



Linguística Quantitativa

Leonardo Araujo

30 de Novembro de 2022

Linguística Quantitativa enquanto ciência

- Lei científica

Uma lei científica é um princípio básico, generalização ou regra que se mantém universalmente verdadeira sob condições particulares (McComas 2014).

• Lei científica

Uma lei científica é um princípio básico, generalização ou regra que se mantém universalmente verdadeira sob condições particulares (McComas 2014).

Uma lei, em ciência, é uma hipótese sistematicamente conectada a outras hipóteses em um campo do conhecimento e, ao mesmo tempo, corroborada por dados empíricos. Uma lei é dita universal quando é válida a todo tempo, em qualquer lugar e para todos objetos sob seu escopo (Köhler 2014; Bunge 2012).

Exemplos

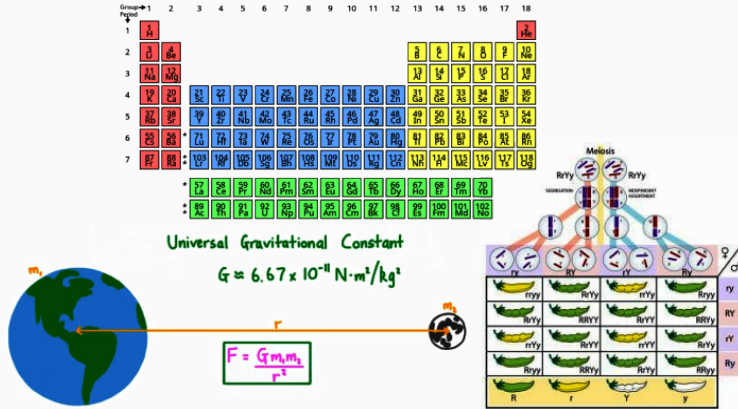


Figura 1: Lei da gravidade (<https://www.nagwa.com/en/videos/346134549365/>). Lei periódica (https://en.wikipedia.org/wiki/Periodic_table). Lei de Mendel (<https://biologydictionary.net/mendels-law-of-heredity/>).

Linguística Quantitativa

└ Linguística Quantitativa enquanto ciência

└ Exemplos

Exemplos

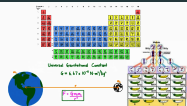


Figura 1: Lei da gravidade (<https://www.nasa.com/en/videos/24613454305/>) Lei periódica (https://en.wikipedia.org/wiki/Periodic_table) Lei de Mendel (<https://biologydictionary.net/mendels-law-of-heredity/>)

A **lei da gravidade** prediz a força de atração entre corpos com base em suas massas e distância entre eles. Entretanto, não explica o motivo da existência de tal atração ou busca explicar o motivo pelo qual tal comportamento é descrito pela equação $F = G \frac{m_1 m_2}{r^2}$.

A **lei periódica** estabelece que certas propriedades físicas e químicas dos elementos serão recorrentes de maneira sistemática e previsível quando os elementos são ordenados de forma crescente de seus número atômico. As principais propriedades em que observamos uma tendência devido à lei da periodicidade são: raio atômico, raio iônico, energia de ionização, eletronegatividade e eletroafinidade.

As **leis de Mendel** (ou leis da herança) são as três leis a seguir: 1) lei da dominância (quando pais com traços puros e contrastantes são cruzados, apenas uma forma de traço aparece na próxima geração; assim, os descendentes híbridos exibirão apenas a característica dominante no fenótipo); 2) lei da segregação (uma única característica associada a um único gene é herdada); e 3) lei da variação independente (os alelos de dois ou mais genes diferentes são distribuídos para os gametas independentemente um do outro).

- Teoria: um sistema de leis
 - teorias axiomáticas
 - teorias empíricas

- Teoria: um sistema de leis
 - teorias axiomáticas
 - teorias empíricas

Compreendemos ciência como um esforço sistemático de construir e organizar conhecimento na forma de explicações e predições testáveis sobre o universo.

Um sistema de leis é chamado de Teoria. A construção de uma teoria é principal objetivo de pesquisas científicas.

A filosofia da ciência distingue dois tipos de teorias: 1) as teorias axiomáticas da lógica e matemática (utiliza um sistema axiomático para construir verdades analíticas); 2) as teorias empíricas da ciência factual (faz afirmações sobre partes do mundo; a veracidade dependerá da correção interna e correspondência com fatos da realidade; teorias empíricas devem também ter um núcleo axiomático).

- Linguística formal - matemática qualitativa (álgebra, teoria de conjuntos e lógica)

Exemplo para o português brasileiro:

$$\begin{bmatrix} +\text{silábico} \\ +\text{alto} \\ -\text{arredondado} \end{bmatrix} \rightarrow \emptyset / \# _ \begin{bmatrix} +\text{consonantal} \\ +\text{contínuo} \\ +\text{coronal} \end{bmatrix} \left[\begin{bmatrix} +\text{consonantal} \end{bmatrix} \right]$$

A vogal [i] pode ser apagada em início de palavra quando seguida de uma sibilante e outra consoante. Por exemplo: esperar, estragar, espelho, estante, etc.

- Linguística quantitativa (LQ) - propriedades quantitativas (quantidades, probabilidades e tendências)

Linguística Quantitativa

└─ Linguística Quantitativa enquanto ciência

└─ Linguística Quantitativa

- Linguística formal - matemática qualitativa (álgebra, teoria de conjuntos e lógica)

Exemplo para o português brasileiro:
 $\left[\begin{array}{l} \text{«vibração»} \\ \text{«alto»} \\ \text{«arredondado»} \end{array} \right] \rightarrow \text{O} / \# \left[\begin{array}{l} \text{«consonantal»} \\ \text{«contínuo»} \\ \text{«coronal»} \end{array} \right] \left\{ \begin{array}{l} \text{«consonantal»} \\ \text{«consonantal»} \end{array} \right\}$
 A vogal [i] pode ser apaga no início de palavra quando seguida de uma sílaba e
 outra consoante. Por exemplo: *espere*, *espelho*, *enxerto*, etc.

- Linguística quantitativa (LQ) - propriedades quantitativas (quantidades, probabilidades e tendências)

Linguística formal utiliza-se de princípios da matemática qualitativa (álgebra, teoria de conjuntos e lógica) para modelar propriedades estruturais da linguagem.

A linguística quantitativa (LQ) estuda as diversas propriedades quantitativas, essenciais para a descrição e compreensão dos sistemas linguísticos. A LQ lida com quantidades, probabilidades e tendências, analisando comprimento, frequência, distância, grau de polissemia, idade, etc. As propriedades de elementos linguísticos e suas inter-relações são formuladas matematicamente, estabelecendo leis estocásticas (ou seja, não capturam casos singulares, mas predizem eventos e condições gerais). A abordagem quantitativa permite uma descrição mais adequada da realidade, permitindo a distinção em diversos níveis ao invés de uma distinção extrema em apenas dois extremos (sim/não).

- George Kingley Zipf - a lei de Zipf
 - relação entre ordem (rank) e frequência – princípio de auto-organização e economicidade
 - “The Psycho-Biology of Language. An Introduction to Dynamic Philology” (1935)
 - “Human Behavior and the Principle of Least Effort” (1949)

Linguística Quantitativa

└─Linguística Quantitativa enquanto ciência

└─Surgimento da linguística quantitativa

- George Kingley Zipf - a lei de Zipf
 - relação entre ordem (rank) e frequência - princípio de auto-organização e economicidade
- "The Psycho-Biology of Language: An Introduction to Dynamic Philology" (1935)
- "Human Behavior and the Principle of Least Effort" (1949)

Embora diversos trabalho abordasse a quantificação de unidades linguísticas, antes mesmo no século XIX, George Kingley Zipf é considerado o pai da linguística quantitativa por ter sistematicamente investigado e criado um modelo teórico para explicar suas observações e propor uma formulação matemática.

Zipf propôs as ideias de

- princípio do esforço mínimo (esforço mínimo individual e esforço mínimo coletivo)
- forças de unificação e diversificação

Para o falante, o princípio da economia busca uma maior unificação e simplificação (o que dificulta a tarefa do ouvinte). Para o ouvinte, o princípio da economia busca diversificação (o que simplifica sua tarefa de decodificação da mensagem, mas dificulta a tarefa do falante).

Leis da Linguística Quantitativa

Leis da Linguística Quantitativa

Leis da Linguística Quantitativa

- Leis de distribuição
- Leis funcionais
- Leis de desenvolvimento

(Köhler 2014)

Linguística Quantitativa

└ Leis da Linguística Quantitativa

└ Leis da Linguística Quantitativa

└ Leis da Linguística Quantitativa

- Leis de distribuição
- Leis funcionais
- Leis de desenvolvimento

(Köhler 2014)

- As leis de distribuição fazem predição sobre o número observado de certas características. O exemplo mais conhecido é a lei de Zipf-Mandelbrot. Outros exemplos são leis sobre distribuição de comprimento, leis de polissemia, leis de sinônimos, leis de frequência de estruturas sintáticas, etc.
- As leis funcionais são leis que estabelecem relações entre duas ou mais propriedades. A lei de Menzerath, que relaciona o tamanho do constituinte com o respectivo construto, é um exemplo de lei funcional.
- As leis de desenvolvimento descrevem como certas características evoluem ao longo do tempo. O exemplo mais conhecido é a lei de Piotrowski, que descreve o crescimento ou decaimento de certas unidades ao longo do tempo.

Leis de distribuição

Leis de distribuição

Lei de Zipf

Zipf (1949) - Ulysses de James Joyce - 260.430 palavras (29.899 palavras diferentes).

Relação entre rank e frequência de ocorrência.

$$r \times f = C \quad (1)$$

Ao visualizar tal relação em um gráfico log-log, esperamos ver uma reta com inclinação -1.

$$f = \frac{C}{r} \quad (2)$$

$$\log f = -\log r + \log C \quad (3)$$

Linguística Quantitativa

- Leis de distribuição
 - Lei de Zipf
 - Lei de Zipf

Zipf (1949) - *Ulysses* de James Joyce - 260.430 palavras (29.899 palavras diferentes).
Relação entre rank e frequência de ocorrência.

$$r \times f = C \quad (1)$$

Ao visualizar tal relação em um gráfico log-log, esperamos ver uma reta com inclinação -1.

$$f = \frac{C}{r} \quad (2)$$

$$\log f = -\log r + \log C \quad (3)$$

Zipf analisou o livro *Ulysses* de James Joyce. O texto possui 260.430 palavras, sendo 29.899 palavras diferentes.

- palavras foram consideradas diferentes se diferem-se foneticamente na forma flexionada que ocorrem no texto (desta forma, give, gives, gave, given, giving, giver e gift representam sete palavras diferentes e não uma única palavras com sete formas diferentes).
- Zipf observou uma correlação entre o número de diferentes palavras e sua frequência de uso, aproximando-se à equação de uma hipérbole equilátera¹ $y = 1/x$ (lei de Zipf inversa).

¹Dizemos que uma hipérbole é equilátera se o comprimento do eixo focal é igual ao comprimento do eixo não-focal.

I Rank (<i>r</i>)	II Frequency (<i>f</i>)	III Product of I and II (<i>r</i> × <i>f</i> = <i>C</i>)	IV Theoretical Length of Ulysses (<i>C</i> × 10)
10	2,653	26,530	265,500
20	1,311	26,220	262,200
30	926	27,780	277,800
40	717	28,680	286,800
50	556	27,800	278,800
100	265	26,500	265,000
200	133	26,600	266,000
300	84	25,200	252,000
400	62	24,800	248,000
500	50	25,000	250,000
1,000	26	26,000	260,000
2,000	12	24,000	240,000
3,000	8	24,000	240,000
4,000	6	24,000	240,000
5,000	5	25,000	250,000
10,000	2	20,000	200,000
20,000	1	20,000	200,000
29,899	1	29,899	298,990

Figura 2: Tabela com alguns exemplos de rank e frequência de ocorrência de palavras em Ulysses (Zipf 1949).

