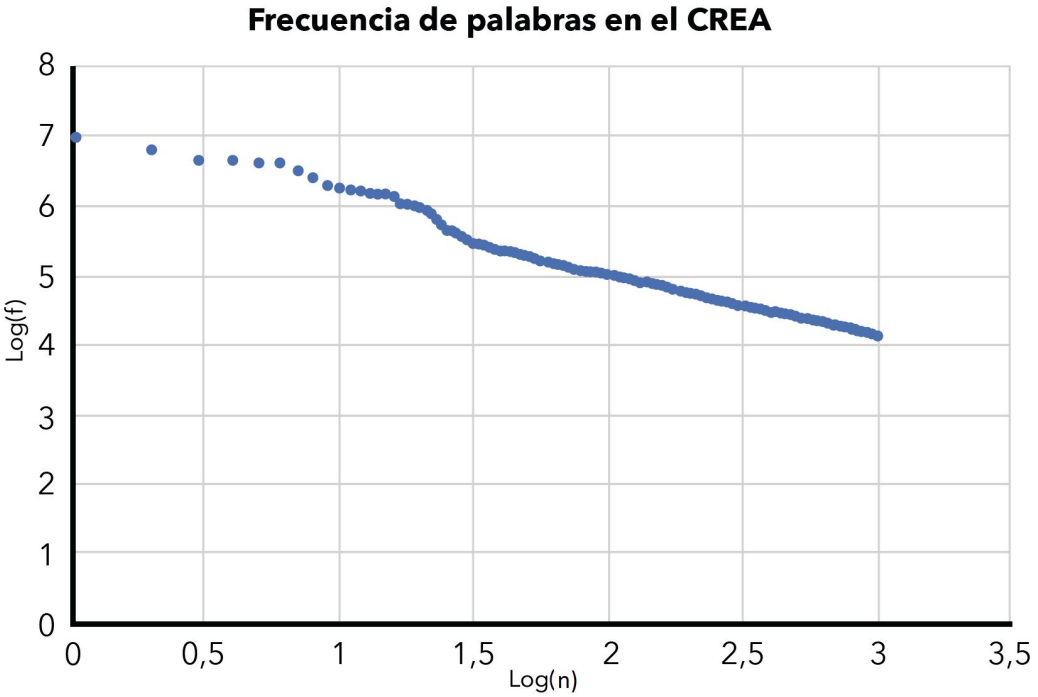
¿Qué pasa si usamos la escala logarítmica en ambos ejes?



