

Lei de Zipf

Anthropic & OpenAI & awm Chat multimodel 22 fev. 2026

Num determinado contexto, ao observarmos a fala ou a escrita das pessoas, percebemos que algumas palavras surgem com uma frequência impressionante, enquanto outras aparecem raramente. Essa repetição não é aleatória: há uma tendência clara, quase uma “ordem invisível” na linguagem. Foi justamente essa curiosidade que levou George Kingsley Zipf, no início do século XX, a estudar padrões linguísticos.

Em seu livro *The Psycho-Biology of Language* (1935), Zipf demonstrou que, em qualquer idioma, uma pequena quantidade de palavras domina o uso cotidiano, enquanto a maioria das palavras é empregada apenas esporadicamente. Quando organizamos as palavras pelo número de vezes que aparecem, a frequência de cada uma decai rapidamente conforme sua posição na lista. Zipf descreveu com precisão matemática essa distribuição de frequência, mostrando que a ocorrência de uma palavra é aproximadamente inversamente proporcional à sua classificação no ranking linguístico — o que ficou conhecido como a **Lei de Zipf**.

Essa lei revela que o uso das palavras não é aleatório e oferece uma base conceitual poderosa: ela explica, por exemplo, como sistemas modernos podem priorizar e filtrar informações, transformando radicalmente a forma como motores de busca processam enormes volumes de dados digitais.

De forma simplificada, a lógica da Lei de Zipf pode ser expressa assim:

$$frequência \approx 1 / posição$$

Ou seja, quanto maior a posição de uma palavra no ranking de frequência, menor será sua ocorrência.

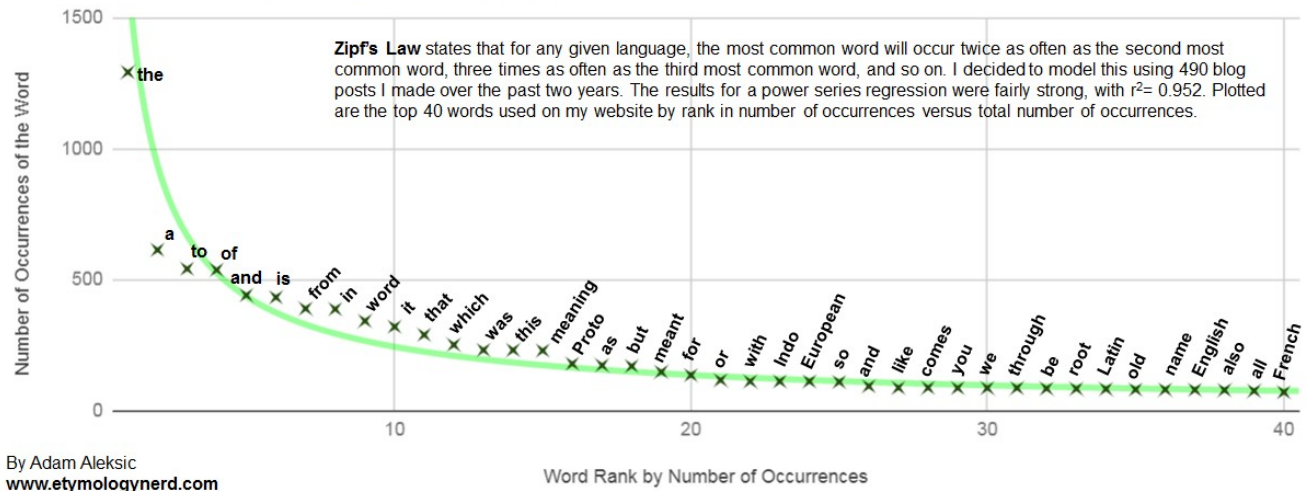
Exemplo

Ranking	Palavra	Frequência
1	"de"	10.000
2	"a"	5.000
3	"o"	3.300
4	"que"	2.500
5	"e"	2.000
6	"do"	1.650
7	"da"	1.430
8	"em"	1.250
9	"um"	1.110
10	"para"	1.000

Ao plotar **frequência × ordenamento**, obtém-se essa curva com cauda longa.

Exemplo

Zipf's Law and My Blog on Language



Grandes sistemas como Google e Amazon exploram isso intensamente, embora a Lei de Zipf seja aplicável a outros sistemas que os linguísticos:

Frequência de palavras ... Acessos a páginas web ... Termos de busca ... Vendas de produtos ...
Uso de APIs ... Erros mais comuns em logs ... Cache hits

A idéia mais importante é que pouquíssimos itens concentram grande parte do volume, algo parecido com a Lei de Pareto (80/20).

Exemplo típico:

- Top 5% → perfazem 70% das consultas
- Resto → enorme cauda longa de raridades

Outro aspecto interessante da Lei de Zipf, menos óbvio, pode ser observado no cotidiano: quando os alunos se divertem registrando, nas margens de seus cadernos, quantas vezes o professor repete certas palavras. Eles acabam se fixando nessas poucas palavras, enquanto perdem vários detalhes do conhecimento transmitido!

Esse é um ponto central para estruturar novos sistemas e otimizar os existentes:

***ao otimizar a resposta do sistema às solicitações mais frequentes,
sobram recursos para tratar as ocorrências menos frequentes.***