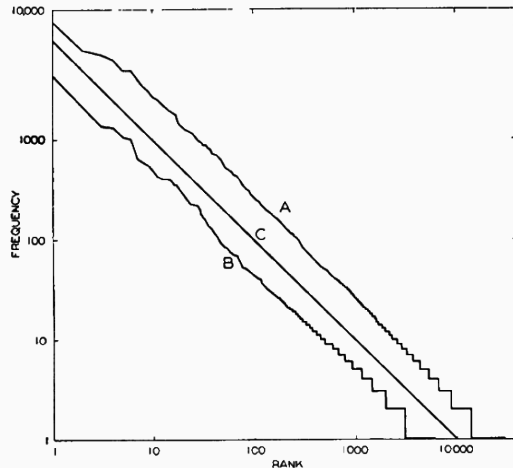


Figura 3: Distribuição rank-frequência para palavras no inglês. (A) dados de James Joyce; (B) dados de Eldridge (43.989 palavras de jornais, sendo 6.002 palavras diferentes); (C) curva ideal com inclinação igual a 1 (Zipf 1949).

Linguística Quantitativa

- Leis de distribuição
 - Lei de Zipf

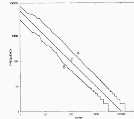


Figura 2: Distribuição rank-frequência para palavras no inglês. (A) dados de James Joyce; (B) dados de Elbridge (43.989 palavras de jornais, sendo 6.062 palavras diferentes); (C) curva ideal com inclinação igual a -1 (Zipf 1940).

Os degraus ao final representam palavras com baixa frequência de ocorrência. Como apenas é possível observarmos as palavras um número inteiro de vezes, o valor da relação dada na equação 1 é arredondado para o inteiro mais próximo. O último degrau representa 16.432 palavras que ocorreram uma única vez em todo o texto (hápax legómenon). O degrau anterior representa 4.776 palavras que ocorreram duas vezes (dis legomenon) e o anterior representa 2.194 palavras que ocorreram três vezes (tris legomenon).

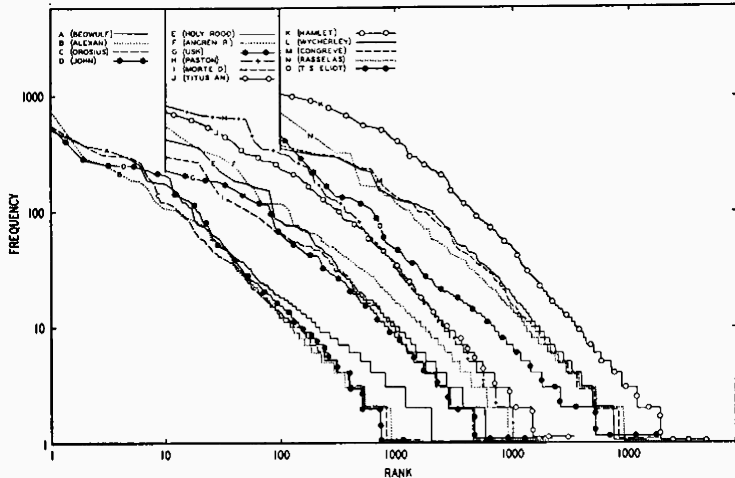


Figura 4: De Beowulf a T.S. Eliot. Distribuição rank-frequência de palavras em 15 escritores da língua inglesa, do inglês antigo ao atual. Os gráficos estão deslocados na abscissa para melhor visualização (Zipf 1949).

A r -ésima palavra mais frequente possui frequência de ocorrência $f(r)$ que varia da seguinte forma com r :

$$f \propto \frac{1}{r^\alpha}$$

onde temos $\alpha \approx 1$ (Zipf 1935, 1949).

A lei de Zipf pode também ser expressa por

$$f(r; \alpha, N) = \frac{1}{H_{N,\alpha} \cdot r^\alpha},$$

onde $H_{N,\alpha}$ é o N -ésimo número harmônico generalizado e N é o tamanho do vocabulário.

Linguística Quantitativa

- Leis de distribuição
 - Lei de Zipf
 - Lei de Zipf

A r -ésima palavra mais frequente possui frequência de ocorrência $f(r)$ que varia da seguinte forma com r :

$$f \propto \frac{1}{r^\alpha}$$

onde temos $\alpha \approx 1$ (Zipf 1935, 1949).

A lei de Zipf pode também ser expressa por

$$f(r; \alpha, N) = \frac{1}{H_{N, \alpha} \cdot r^\alpha},$$

onde $H_{N, \alpha}$ é o N -ésimo número harmônico generalizado e N é o tamanho do vocabulário.

Outros trabalhos mostram que o valor do expoente α está entre 0.6 e 1.5 para o inglês falado (Bian et al. 2016; Baixeries, Elvevåg, e Ferrer-i-Cancho 2013), entre 0.765 e 1.357 para traduções da bíblia em diversas línguas (Mehri e Jamaati 2017). O N -ésimo número harmônico é a soma dos recíprocos dos N primeiros números naturais:

$$H_N = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{N} = \sum_{k=1}^N \frac{1}{k}.$$

O N -ésimo número harmônico generalizado de ordem α é dada por

$$H_{N, \alpha} = \sum_{k=1}^N \frac{1}{k^\alpha}.$$

O limite para $N \rightarrow \infty$, quando $\alpha > 1$, converge para a função zeta de Riemann:

$$\lim_{N \rightarrow \infty} H_{N, \alpha} = \zeta(\alpha).$$

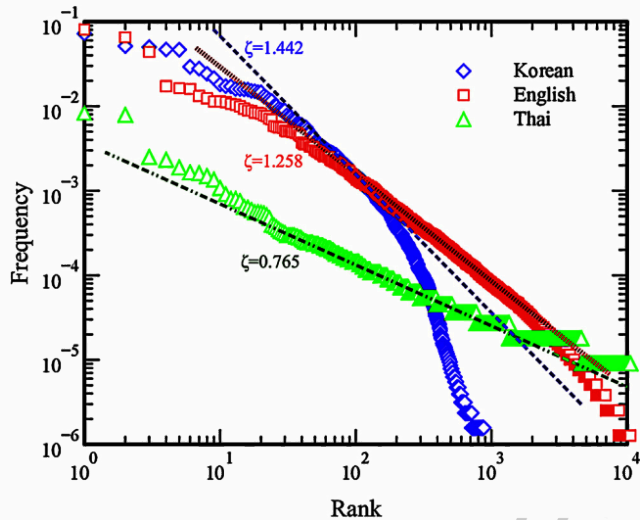


Figura 5: Gráfico Zipf para tradução da bíblia no coreano, inglês e tailandês (Mehri e Jamaati 2017).

Linguística Quantitativa

- Leis de distribuição
 - Lei de Zipf

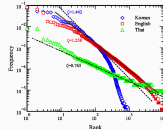


Figure 5: Gráfico Zipf para tradução da Bíblia no coreano, inglês e tailandês (Mehri e Jamaati 2017).

Em seu trabalho, Mehri e Jamaati (2017) fazem o ajuste da distribuição de Zipf para a tradução da Bíblia em 100 línguas diferentes. Mehri e Jamaati (2017) enfatizam que estruturas sintáticas distintas levam a expoentes de Zipf diferentes e analisam o comportamento do coeficiente de Zipf em diferentes famílias linguísticas.

Inglês e Coreano possuem coeficiente de Zipf próximos, enquanto o tailandês possui coeficiente bem diferente, o que pode ser explicado pela sua estrutura gramatical distinta. A lei de Zipf não é compatível com dados empíricos de palavras de tipo raro. Benoit Mandelbrot (1966) generalizou a lei de Zipf para superar tal limitação.

Esta relação de potência é também observada em outros diferentes fenômenos, tais como:

- magnitude de terremotos (Abe e Suzuki 2005);
- população de cidades (Gabaix 1999);
- variações de preços (Benoît Mandelbrot 1963);
- distribuição do passivo total de empresas falidas (Fujiwara 2004);
- expressão gênica (Furusawa e Kaneko 2003);
- sistemas dinâmicos caóticos (Nicolis, Nicolis, e Nicolis 1989);
- magnitude de avalanches (Bak 1996);
- tráfegos de dados na Internet (Crovella e Bestavros 1996);
- requisições de páginas web (Adamic e Huberman 2002);
- número de citações de artigos científicos (Solla Price 1965);
- tamanho das famílias linguísticas (Wichmann 2005);
- tiragem de livros e discos (Kohli e Sah 2003; Cox, Felton, e Chung 1995);

e muitos outros.

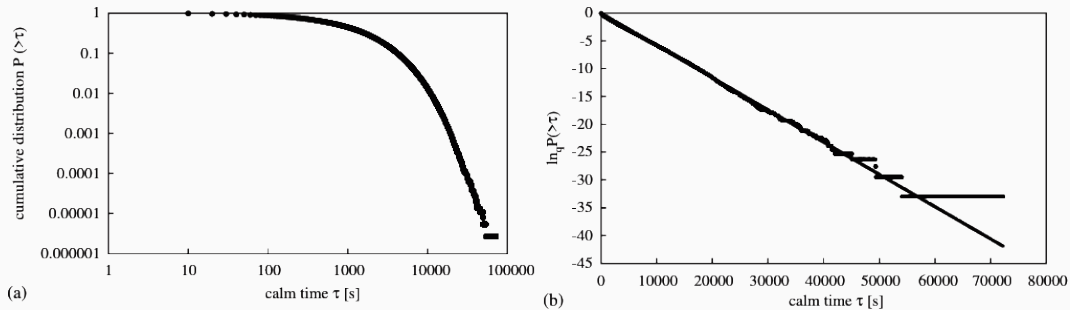


Figura 6: Distribuição cumulativa dos tempos calmos (entre terremotos) na Califórnia (Abe e Suzuki 2005).

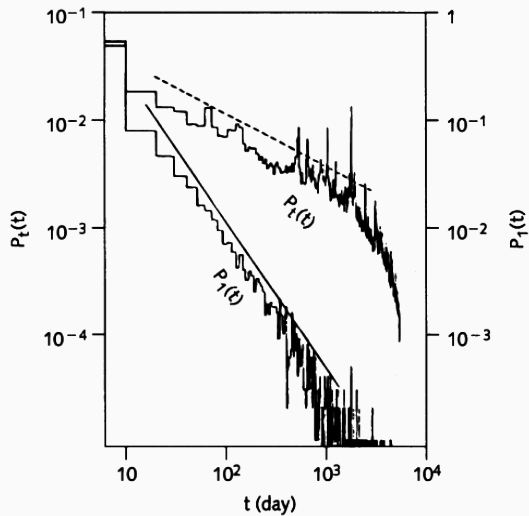


Figura 7: Distribuição do tempo entre terremotos (tempo de primeiro retorno e todos) (Bak 1996; Ito e Matsuzaki 1990).

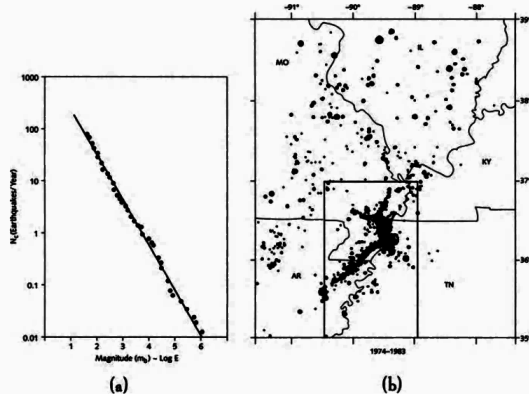


Figura 8: Distribuição da magnitude de terremotos. Os pontos representam o número de terremotos N_c com magnitude maior que uma determinada magnitude m . A figura (b) apresenta a localização dos terremotos em Nova Madrid (Missouri) nos anos de 1974 a 1983. O tamanho dos pontos representa a magnitude dos terremotos (Bak 1996).

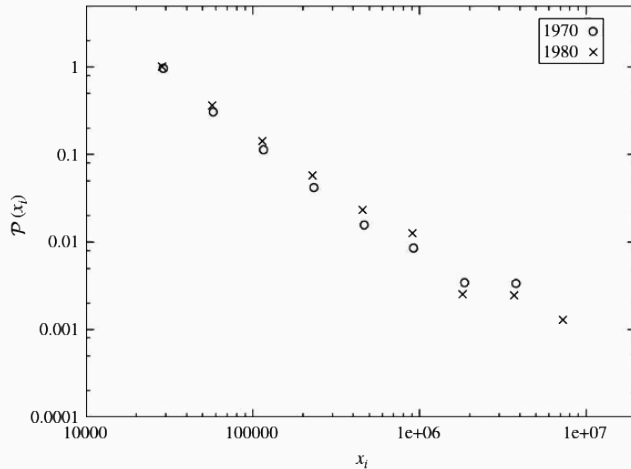


Figura 9: Distribuição cumulativa $P(x_i)$ das cidades brasileiras com população x_i (Moura Jr e Ribeiro 2006).