Analizando el nuevo gráfico

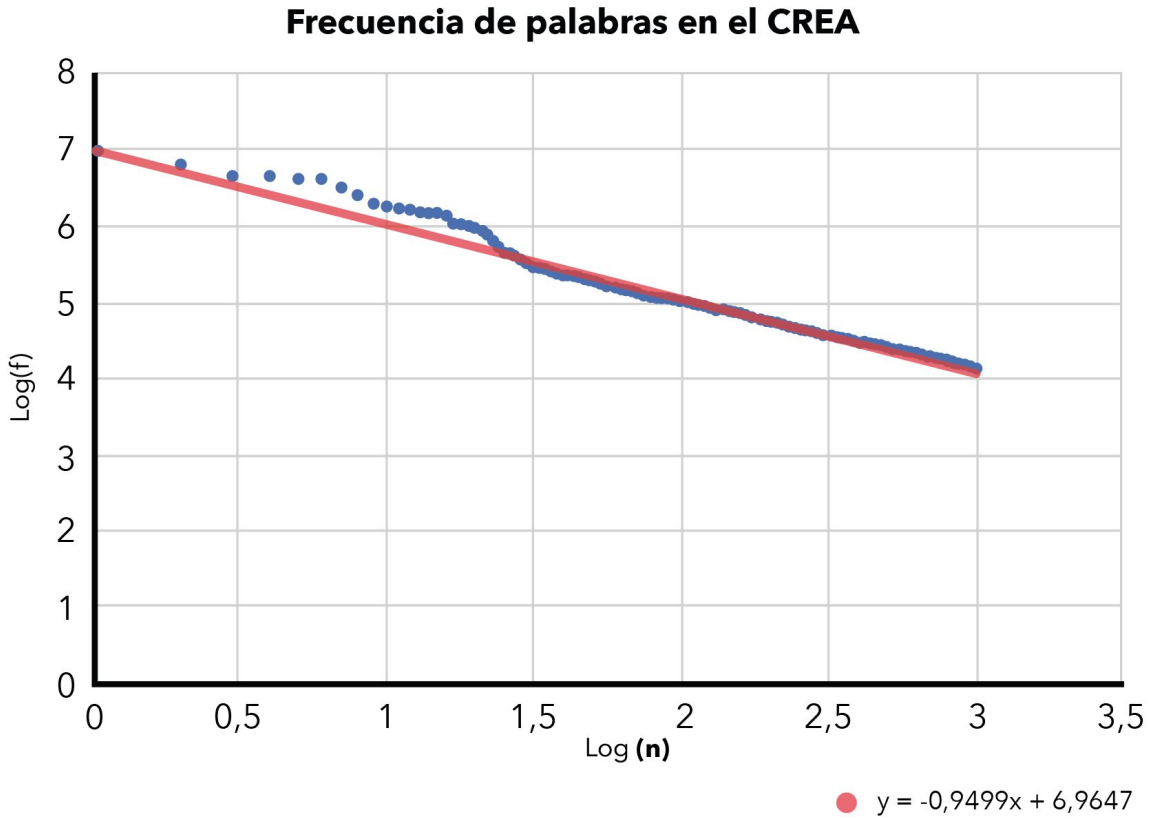


Analizando el nuevo gráfico

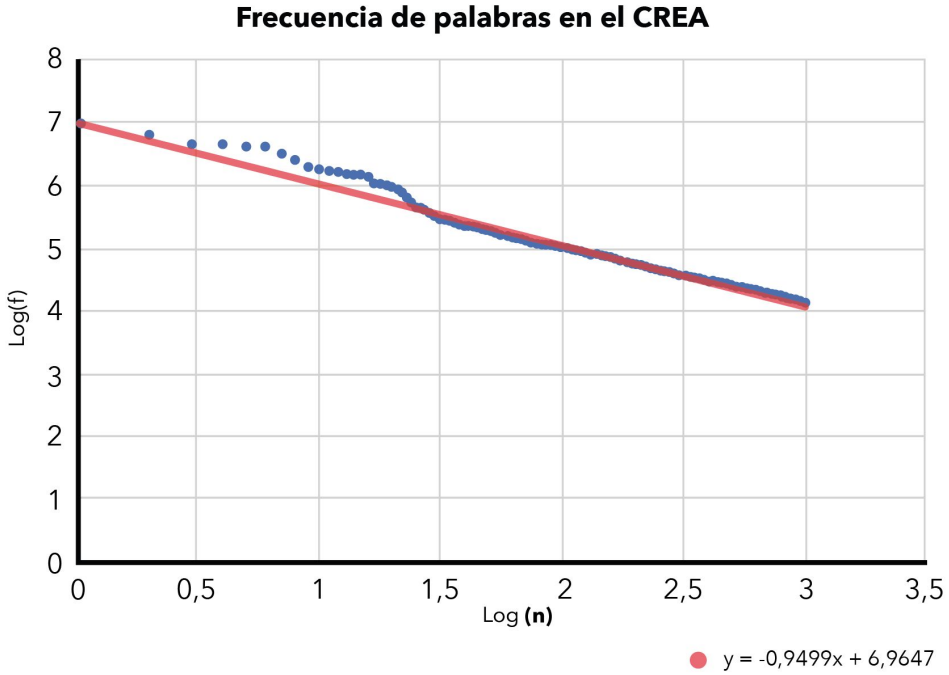


- ¿Qué características tiene este gráfico?
- ¿Cómo funciona la escala logarítmica en el eje x? ¿Qué significa x=3?
- ¿Qué tipo de función podría modelar lo que se observa en el gráfico?

Analizando la línea de tendencia



Analizando la línea de tendencia



- ¿Qué tan bien se ajusta la línea de tendencia a los datos?
- ¿La pendiente del modelo es positiva o negativa? ¿Cuál es su valor?

Tenemos que:

$$y = -a \cdot x + b$$

$$log(f) = -a \cdot log\ (n) + b$$

¿Cómo escribir esta relación con potencias?