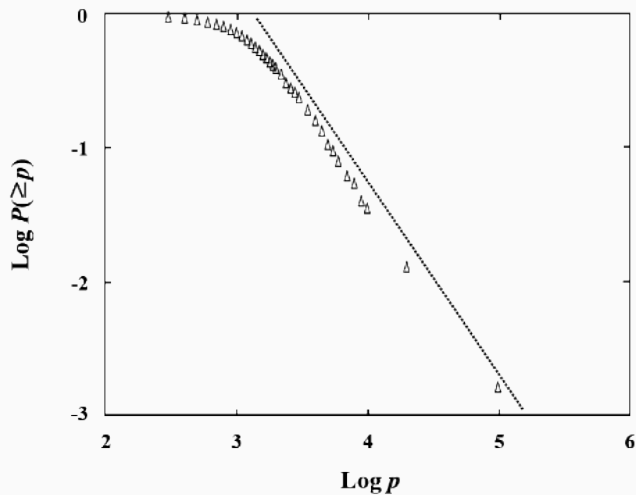


Figura 10: Distribuição cumulativa dos preços de ações negociadas no KOSDAQ (Korean Securities Dealers Automated Quotations) (Choi et al. 2005).

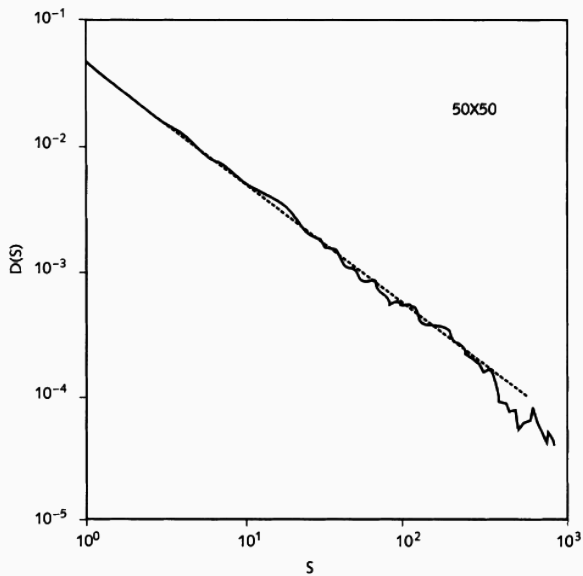
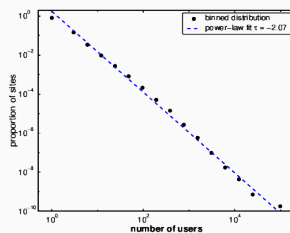
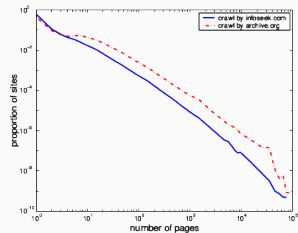
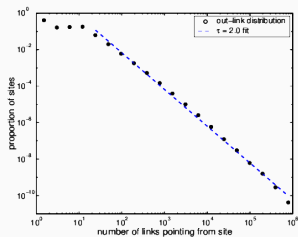


Figura 11: Distribuição de avalanches (s representa o tamanho das avalanches) (Bak 1996).



c)



d)

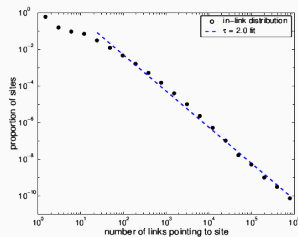


Figura 12: Distribuição de lei de potência para o número de a) páginas, b) visitantes, c) links externos e d) links internos de um site (rastreadas pelo infoseek.com e archive.com em 1997) (Adamic e Huberman 2002).

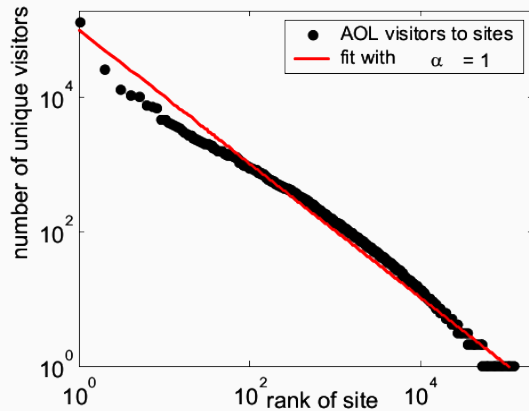


Figura 13: Sites ordenados pelo número de visitantes AOL únicos recebidos em 01/12/1997 (Adamic e Huberman 2002).

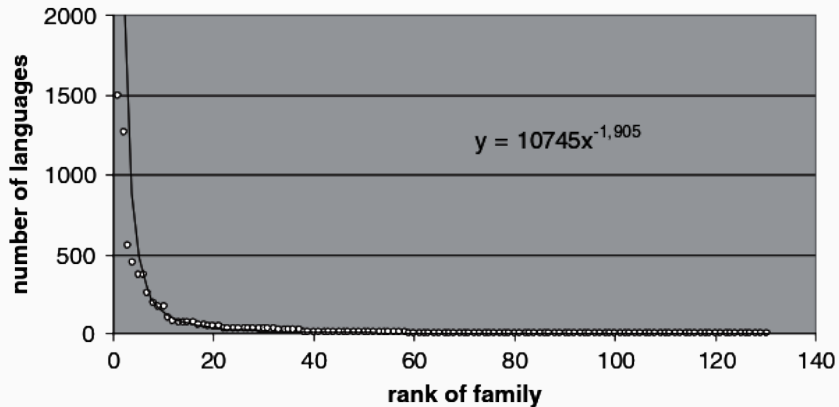


Figura 14: Tamanho das famílias linguísticas, segundo Grimes (2000; Wichmann 2005).

Benoît Mandelbrot (1963) propôs um deslocamento para levar em consideração o achatamento observado na região do baixo ranque

$$f \propto \frac{1}{(r + \beta)^\alpha}$$

em que $\alpha \approx 1$ e $\beta \approx 2.7$ (Zipf 1935, 1949; Benoît Mandelbrot 1963).

Por que as línguas seguem a lei de Zipf? i

Algumas possíveis explicações:

- processos aleatórios concatenativos (Conrad e Mitzenmacher 2004; Li 1992; Miller 1957)
- mistura de distribuições exponenciais (Farmer e Geanakoplos 2008)
- invariância à escala (Chater e Brown 1999)
- otimização (com restrição) da entropia (B. B. Mandelbrot 1953)
- otimização da informação de Fisher (Hernando et al. 2009)
- invariância das leis de potência sob agregação (Farmer e Geanakoplos 2008)
- processos estocásticos multiplicativos (Mitzenmacher 2004)
- reuso preferencial (Simon 1955; Yule 1944)
- descrição simbólica de sistemas complexos de processos estocásticos (Corominas-Murtra e Solé 2010)

Linguística Quantitativa

└ Leis de distribuição

└ Lei de Zipf

└ Por que as línguas seguem a lei de Zipf?

Algumas possíveis explicações:

- processos aleatórios concatenativos (Conrad e Mitzemacher 2004; Li 1962; Miller 1957)
- mistura de distribuições exponenciais (Farmer e Geanakoplos 2008)
- invariância à escala (Chater e Brown 1999)
- otimização (com restrição) da entropia (B. B. Mandelbrot 1953)
- otimização da informação de Fisher (Hernando et al. 2009)
- invariância das leis de potência sob agregação (Farmer e Geanakoplos 2008)
- processos estocásticos multiplicativos (Mitzemacher 2004)
- reuso preferencial (Simon 1955; Yule 1944)
- descrição simbólica de sistemas complexos de processos estocásticos (Corominas-Murtra e Solé 2010)

É razoável esperar que as palavras não sejam igualmente distribuídas, entretanto, dado que as palavras tem frequências distintas, por qual razão seguiriam uma regra matemática particular? Sobretudo, por que seguiriam uma regra que não leva em consideração o significado ou função sintática das palavras?

Por que as línguas seguem a lei de Zipf? ii

- passeio aleatório em escalas logarítmicas (Kawamura e Hatano 2002)
- organização semântica (Guiraud 1968; Manin 2008)
- otimização da comunicação (Cancho e Solé 2003; Cancho 2005; R. Ferrer-i-Cancho 2005; Benoît Mandelbrot 1963; Salge et al. 2015; Zipf 1935, 1949)
- divisão aleatória de elementos em grupos (Baek, Bernhardsson, e Minnhagen 2011)
- aproximação de primeira e segunda ordem de distribuições comuns (normal, por exemplo) (Baek, Bernhardsson, e Minnhagen 2011)
- busca otimizada em memória (Parker-Rhodes e Joyce 1956)
- etc.

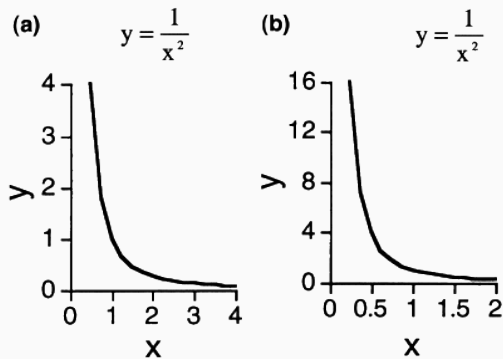


Figura 15: Conceito de invariância à escala. A mesma função é apresentada em diferentes escalas (Chater e Brown 1999).

Linguística Quantitativa

└ Leis de distribuição

└ Lei de Zipf

└ Invariância à escala

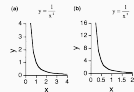


Figura 10: Conceito de invariância à escala. A mesma função é apresentada em diferentes escalas (Chater e Brown 1999).

Para Chater e Brown (1999), o sistema perceptual-motor reflete e preserva as características de invariância à escala presentes em vários aspectos ambientais. Isto permite o surgimento de diversas leis da psicologia governando percepção e ação em vários domínios e espécies (exemplos: lei de Weber-Fechner, lei de Stevens, lei de Fitts e lei de Piéron).

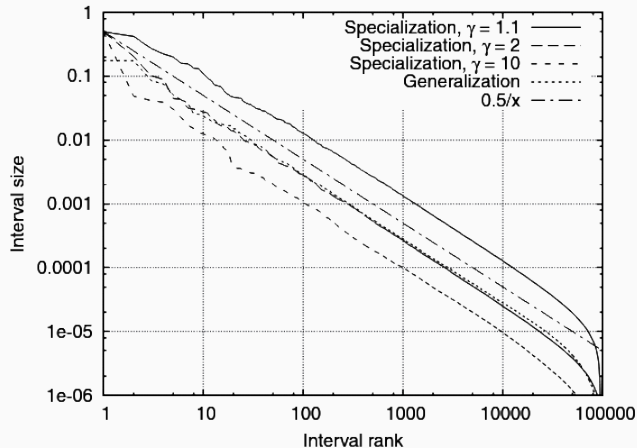


Figura 16: Lei de Zipf gerada por modelos de especialização e generalização. O parâmetro γ determina o quanto duas palavras podem diferir em extensão e ainda competir entre elas (Manin 2008).

Linguística Quantitativa

└ Leis de distribuição

└ Lei de Zipf

└ Evitando sinonímia excessiva

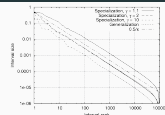


Figura 16: Lei de Zipf gerada por modelos de especialização e generalização. O parâmetro γ determina o quanto duas palavras podem diferir em extensão e ainda competir entre elas (Manin 2008).

Manin (2008) sugere que a lei de Zipf é resultante da organização hierárquica dos significados de palavras no espaço semântico. Manin (2008) parte da proposta de matriz semântica de Guiraud (1968) em que o significado de uma palavra é representado pela superposição de significados elementares. Manin (2008) propõe um modelo em que os significados de palavras são associados a intervalos numéricos e estão sujeito ao processo de generalização e especialização, sendo regidos por regras simples. Manin (2008) mostra que este modelo simples leva à distribuição de Zipf.

Os significados de palavras são associados a subintervalos do intervalo $S = [0, 1]$. **Modelo de generalização:** Se não estiverem congelados, os intervalos vão crescendo paulatinamente por uma quantidade δ . Quando surgir uma sobreposição de intervalos, um deles será escolhido aleatoriamente e será congelado. **Modelo de especialização:** Se dois intervalos, r_i e r_j , se interceptam e seus comprimentos, l_i e l_j , satisfazem $1/\gamma < l_i/l_j < \gamma$, diminui-se o menor intervalo pelo tamanho da sobreposição (competição entre os significados de palavras).

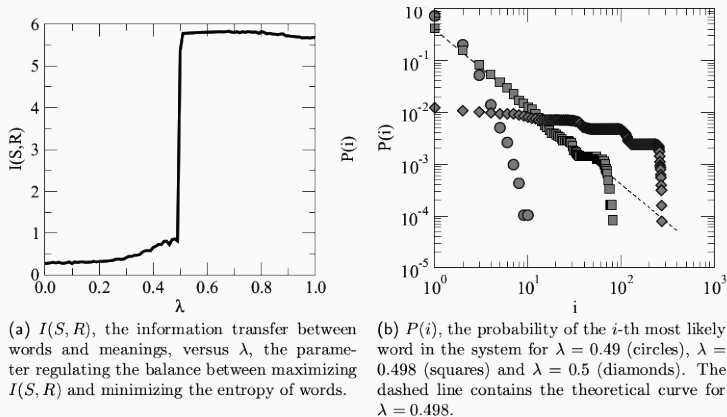


Figura 17: Resultado de um modelo computacional onde as probabilidades dos significados são governadas por estruturas internas do sistema de comunicação. Função minimizada:

$$\Omega(\lambda) = -\lambda I(S, R) + (1 - \lambda)H(S) \text{ (Cancho e Solé 2003; Cancho 2007).}$$

Linguística Quantitativa

└ Leis de distribuição

└ Lei de Zipf

└ Maximização da informação mútua

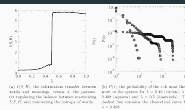


Figura 17: Resultado de um modelo computacional onde as probabilidades dos significados são governadas por estruturas internas do sistema de comunicação. Função minimizada: $\Omega(\lambda) = -\lambda I(S, R) + (1 - \lambda)H(S)$ (Cancho e Solé 2003; Cancho 2007).

As diversas línguas diferem-se muito, mas todos tem em comum o fato de serem utilizadas para a comunicação. Um sistema de comunicação confiável deve maximizar a transferência de informação. Além disto, a comunicação falada é um processo cognitivo e, portanto, busca-se economia de energia (para o falante e ouvinte). Cancho e Solé (2003) utilizam um modelo em que a comunicação visa à maximização da transferência de informação e a minimização do custo energético do uso das palavras (entropia).

Cancho e Solé (2003); Cancho (2007) propõem uma função Ω que deve ser minimizada pelo sistema de comunicação. Minimizar esta função será um balanço entre a maximização da transferência de comunicação $(I(S, R))^2$ e minimizar o custo da comunicação $(H(S))^3$. A função definida é $\Omega(\lambda) = -\lambda I(S, R) + (1 - \lambda)H(S)$, onde o parâmetro $\lambda \in [0, 1]$ controla o balanço entre a transferência de informação e o custo. Utilizou-se um algoritmo de Monte Carlo para realizar a minimização e encontrou-se o valor crítico de $\lambda = \lambda^* = 1/2 - \epsilon$, onde ϵ é um valor positivo pequeno ($\epsilon \approx 0.002$ na figura 17). A lei de Zipf ocorre na transição abrupta observada em $I(S, R)$, com $\lambda \approx 1/2$.

²Informação mútua entre o sinal e o estímulo.

³Entropia associada ao sinal.

Leis de distribuição

Significado

No balanço entre forças de unificação e diversificação esperamos encontrar palavras que possuam alguns significados.

- F_r : frequência da r -ésima palavra mais frequente
- m_r : número de significados da r -ésima palavras mais frequente
- f_r : frequência média de ocorrência dos m_r significados

$$m_r \times f_r = F_r$$

- forças de unificação: $\uparrow m_r, \downarrow f_r$
- forças de diversificação: $\downarrow m_r, \uparrow f_r$

Linguística Quantitativa

- Leis de distribuição
- Significado
- Significado

No balanço entre forças de unificação e diversificação esperamos encontrar palavras que possuam alguns significados.

- F_r : frequência da r -ésima palavra mais frequente
- m_r : número de significados da r -ésima palavras mais frequente
- f_r : frequência média de ocorrência dos m_r significados

$$m_r \times f_r = F_r$$

- forças de unificação: $\uparrow m_r, \downarrow f_r$
- forças de diversificação: $\downarrow m_r, \uparrow f_r$

A princípio, não sabemos qual é o peso dessas duas forças (unificação e diversificação), porém, pela relação entre o número de palavras distintas em uma amostra e suas respectivas frequências de ocorrência, suspeitamos que as forças de unificação e diversificação estabelecem, em geral, uma relação hiperbólica. Desta forma, resulta-se em m_r e f_r estarem também em uma relação hiperbólica, levado a obtermos $m_r = f_r = \sqrt{F_r}$.

Uma relação hiperbólica entre as forças leva a

$$m_r = f_r = \sqrt{F_r}$$

Em um gráfico log-log da distribuição de frequência dos significados das palavras, esperamos observar uma reta com inclinação -0.5 (Zipf 1945, 1949).

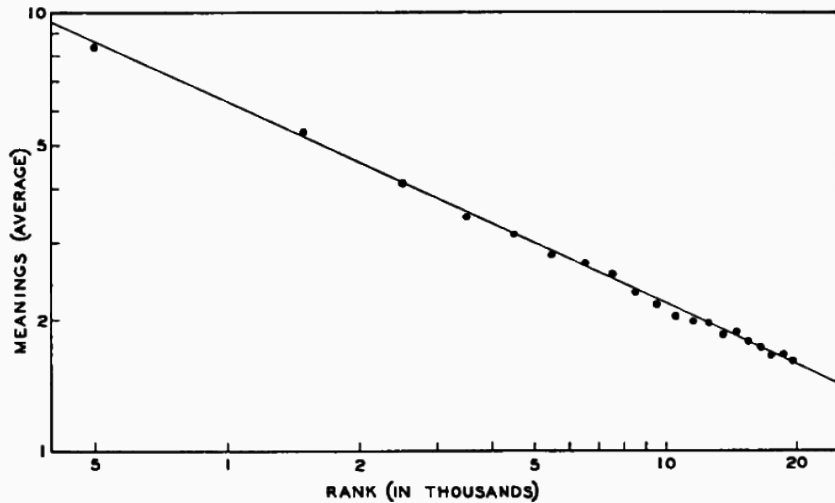


Figura 18: Distribuição de frequência dos significados das palavras (Zipf 1945, 1949).

Linguística Quantitativa

- Leis de distribuição
- Significado

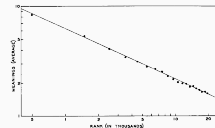


Figure 18: Distribuição de frequência dos significados das palavras (Zipf 1945, 1949).

Zipf (1949) utilizou os dados de E. L. Thorndike para obter a distribuição de frequência de ocorrência das palavras (corpus de 10 milhões de palavras) e o Thorndike-Century Dictionary para obter os m distintos significados de cada palavra.

Leis de distribuição

Lei da brevidade

Lei de abreviação/brevidade de Zipf

“a magnitude das palavras apresenta uma relação inversa ao número de ocorrências” (Zipf 1935)

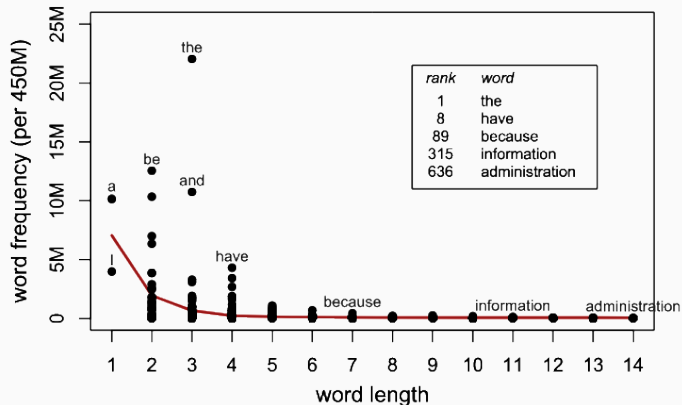


Figura 19: As 1000 palavras mais frequentes no inglês (COCA corpus) (Kanwal et al. 2017).

Linguística Quantitativa

└ Leis de distribuição

└└ Lei da brevidade

└└└ Lei de abreviação/brevidade de Zipf

"a magnitude das palavras apresenta uma relação inversa ao número de ocorrências" (Zipf 1935)

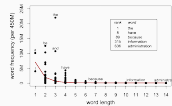


Figura 19: As 1000 palavras mais frequentes no inglês (COCA corpus) (Kernell et al. 2007).

Zipf (1935) supôs que tal padrão seria resultante da relação de compromisso entre uma comunicação sem erros e um código eficiente (menor esforço). Como as línguas utilizam-se de um inventário finito de símbolos discretos para formar palavras, o número de palavras possíveis para um dado comprimento é limitado. Em palavras curtas, há menos espaço para redundâncias, o que acarreta um menor potencial de distinguibilidade entre elas. A solução é associar às palavras curtas os significados mais frequentes e às palavras longas os significados menos frequentes, abordagem similar ao código de Huffman (1952).

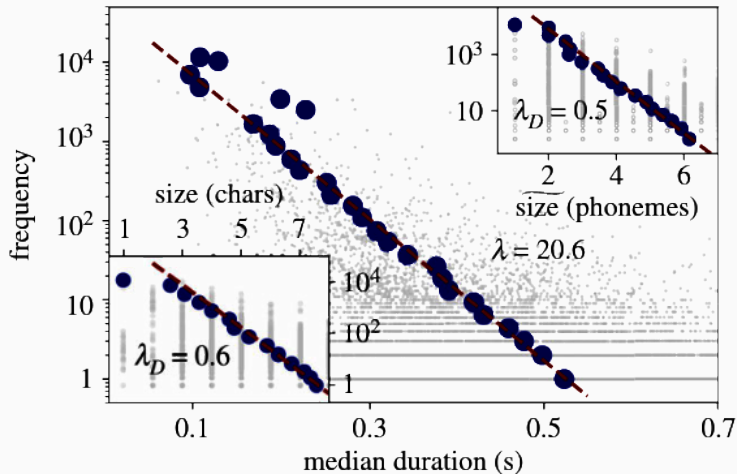


Figura 20: Lei da brevidade para palavras (duração, número de fonemas e número de caracteres) (Torre et al. 2019).

Linguística Quantitativa

- Leis de distribuição
- Lei da brevidade

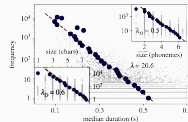


Figura 20: Lei da brevidade para palavras (duração, número de fonemas e número de caracteres) (Torre et al. 2019).

Para palavras, Torre et al. (2019) consideram 3 casos para analisar a lei de brevidade: (1) tendência das palavras mais frequentes serem constituídas por um menor número de caracteres; (2) tendência das palavras mais frequentes serem constituídas por um menor número de fonemas; (3) tendência das palavras mais frequentes serem articuladas pelos falantes em um menor intervalo de tempo.

Os pontos cinza na figura 20 apresentam o espalhamento das palavras em relação ao tempo mediano de duração (em segundos) e a frequência de ocorrência das palavras. Os pontos azuis são gerados através de agrupamento logarítmico nas frequências. O gráfico superior à direita apresenta o mesmo tipo de relação, porém considerando a mediana do número de fonemas e o gráfico inferior à esquerda utiliza o número de caracteres.

Torre et al. (2019) utilizou o corpus Buckeye contendo fala de conversação de falantes nativos de inglês, contendo aproximadamente 8×10^5 fonemas, 3×10^5 palavras e 5×10^4 grupos respiratórios⁴ (breath-groups).