Estrategia 1: Propiedades de la exponencial

$$\log(f) = -a \cdot \log(n) + b$$

$$10^{\log(f)} = 10^{(-a \cdot \log(n) + b)}$$

$$10^{\log(f)} = 10^{-a \cdot \log(n)} \cdot 10^{b}$$

$$10^{\log(f)} = \left(10^{\log(n)}\right)^{-a} \cdot 10^{b}$$

$$f = n^{-a} \cdot 10^{b}$$

Estrategia 2: Propiedades de los logaritmos

$$\log(f) = -a \cdot \log(n) + b$$

$$\log(f) = \log\left(n^{-a}\right) + \log(10^{b})$$

$$\log(f) = \log\left(n^{-a} \cdot 10^{b}\right)$$

$$f = n^{-a} \cdot 10^{b}$$

Ley de Zipf

$$f = n^{-a} \cdot 10^{b} = \frac{10^{b}}{n^{a}}$$

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

Reemplazando…

$$y = -ax + b$$

$$y = -0{,}9499x + 6{,}9647$$

$$f = \frac{10^{b}}{n^{a}} = \frac{10^{6{,}9647}}{n^{0{,}9499}}$$

Hoja de Actividades

1. **Dado el modelo lineal encontrado, ¿cuál sería la frecuencia para la palabra que está en el ranking número 60 y en el ranking número 900?**

Hoja de Actividades

1. **Dado el modelo lineal encontrado, ¿cuál sería la frecuencia para la palabra que está en el ranking número 60 y en el ranking número 900?**

$$f_{60} = \frac{10^{6,9647}}{60^{0,9499}} \approx 188\,640$$

$$f_{900} = \frac{10^{6,9647}}{900^{0,9499}} \approx 14\,403$$

Hoja de Actividades

2. **Considerando la base de datos del CREA, ¿cuáles son las frecuencias de las palabras número 60 y 900 del ranking? ¿Cuáles son esas palabras?**

Hoja de Actividades

MAT CON
MATEMÁTICA CONECTADA

2. Considerando la base de datos del CREA, ¿cuáles son las frecuencias de las palabras número 60 y 900 del ranking? ¿Cuáles son esas palabras?

Palabra	Ranking	Frecuencia absoluta (f)
todos	60	158 168
busca	900	15 280

Hoja de Actividades

3. **Considerando el modelo lineal y la base de datos del CREA, ¿cuál es la diferencia entre las frecuencias absolutas para las palabras de la pregunta anterior? ¿Qué se puede decir de esas diferencias?**

Hoja de Actividades

MAT CON
MATEMÁTICA CONECTADA

3. Considerando el modelo lineal y la base de datos del CREA, ¿cuál es la diferencia entre las frecuencias absolutas para las palabras de la pregunta anterior? ¿Qué se puede decir de esas diferencias?

Ranking	Frecuencia según el modelo	Frecuencia según la base de datos	Diferencia positiva
60	188 640	158 168	30 472
900	14 403	15 280	877

Hoja de Actividades

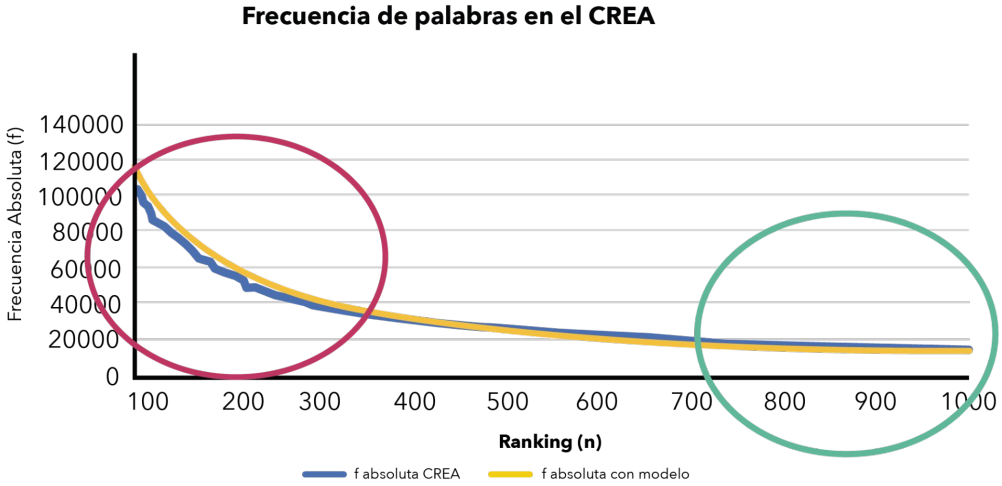
4. **Analiza diferentes rankings (por ejemplo: 1, 11, 25, 500, 800, 1000) y responde: ¿qué tan bien predice las frecuencias el modelo? ¿En qué rangos del ranking funciona mejor el modelo? Justifica.**

Hoja de Actividades

MAT CON
MATEMÁTICA CONECTADA

4. Analiza diferentes rankings (por ejemplo: 1, 11, 25, 500, 800, 1000) y responde: ¿qué tan bien predice las frecuencias el modelo? ¿En qué rangos del ranking funciona mejor el modelo? Justifica.

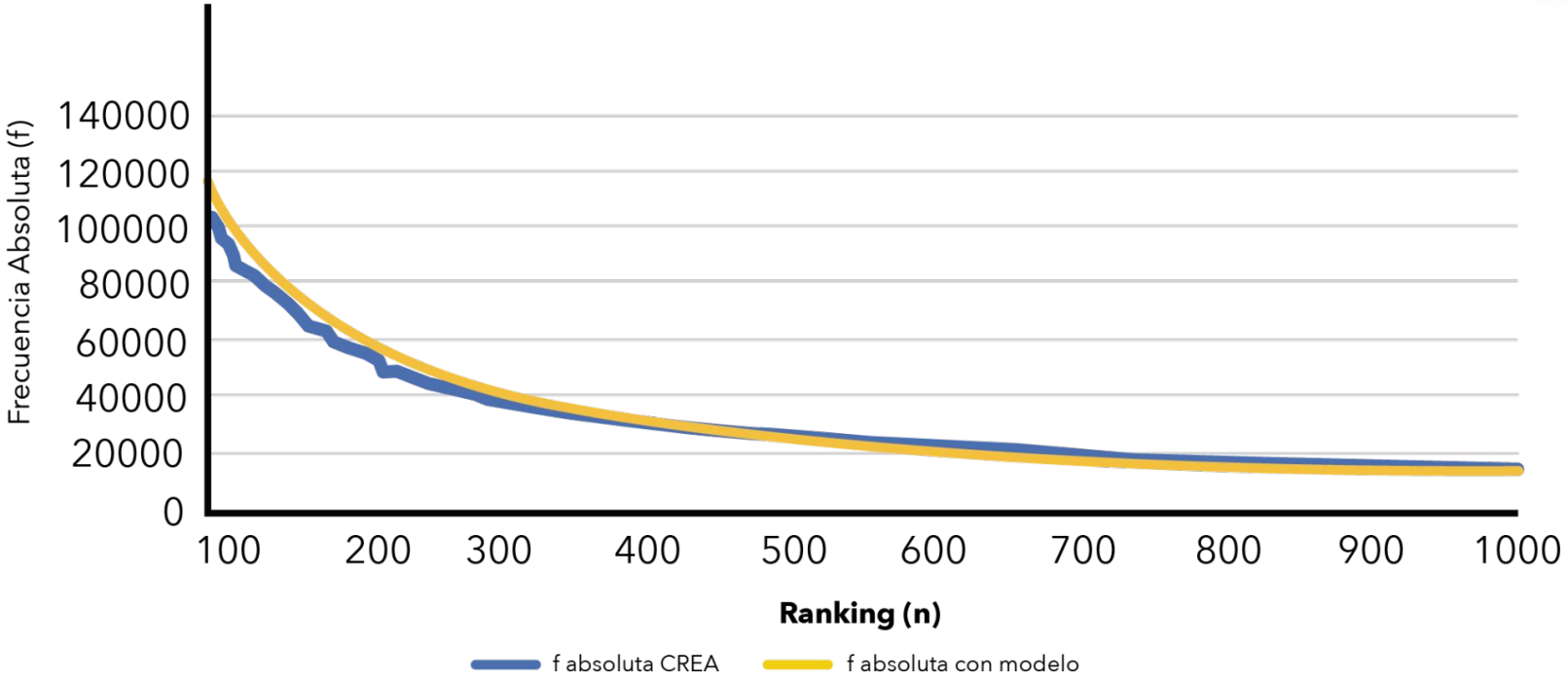
Menor precisión del modelo



Mayor precisión del modelo

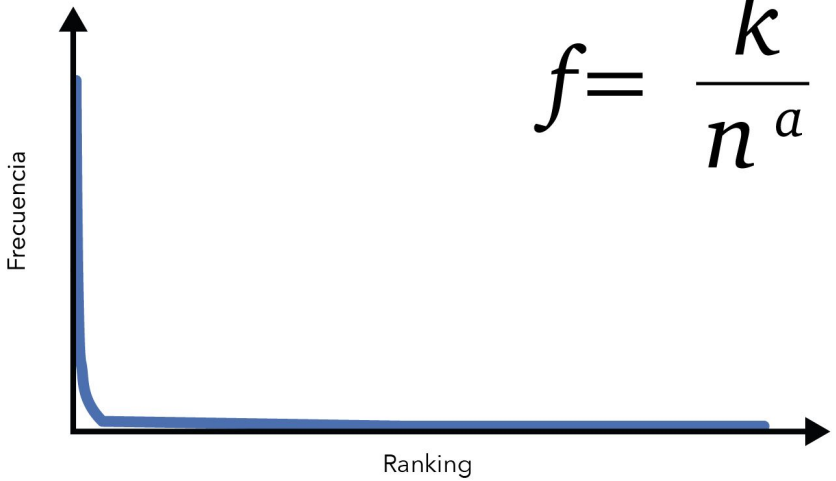
MAT CON
MATEMÁTICA CONECTADA

Frecuencia de palabras en el CREA