⁴Grupo respiratório é uma sequência de sons articulados no decorrer de uma única expiração

Leis de distribuição

Lei da polissemia

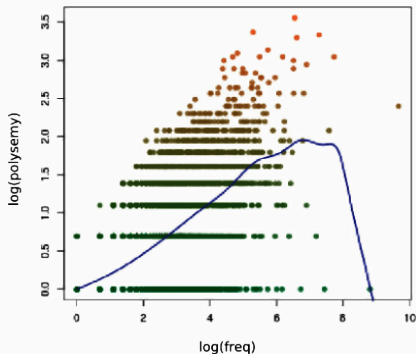


Figura 21: Relação entre frequência e polissemia. A cor verde indica a densidade de pontos (mais escuro, maior densidade). A linha azul é a regressão não-paramétrica. Dados do corpus SemCor (Hernández-Fernández et al. 2016).

Linguística Quantitativa

- Leis de distribuição
 - Lei da polissemia
 - Lei da polissemia

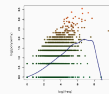


Figura 21: Relação entre frequência e polissemia. A cor verde indica a densidade de pontos (mais escura, maior densidade). A linha azul é a regressão não-paramétrica. Dados do corpus SemCor (Herrández-Fernández et al. 2016).

SemCor é um corpus criado pelo Universidade de Princeton, composto por 352 textos, sendo estes um subconjunto do corpus Brown para o inglês.

A tendência observada pela regressão não paramétrica não é válida para regiões extremas do gráfico, onde a densidade de pontos é muito pequena.

Vários trabalhos analisaram as leis de polissemia e brevidade de Zipf (Zipf 1935, 1945; Hernández-Fernández et al. 2016; Ilgen e Karaoglan 2007; Ramon Ferrer-i-Cancho e Vitevitch 2018; Kanwal et al. 2017; Tomaschek et al. 2017; Bentz e Ferrer Cancho 2016; Piantadosi, Tily, e Gibson 2011; Ramon Ferrer-i-Cancho et al. 2013; Strauss, Grzybek, e Altmann 2007; Sigurd, Eeg-Olofsson, e Van Weijer 2004; Teahan et al. 2000).

Linguística Quantitativa

└ Leis de distribuição

└ Lei da polissemia

└ Leis de polissemia e brevidade de Zipf

Vários trabalhos analisaram as leis de polissemia e brevidade de Zipf (Zipf 1935, 1946; Hernández-Fernández et al. 2016; Ilgen e Karacoglan 2007; Ramon Ferrer-i-Cancho e Vitevitch 2018; Kamwal et al. 2017; Tomaschek et al. 2017; Bentz e Ferrer Cancho 2016; Plantadosi, Tily, e Gibson 2017; Ramon Ferrer-i-Cancho et al. 2013; Strauss, Grzybicki, e Altmann 2007; Sigurd, Eeg-Olofsson, e Van Weijer 2004; Teahan et al. 2000).

Estas leis são presumidamente universais por serem verificadas em todas as línguas em que foram testadas até o momento. Alguns estudos buscam modelar suas origens e traçar mecanismos abstratos ou princípios linguísticos que suportem sua universalidade. Como exemplo temos o trabalho de Ramon Ferrer-i-Cancho et al. (2013), em que busca-se argumentar que a compressão é um princípio geral na comunicação animal, refletindo uma codificação eficiente.

Leis de distribuição

Lei da Lognormalidade

Diversos estudos observam consistentemente uma distribuição lognormal para unidades da fala (fonemas, palavras e grupo respiratório) (Herdan 1958; Sayli 2002; Rosen 2005; Gopinath, Veena, e Nair 2008; Shaw e Kawahara 2019; Hernández-Fernández, G. Torre, et al. 2019; Torre et al. 2019).

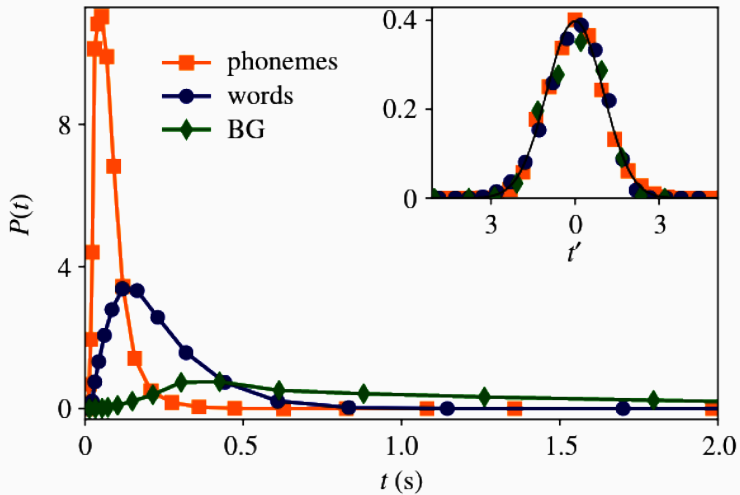


Figura 22: Distribuição lognormal na duração de fonemas, palavras e grupos respiratórios no inglês (Torre et al. 2019).

Linguística Quantitativa

- Leis de distribuição
 - Lei da Lognormalidade

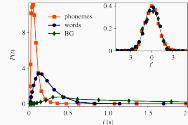


Figura 22: Distribuição lognormal na duração de fonemas, palavras e grupos respiratórios no inglês (Torre et al. 2016)

Dizemos que uma v.a. X possui distribuição lognormal se o logaritmo dela, $Y = \ln(X)$, possui distribuição normal. O gráfico menor da figura 22, no canto superior direito, apresenta o escalonamento dos valores para verificar que de fato seguem uma distribuição normal.

Leis de distribuição

Lei de Heaps

A lei de Heaps/Herdan descreve o crescimento do vocabulário em um texto.

$$V(n) = Kn^{\beta}, \quad \beta < 1$$

K tipicamente está entre 10 e 100, e β entre 0.4 e 0.6.

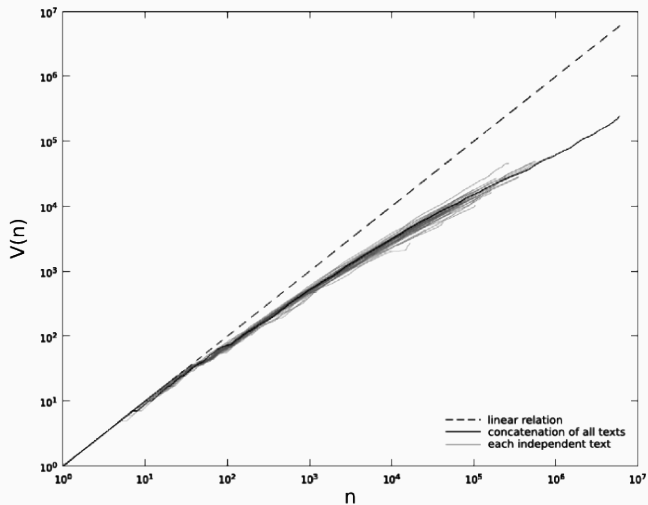


Figura 23: Crescimento do vocabulário em 35 livros do projeto Gutenberg (Araújo 2013).

Leijenhorst, Weide, e Grootjen (2005) mostram que é possível derivar matematicamente a lei de Heaps a partir da lei de Zipf. Neste caso, teremos $\beta = 1/\alpha$, sendo necessário $\alpha > 1$.

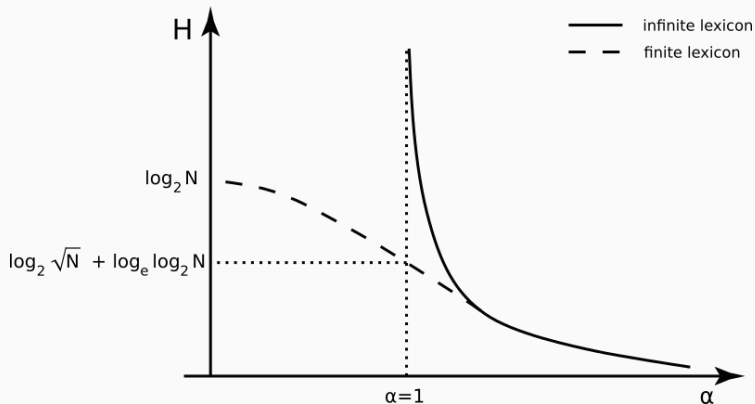


Figura 24: Entropia para uma fonte com distribuição de Zipf. Comparação entre léxico finito e infinito (Benoit Mandelbrot 1953; Araújo, Silva, e Yehia 2013).

Crescimento do vocabulário e das classes de baixa frequência

A lei de potência proposta por Altmann $y = Ax^{-b}$ descreve bem a relação entre crescimento do vocabulário e tamanho das classes.

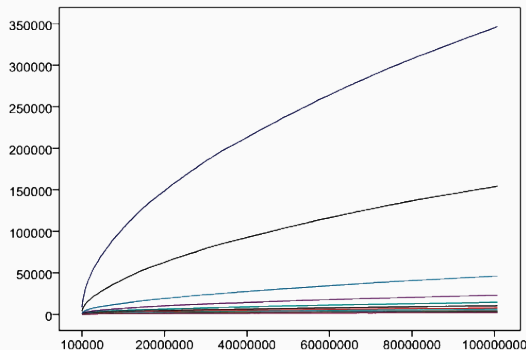


Figura 25: Crescimento do vocabulário e das classes de baixa frequência (1 a 15) (Fan, Wang, e Gao 2014).

Leis de distribuição

Lei de Zipf inversa

Zipf (1935) estabelece a lei inversa, relacionando a frequência de ocorrência e o número de palavras para uma dada frequência.

$$N_f = af^{-b}$$

onde f é a frequência de ocorrência e N_f o número de palavras com uma dada frequência de ocorrência f .

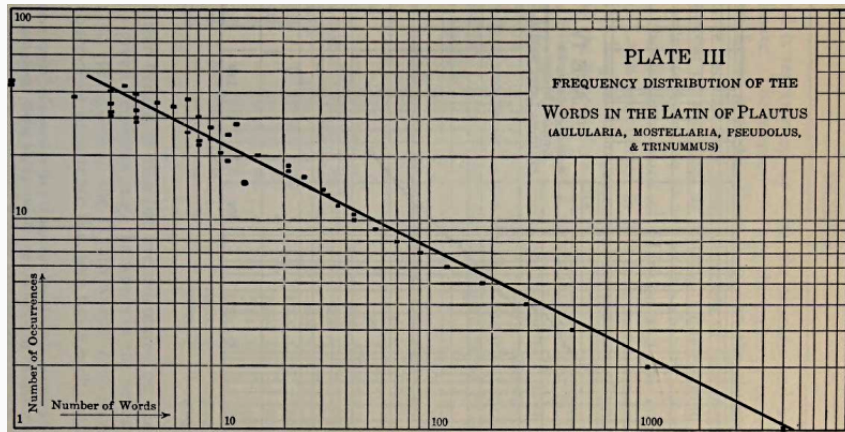
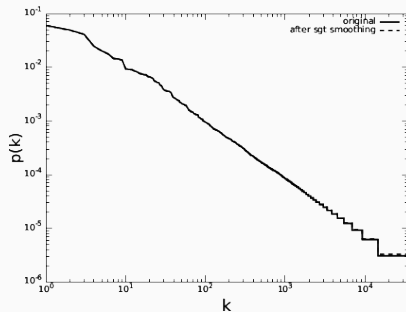
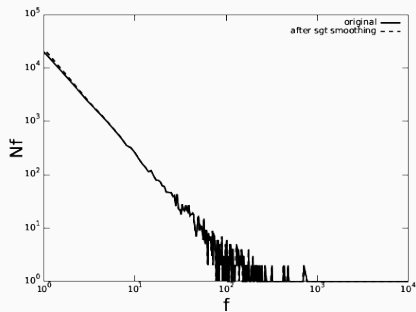


Figura 26: Relação entre frequência de ocorrência e número de palavras para uma dada frequência (Zipf 1935).



(a) Zipf plot



(b) Inverse Zipf plot

Figura 27: Gráfico de Zipf e gráfico inverso para o texto Ulysses (Araújo 2013).

Leis de distribuição

Hapax Legomena

O número de hapax legomena algumas vezes é utilizado como uma medida de riqueza de vocabulário.

O número de hapax legomena (HL) e o tamanho do vocabulário (V) apresentam uma relação linear

$$HL = aV - b$$

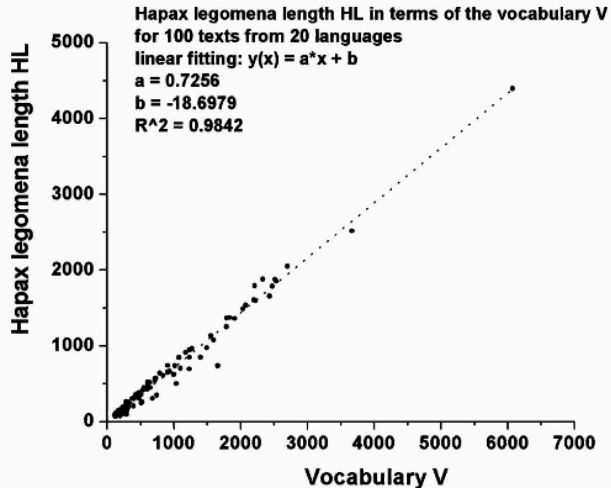


Figura 28: Dependência entre o número de hapax legomena (HL) e o tamanho do vocabulário (V) (Popescu e Altmann 2008).

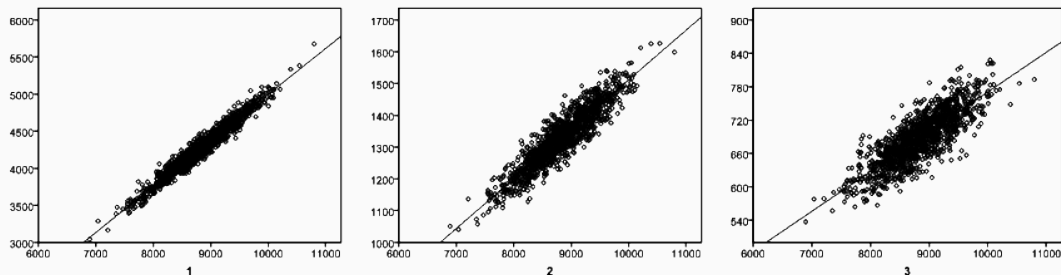


Figura 29: Relação entre o tamanho do vocabulário e o tamanho das classes de baixa frequência (1, 2 e 3). Para esta análise, o British National Corpus foi dividido em 1.000 pedaços de aproximadamente 100.000 palavras. (Fan, Wang, e Gao 2014).

A razão entre o tamanho do vocabulário e o número de hapaxes foi objeto de estudo de diversos linguistas (H. Baayen 1996; Tweedie e Baayen 1998; R. H. Baayen 2001; Kornai 2002; Fengxiang 2010).

Leis de distribuição

Lei de Frumkina

A lei de Frumkina (lei dos blocos de texto) descreve a frequência de ocorrência de certas unidades linguísticas em blocos de texto. A distribuição hipergeométrica negativa é um bom modelo para descrever a ocorrência (Altmann e Burdinski 1982; Karl-Heinz Best 2005; Wikipedia 2021b).

Tabelle 1
Die Verteilung von <a> in Textblöcken

x	<a> in 50-Wort-Blöcken		<a> in 100-Wort-Blöcken	
	n_x	NP_x	n_x	NP_x
0	12	12.09		
1	27	27.95	4	3.82
2	36	34.36	6	7.26
3	31	28.60	10	9.63
4	13	17.19	14	10.74
5	8	7.41	9	10.58
6	2	2.11	7	9.31
7	1	0.30	8	7.18
8			5	4.55
9			2	1.93
	$K = 17.5773$ $M = 5.8527$ $n = 7$ $\chi^2 = 1.53$ $FG = 3$ $P = 0.68$		$K = 5.0624$ $M = 2.3170$ $n = 8$ $\chi^2 = 2.18$ $FG = 5$ $P = 0.82$	

Figura 30: Ocorrências de ‘a’ no capítulo 1 de *Die Bäder von Lucca* (Heinrich Heine) com blocos de tamanho 50 e 100. x : número de ocorrências da letra ‘a’ nos blocos de texto; n_x : número observado de blocos de texto com x ocorrências de ‘a’; NP_x : número de blocos de texto com x ocorrências de ‘a’, calculado de acordo com a distribuição hipergeométrica negativa (Karl-Heinz Best 2005).

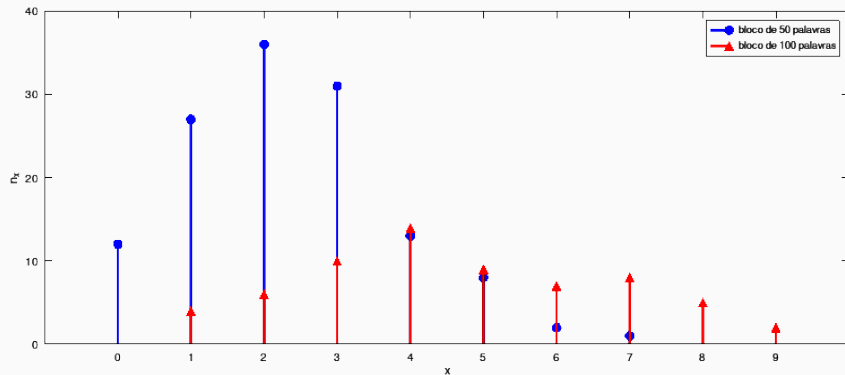


Figura 31: x : número de ocorrências da letra ‘a’ nos blocos de texto; n_x : número observado de blocos de texto com x ocorrências de ‘a’.

Leis de distribuição

Lei de Martin

A lei de Martin diz respeito à estruturação hierárquica do léxico.

Exemplo:

Sessel(1) – Sitzmöbel(2) – Möbel(3) – Einrichtungsgegenstand(4) – Gegenstand(5)

(Poltrona - móveis de assento - móveis - item de decoração - objeto)

Linguística Quantitativa

- Leis de distribuição
 - Lei de Martin
 - Lei de Martin

A lei de Martin diz respeito à estruturação hierárquica do léxico.

Exemplo:

Sessel(1) – Sitzmöbel(2) – Möbel(3) – Einrichtungsgegenstand(4) – Gegenstand(5)

(Poltrona - móveis de assento - móveis - item de decoração - objeto)

A posição que uma palavra ocupa na cadeia lexical diz respeito a quantas definições são ligadas a ela, indo do termo mais específico ao mais geral, formando uma hierarquia de significados cada vez mais abrangentes. O número de definições diminui com o aumento da generalidade.

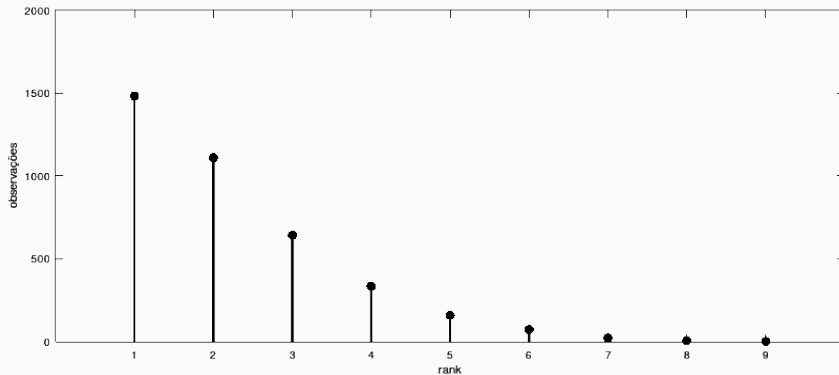


Figura 32: A distribuição de Poisson mista é um bom modelo para os dados de Bagheri (2002; Wikipedia 2021a)

Leis funcionais

Leis funcionais

Lei de Menzerath

Menzerath (1954) observou a existência de uma relação inversa entre o tamanho de um construto e o tamanho de seus constituintes.

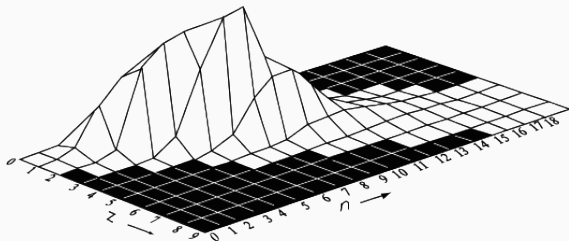
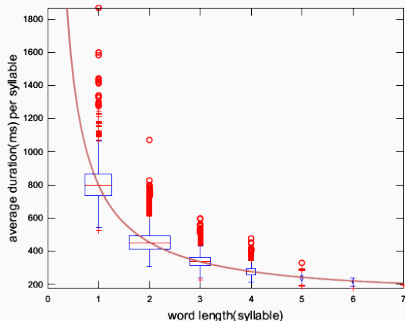


Abb. 47.1: Typenhäufigkeitsgebirge des deutschen Wortschatzes mit z = Silbenzahl und n = Lautzahl (entnommen aus Menzerath 1954, 100)

Figura 33: Frequência de tipo das palavras alemãs em relação ao número de sílabas (z) e o número de sons (n) (Menzerath 1954).

(a) Average duration of syllables as the number of syllables in a word increases. The parameters found were: $a = 743$, $b = -0.916$ and $c = 0.072$. The correlation between the model and data was 0.93.



(b) Average duration of phones as the number of phones in a word increases. The parameters found were: $a = 619$, $b = -0.901$ and $c = 0.039$. The correlation between the model and data was 0.91

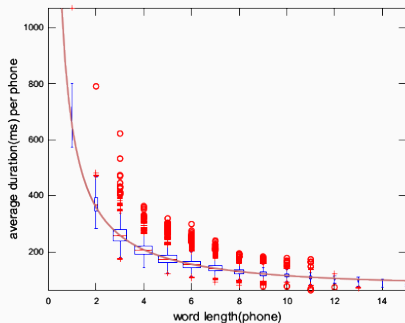


Figura 34: Relação entre o comprimento das palavras e a duração de seus constituintes. Foram analisadas 10.086 palavras do inglês, com dados obtidos de dicionários online (L. Araujo, Cristófar-Silva, e Yehia 2014).

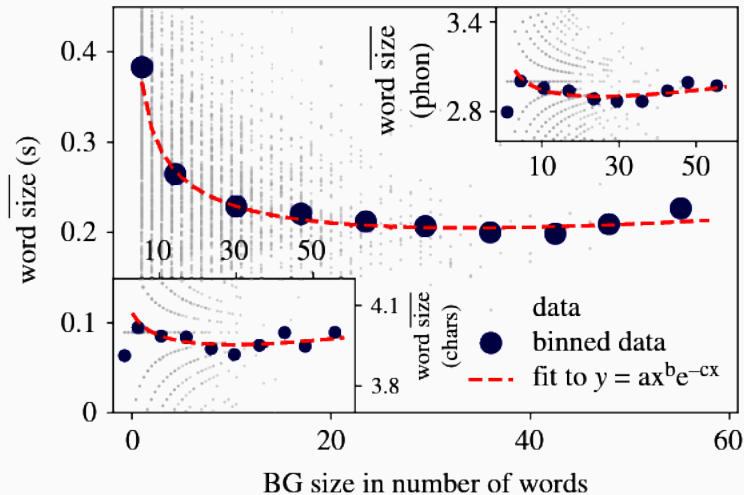


Figura 35: Lei de Menzerath-Altmann. Relação entre o tamanho de grupo respiratório (número de palavras) versus tamanho médio das palavras (duração, número de fonemas, número de caracteres) (Torre et al. 2019).

Altmann (1980) observou que o conceito de *tamanho* poderia referir-se também à complexidade e número de elementos utilizados na composição. Altmann (1980) propôs o modelo matemático

$$\frac{y'}{y} = \frac{b}{x} + c$$

para o qual a solução é

$$y = ax^b e^{cx}$$

Para Köhler (1989), a lei de Menzerath-Altmann é uma manifestação característica de sistemas complexos.

Outros estudos analisam a lei de Menzerath-Altmann em textos (Hřebíček 1995; Andres 2010; L. C. Araujo, Benevides, e Pereira 2020; Torre, Dębowski, e Hernández-Fernández 2021), fala (Hernández-Fernández, Torre, et al. 2019; Torre et al. 2019), genoma (Ramon Ferrer-i-Cancho et al. 2014), música (Boroda e Altmann 1991), comunicação gestual de chimpanzés (Heesen et al. 2019), etc.