Sistematización

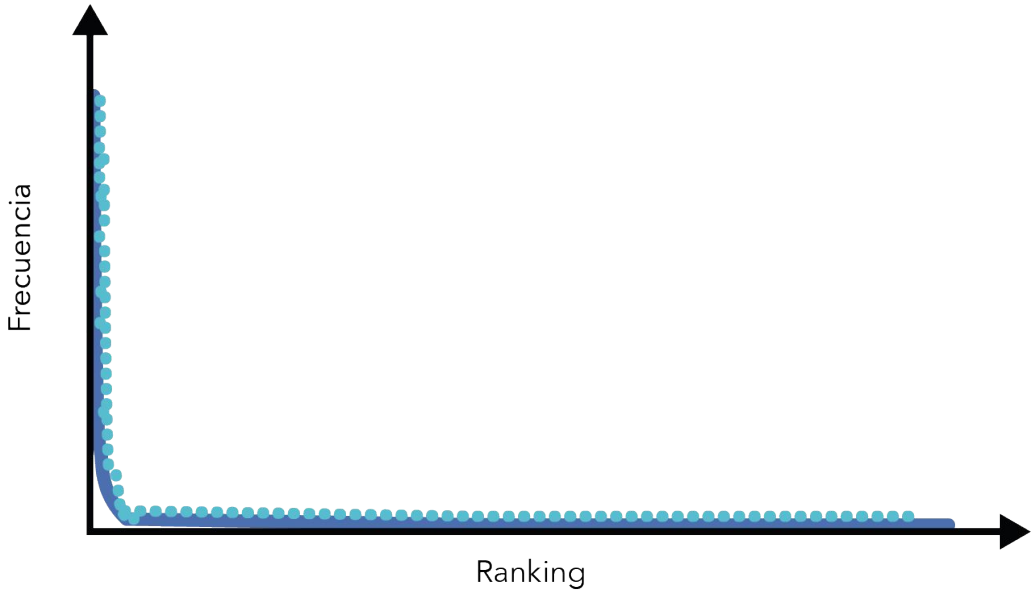
- La ley de Zipf es una ley empírica que describe la distribución de frecuencia de las palabras en corpus lingüísticos.



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

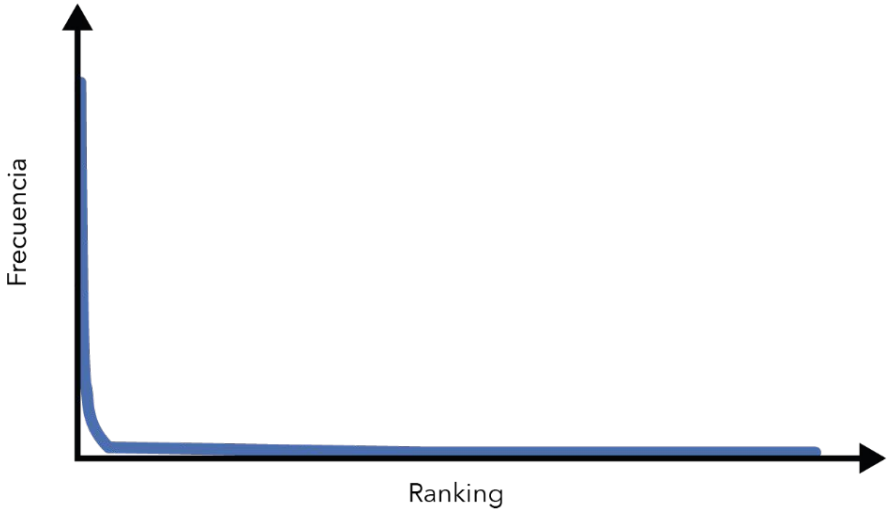
Sistematización

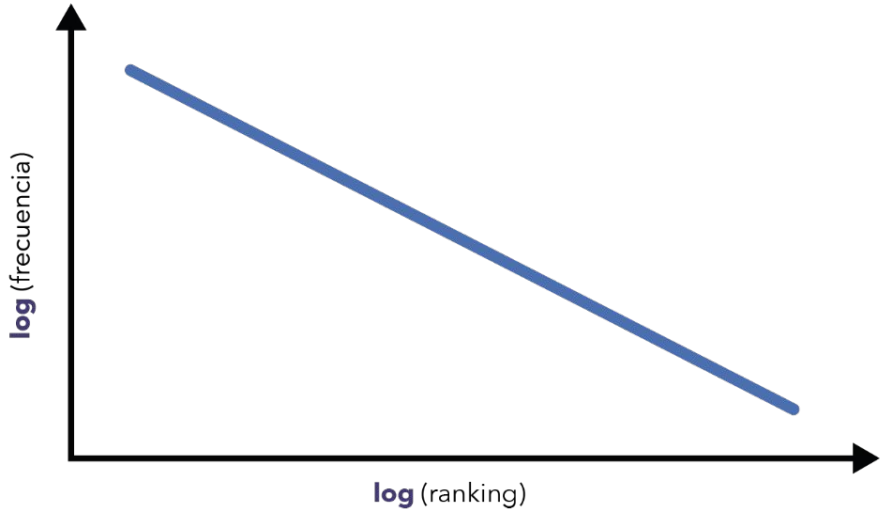
- La ley de Zipf se comprueba y se valida a través de análisis estadísticos y empíricos de grandes corpus lingüísticos.



Sistematización

- Al trabajar con los logaritmos de las variables involucradas en la ley de Zipf, se facilita la tarea de encontrar el modelo que describe la situación.





Sistematización

- La función inversa del logaritmo se llama exponencial.

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

Función logaritmo

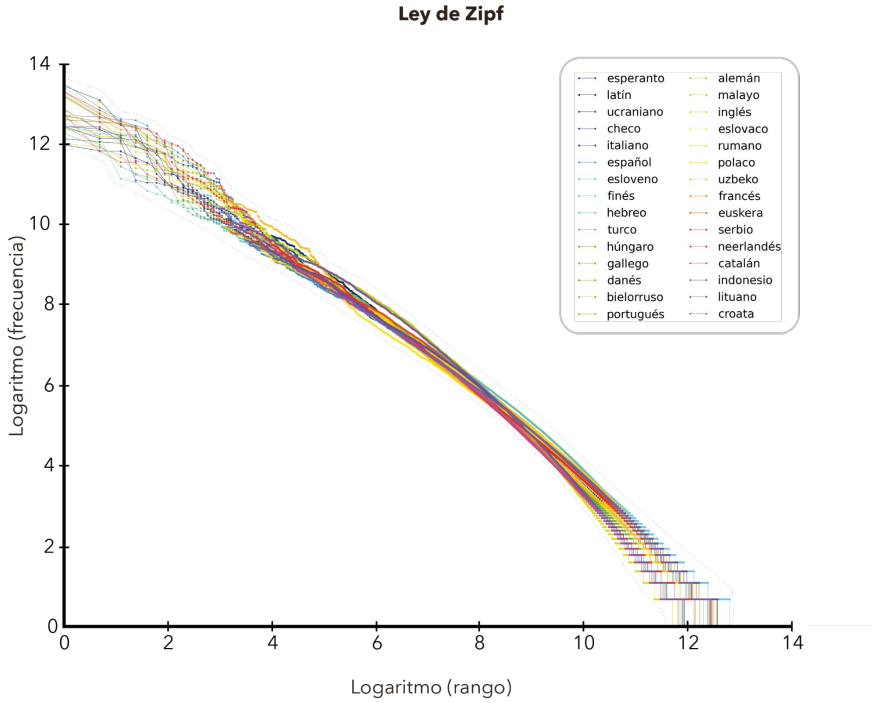
$$f^{-1}(x) = a^x$$

Función exponencial

Sistematización

- Es interesante ver 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
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
SEAMOS COMUNIDAD