Leis de desenvolvimento

Leis de desenvolvimento

Lei de Piotrowski

As línguas mudam pela interação entre formas antigas e novas.

- Mudanças qualitativas: mudanças em entidades individuais, mudanças sonoras.
- Mudanças em volume: crescimento/decaimento lexical.

O influxo de novos elementos em uma língua, ao longo do tempo, é descrito por

$$p(t) = \frac{c}{1 + ae^{-bt}}$$

(Altmann 1983b, 1983a; Karl-Heinz Best 2016a; Stachowski 2020)

Karl-Heinz Best (2016b) reúne uma extensa lista de publicações sobre a lei de Piotrowski.

Table 1

Fitting formula (2) to the replacement of $-(t)$ by $-(st)$ in German (Best 2003a)

Time interval	t	Frequency of $-(t)$	Frequency of $-(st)$	f_t	$p_t(2)$
1450-1479	1	853	3	0.0035	0.0016
1480-1509	2	683	1	0.0015	0.0040
1510-1539	3	950	4	0.0042	0.0099
1540-1569	4	792	27	0.0330	0.0244
1570-1599	5	426	45	0.0955	0.0589
1600-1629	6	341	56	0.1411	0.1357
1630-1659	7	421	247	0.3698	0.2824
1660-1689	8	266	332	0.5552	0.4966
1690-1719	9	178	362	0.6741	0.7121
1720-1749	10	69	437	0.8636	0.8611
1750-1779	11	59	558	0.9044	0.9395
1780-1809	12	10	645	0.9847	0.9750
1810-1839	13	7	494	0.9860	0.9899

a = 1581.7962; b = -0.9191, D = 0.99

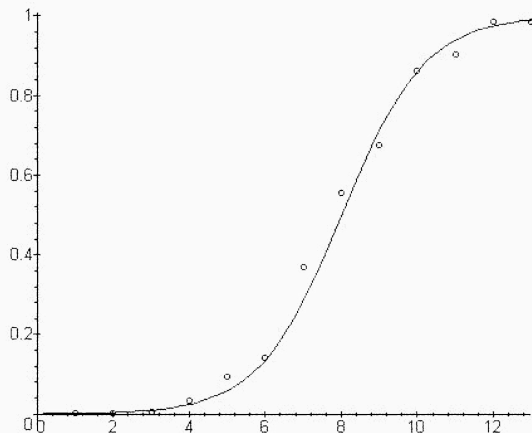


Figura 36: Substituição de $-(t)$ por $-(st)$ na 2a pessoa do singular do presente do indicativo no verbo alemão 'wollen' (*wilt* → *willst*) (K. H. Best 2006).






Variant	Formula	Shape
Complete change (Beöthy and Altmann 1982)	$p(t) = \frac{1}{1+e^{\frac{t-\mu}{s}}}$	
Partial change (Beöthy and Altmann 1982)	$p(t) = \frac{c}{1+e^{\frac{t-\mu}{s}}}$	
Reversible change (Altmann 1983)	$p(t) = \frac{1}{1+ae^{-bt+ct^2}}$	
Punctuated change (Stachowski 2013)	$p(t) = \sum \frac{c_i}{1+e^{\frac{t-\mu_i}{s_i}}}$	
Multivariate change (Yokoyama and Wada 2006; Vulcanović 2013)	$p(t) = \frac{1}{1+e^{-(a_1 t_1 + a_2 t_2 + b)}}$	

Figura 37: Variantes da lei Piotrowski-Altmann (Stachowski 2020).

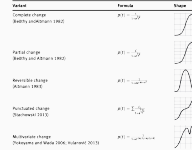


Figura 22: Variantes da lei de Piotrowski-Altmann (Stachowicz 2000).

A variante da **mudança completa** é a mais básica, descrevendo a substituição completa de uma forma por outra. Exemplo: desaparecimento do sufixo -ov do plural genitivo dos nomes de unidades nos textos técnicos russos entre 1880 e 1920. A **mudança parcial** descreve mudanças que devem ser descritas pelo volume ao invés de percentual. Por exemplo: empréstimos de outras línguas no alemão. A variante de **mudança reversível** descreve uma mudança que se inicia, tem um pico e reverte-se. Por exemplo: epêntese do -e em verbos fortes no alemão.

Starke Verben haben in der 1. Person Singular Präteritum Indikativ keine Endung -e: ich ging (gehen), sah (sehen), kam (kommen), nahm (nehmen), fand (finden), half (helfen), blieb (bleiben), ... Die Ausnahme ist werden: ich wurde (veraltet: ward)

Die e-Epithese bei den starken Verben ist ein Sprachwandelprozeß, der sich nicht durchgesetzt hat. Im Verlauf des 17. Jahrhunderts erlangte sie ihre größte Beliebtheit, war jedoch nie obligatorisch. Man findet also auch in ihrer Hochzeit bei ein und demselben Autor Formen ohne -e neben solchen mit -e. Die Formen mit -e sind immer seltener als die ohne. (<https://german.stackexchange.com/questions/48263/die-form-fande-als-1-person-singular-pr%C3%A4teritum-indikativ-e-epithese>)

Softwares

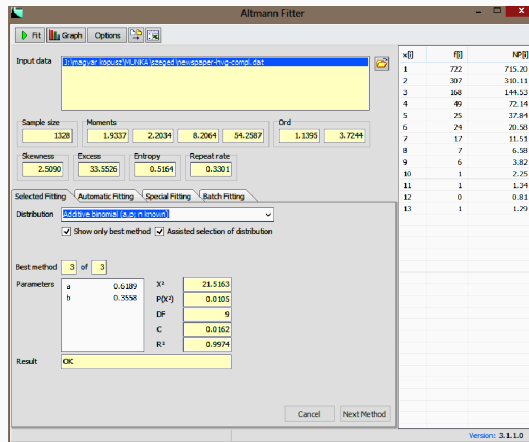


Figura 38: Tela do software Altmann-Fitter.

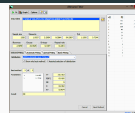


Figura 38: Tela do software Altmann-Fitter.

O Altmann Fitter é um software interativo para ajuste de distribuições de probabilidade discretas a dados de frequência. O algoritmo utilizado é baseado no Nelder-Mead Simplex. Mais de 200 distribuições estão definidas e implementadas no software. Os procedimentos matemáticos são automatizados. O software iterativamente busca o melhor ajuste, buscando a distribuição que melhor explique os dados observados.

Selected Fitting	Automatic Fitting	Special Fitting	Batch Fitting			
Distribution	χ^2	$P(\chi^2)$	C	DF	$R^2 \nabla$	▲
Mixed Poisson (a,b,a)	2.65	0.4479	0.0035	3	0.9991	
Thomas (a,b)	2.24	0.8144	0.0030	5	0.9991	
Cernuschi-Castagnetto-Poisson (a,b)	1.81	0.9366	0.0024	6	0.9989	
Gegenbauer (a,b,k)	5.95	0.3115	0.0079	5	0.9988	
Polya-Aeppli (a,p)	3.33	0.7666	0.0044	6	0.9988	
Positive Singh-Poisson (a,a)	4.99	0.2886	0.0067	4	0.9987	
Negative hypergeometric (K,M,n)	3.92	0.4163	0.0052	4	0.9986	
Polya (s,p,n)	4.10	0.3922	0.0055	4	0.9985	
Non-central negative binomial (a,k,p)	4.47	0.4834	0.0060	5	0.9981	
Gross-Harris-geometric I (q,a)	7.65	0.2653	0.0102	6	0.9980	
Darwin (B; M = x-max)	5.27	0.3833	0.0070	5	0.9976	
Consul-Mittal-binomial with 3 parameters (n,p,θ)	6.42	0.1702	0.0086	4	0.9974	
Right truncated modified Zipf-Alekseev (a,b; n = x...	10.75	0.0295	0.0144	4	0.9970	
Doubly truncated logarithmic (q; L = x-min,R = x-...	13.67	0.0178	0.0183	5	0.9966	
Right truncated logarithmic (q; R = x-max)	13.67	0.0335	0.0183	6	0.9966	
Bissinger-geometric (p)	12.79	0.0773	0.0171	7	0.9966	▼

Figura 39: Tela do software Altmann-Fitter.

Selected Fitting
Automatic Fitting
Special Fitting
Batch Fitting

Distribution
Altmann-Wimmer (a0,a1,a2,b1,b2)

Preflight

a0

☐

0.00000000

a1

☐

0.00000000

a2

☐

0.00000000

b1

-8.00000000

8.00000000

0.16000000

Default

b2

-8.00000000

8.00000000

0.16000000

Default

Parameters

X²

0.0000

P(X²)

0.0000

DF

0

C

0.0000

R²

0.0000

Result

Cancel

Figura 40: Tela do software Altmann-Fitter.

Pacotes

languageR Analyzing Linguistic Data: A Practical Introduction to Statistics

Autores: R. H. Baayen, Elnaz Shafaei-Bajestan

<https://cran.r-project.org/web/packages/languageR/>

zipfR Statistical Models for Word Frequency Distributions

Autores: Stefan Evert, Marco Baroni

<https://cran.r-project.org/web/packages/zipfR/>

fitdistrplus Help to Fit of a Parametric Distribution to Non-Censored or Censored Data

Autores: Marie-Laure Delignette-Muller, Christophe Dutang, Regis Pouillot,
Jean-Baptiste Denis, Aurelie Siberchicot

<https://cran.r-project.org/web/packages/fitdistrplus/>

```
> ItaRi.spc
```

	m	V _m
1	1	346
2	2	105
3	3	74
4	4	43
5	5	39
6	6	25
7	7	27
8	8	15
9	9	17
10	10	9

...

N	V
1399898	1098



```
> summary(ItaRi.spc)
```

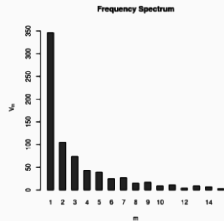
zipfR object for frequency spectrum

Sample size: N = 1399898

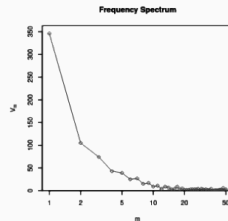
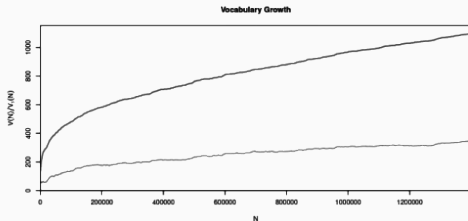
Vocabulary size: V = 1098

Class sizes: V_m = 346 105 74 43 39 25 27 15 ...

```
> plot(ItaRi.spc)
> plot(ItaRi.spc, log="x")
```



```
> plot(ItaRi.emp.vgc, add.m=1)
```



Teoria

Abordagens:

1. linguística sinérgica (Köhler 1986, 1987, 2005)
2. teoria unificada (Wimmer e Altmann 2005)

Abordagens:

1. linguística sinérgica (Köhler 1986, 1987, 2005)
2. teoria unificada (Wimmer e Altmann 2005)

Segundo a filosofia da ciência, a pesquisa científica ocorre em três níveis: observação, descrição e explicação. O nível mais alto, explicação, não é possível sem o estabelecimento de leis. Vimos alguns exemplos que ilustram o que é e o que faz a linguística quantitativa, uma área de estudo ainda incipiente, almejando a construção de uma teoria que seja capaz de descrever e explicar a diversidade da comunicação humana. As leis da linguística quantitativa são observadas na comunicação escrita, na comunicação oral, em certos aspectos da comunicação animal e, algumas vezes, também em outros fenômenos da natureza. Algumas delas podem ser ligadas à emergência em sistemas complexos, aspectos associados à sistemas de comunicação, ou aspectos cognitivos e psicológicos que devem ser explorados.

1. A proposta sinérgica busca integrar leis e hipóteses em um modelo complexo para descrever o fenômeno linguístico. Utilizam para tanto um axioma central: as línguas são sistemas auto-organizativos. Estabelece também alguns requisitos que devem ser seguidos por um sistema semiótico: ser possível realizar codificação e decodificação com eficiência, economia de memória, minimização de esforço, dentre outros.
2. A teoria unificada busca integrar as leis e hipóteses a partir de equações diferenciais gerais (ou equações de diferenças, no caso discreto), assim como dois pressupostos gerais: 1) a dinâmica de uma variável linguística y será modelada matematicamente em termos de sua mudança relativa (dy/y); 2) uma outra variável x que tenha efeito sobre y deverá ser considerada da mesma forma em termos de sua mudança relativa (dx/x).

Referências

- Abe, Sumiyoshi, e Norikazu Suzuki. 2005. "Scale-free statistics of time interval between successive earthquakes". *Physica A: Statistical Mechanics and its Applications* 350 (2): 588–96.
- Adamic, Lada A., e Bernardo A. Huberman. 2002. "Zipf's law and the Internet". *Glottometrics* 3: 143–50.
- Altmann, Gabriel. 1980. "Prolegomena to Menzerath's law". *Glottometrika* 2: 1–10.
- . 1983a. "A Law of Change in Language, Historical Linguistics". *Quantitative Linguistics* 18: 104–15.
- . 1983b. "Das Piotrowski-Gesetz und seine Verallgemeinerungen". *Exakte Sprachwandelforschung*, 54–90.
- Altmann, Gabriel, e Violetta Burdinski. 1982. "Towards a law of word repetitions in text-blocks". *Glottometrika* 4: 147–67.

- Andres, Jan. 2010. "On a Conjecture about the Fractal Structure of Language". *Journal of Quantitative Linguistics* 17 (2): 101–22. <https://doi.org/10.1080/09296171003643189>.
- Araujo, Leonardo Carneiro, Aline Benevides, e Marcos Pereira. 2020. "Análise da Lei de Menzerath no Português Brasileiro". *Linguamática* 12 (1): 31–48. <https://doi.org/10.21814/lm.12.1.300>.
- Araujo, Leonardo, Thaïs Cristófar-Silva, e Hani Yehia. 2014. "Menzerath's law on word duration". Em *1st International DINAFON Meeting*.
- Araújo, Leonardo Carneiro de. 2013. "Statistical analyses in language usage". Tese de doutorado, Universidade Federal de Minas Gerais.
- Araújo, Leonardo Carneiro de, Thaïs Cristófar-Silva, e Hani C. Yehia. 2013. "Entropy of a Zipfian Distributed Lexicon". *Glottometrics* 26: 38–49.
- Baayen, Harald. 1996. "The effects of lexical specialization on the growth curve of the vocabulary". *Computational Linguistics* 22 (4): 455–80.

- Baayen, R Harald. 2001. *Word frequency distributions*. Vol. 18. Springer Science & Business Media.
- Baek, Seung Ki, Sebastian Bernhardsson, e Petter Minnhagen. 2011. "Zipf's law unzipped". *New Journal of Physics* 13 (4): 043004.
- Bagheri, Dariusch. 2002. "Definitionsfolgen und Lexemnetze". Em *Einführung in die quantitative Lexikologie*, editado por Gabriel Altmann, Dariusch Bagheri, Hans Goebel, Reinhard Köhler, e Claudia Prün, 124. Peust & Gutschmidt.
- Baixeries, Jaume, Brita Elvevåg, e Ramon Ferrer-i-Cancho. 2013. "The Evolution of the Exponent of Zipf's Law in Language Ontogeny". Editado por Satoru Hayasaka. *PLoS ONE* 8 (3): e53227. <https://doi.org/10.1371/journal.pone.0053227>.
- Bak, Per. 1996. *How Nature Works: the science of self-organized criticality*. Copernicus.

- Bentz, Chris, e Ramon Ferrer Cancho. 2016. "Zipf's law of abbreviation as a language universal". Em *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, 1–4. University of Tübingen.
- Best, K. H. 2006. *Quantitative Linguistik: eine Annäherung*. Göttinger linguistische Abhandlungen. Peust & Gutschmidt. https://books.google.com.br/books?id=_iCwMQAACAAJ.
- Best, Karl-Heinz. 2005. "Sprachliche Einheiten in Textblöcken." *Glottometrics* 9: 1–12.
- . 2016a. "Bibliography–Piotrowski's law". *Glottology* 7 (1): 89–93.
- . 2016b. "Bibliography – Piotrowski's law". *Glottology* 7 (1).
<https://doi.org/10.1515/glot-2016-0006>.
- Bian, Chunhua, Ruokuang Lin, Xiaoyu Zhang, Qianli D. Y. Ma, e Plamen Ch. Ivanov. 2016. "Scaling laws and model of words organization in spoken and written language". *EPL (Europhysics Letters)* 113 (1): 18002. <https://doi.org/10.1209/0295-5075/113/18002>.

- Boroda, M., e Gabriel Altmann. 1991. "Menzerath's law in musical texts". *Musikometrika* 3: 1–13.
- Bunge, Mario. 2012. *Scientific research II: The search for truth*. Springer Science & Business Media.
- Cancho, Ramon Ferrer i. 2005. "Decoding least effort and scaling in signal frequency distributions". *Physica A: Statistical Mechanics and its Applications* 345 (1-2): 275–84.
- . 2007. "On the universality of Zipf's law for word frequencies". Em *Exact Methods in the Study of Language and Text*, editado por Peter Grzybek e Reinhard Köhler, 131–40. De Gruyter Mouton. <https://doi.org/10.1515/9783110894219.131>.
- Cancho, Ramon Ferrer i, e Ricard V. Solé. 2003. "Least effort and the origins of scaling in human language". *Proceedings of the National Academy of Sciences* 100 (3): 788–91. <https://doi.org/10.1073/pnas.0335980100>.
- Chater, Nick, e Gordon DA Brown. 1999. "Scale-invariance as a unifying psychological principle". *Cognition* 69 (3): B17–24.

- Choi, JS, Kyungsik Kim, SM Yoon, KH Chang, e C Christopher Lee. 2005. "Zipf's law distributions in Korean financial markets". *Journal of the Korean Physical Society* 47 (1): 171.
- Conrad, Brian, e Michael Mitzenmacher. 2004. "Power laws for monkeys typing randomly: the case of unequal probabilities". *IEEE Transactions on information theory* 50 (7): 1403–14.
- Corominas-Murtra, Bernat, e Ricard V Solé. 2010. "Universality of Zipf's law". *Physical Review E* 82 (1): 011102.
- Cox, Raymond A K, James M. Felton, e Kee C. Chung. 1995. "The concentration of commercial success in popular music: an analysis of the distribution of gold records". *Journal of Cultural Economics* 19: 333–40.

- Crovella, Marl E., e Azer Bestavros. 1996. "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes". Em *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 160–69.
<http://www.cs.bu.edu/faculty/crovella/paper-archive/self-sim/sigmetrics-version.ps>.
- Fan, Fengxiang, Yaqin Wang, e Zhao Gao. 2014. "Some macro quantitative features of low-frequency word classes." *Glottometrics* 28: 1–12.
- Farmer, J Doyne, e John Geanakoplos. 2008. "Power laws in economics and elsewhere". Em *Santa Fe Institute*.
- Fengxiang, Fan. 2010. "An asymptotic model for the English hapax/vocabulary ratio". *Computational Linguistics* 36 (4): 631–37.
- Ferrer-i-Cancho, R. 2005. "Hidden communication aspects inside the exponent of Zipf's law". *Glottometrics* 11 (janeiro): 96–117.

- Ferrer-i-Cancho, Ramon, Antoni Hernández-Fernández, Jaume Baixeries, Łukasz Dębowski, e Ján Mačutek. 2014. "When is Menzerath-Altmann law mathematically trivial? A new approach". *Statistical applications in genetics and molecular biology* 13 (6): 633–44.
- Ferrer-i-Cancho, Ramon, Antoni Hernández-Fernández, David Lusseau, Govindasamy Agoramoorthy, Minna J Hsu, e Stuart Semple. 2013. "Compression as a universal principle of animal behavior". *Cognitive Science* 37 (8): 1565–78.
- Ferrer-i-Cancho, Ramon, e Michael S Vitevitch. 2018. "The origins of Zipf's meaning-frequency law". *Journal of the Association for Information Science and Technology* 69 (11): 1369–79.
- Fujiwara, Yoshi. 2004. "Zipf law in firms bankruptcy". *Physica A: Statistical and Theoretical Physics* 337 (1-2): 219–30.
- Furusawa, Chikara, e Kunihiro Kaneko. 2003. "Zipf's law in gene expression". *Physical review letters* 90 (8): 088102.

- Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation". *Quarterly Journal of Economics* 114 (3): 739–67.
- Gopinath, Deepa P, S Veena, e Achuthsankar S Nair. 2008. "Modeling of vowel duration in Malayalam speech using probability distribution". *Proceedings of the speech prosody, Campinas, Brazil*, 6–9.
- Grimes, Barbara F. 2000. "Ethnologue: Languages of the World". Texas: <https://www.ethnologue.com/14/>; SIL International.
- Guiraud, Pierre. 1968. "The semic matrices of meaning". *Social science information* 7 (2): 131–39.
- Heesen, Raphaela, Catherine Hobaiter, Ramon Ferrer-i-Cancho, e Stuart Semple. 2019. "Linguistic laws in chimpanzee gestural communication". *Proceedings of the Royal Society B: Biological Sciences* 286 (1896): 20182900. <https://doi.org/10.1098/rspb.2018.2900>.

- Herdan, Gustav. 1958. "The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics". *Biometrika* 45 (1-2): 222–28.
- Hernández-Fernández, Antoni, Bernardino Casas, Ramon Ferrer-i-Cancho, e Jaume Baixeries. 2016. "Testing the robustness of laws of polysemy and brevity versus frequency". Em *International Conference on Statistical Language and Speech Processing*, 19–29. Springer.
- Hernández-Fernández, Antoni, Iván G. Torre, Juan-María Garrido, e Lucas Lacasa. 2019. "Linguistic laws in speech: the case of Catalan and Spanish". *Entropy* 21 (12): 1153.
- Hernández-Fernández, Antoni, Iván G. Torre, Juan-María Garrido, e Lucas Lacasa. 2019. "Linguistic Laws in Speech: The Case of Catalan and Spanish". *Entropy* 21 (12): 1153.
<https://doi.org/10.3390/e21121153>.

- Hernando, A, D Puigdomènech, D Villuendas, C Vesperinas, e A Plastino. 2009. "Zipf's law from a Fisher variational-principle". *Physics Letters A* 374 (1): 18–21.
- Hřebíček, Ludek. 1995. *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*. Quantitative linguistics. Wissenschaftlicher Verlag Trier.
- Huffman, David A. 1952. "A method for the construction of minimum-redundancy codes". *Proceedings of the IRE* 40 (9): 1098–1101.
- Ilgen, Bahar, e Bahar Karaoglan. 2007. "Investigation of Zipf's 'law-of-meaning' on Turkish corpora". Em *2007 22nd international symposium on computer and information sciences*, 1–6. IEEE.
- Ito, Keisuke, e Mitsuhiro Matsuzaki. 1990. "Earthquakes as self-organized critical phenomena". *Journal of Geophysical Research: Solid Earth* 95 (B5): 6853–60.

- Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson, e Simon Kirby. 2017. "Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication". *Cognition* 165 (agosto): 45–52. <https://doi.org/10.1016/j.cognition.2017.05.001>.
- Kawamura, Kenji, e Naomichi Hatano. 2002. "Universality of Zipf's law". *Journal of the Physical Society of Japan* 71 (5): 1211–13.
- Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Vol. 31. N. Brockmeyer.
- . 1987. "System theoretical linguistics". *Theoretical Linguistics* 14 (2/3): 241–57.
- . 1989. "Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus". *Das Menzerathsche Gesetz in informations-verarbeitenden Systemen*. Olms, Hildesheim, 108–16.

- . 2005. "Synergetic linguistic". Em *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, editado por Reinhard Köhle, Gabriel Altmann, e Rajmond G. Piotrowski. Berlin, New York: de Gruyter.
- . 2014. "Laws of language and text in quantitative and synergetic linguistics". Em *Aggregating Dialectology, Typology, and Register Analysis*, 426–50. Berlin, Boston: Gruyter.
- Kohli, Rajeev, e Raaj Kumar Sah. 2003. "Market Shares: Some Power Law Results and Observations". *Management Science* 52 (11): 1792–98.
- Kornai, András. 2002. "How many words are there?" *Glottometrics* 4: 61–86.
- Leijenhorst, D. C. van, Th. P. van der Weide, e F. A. Grootjen. 2005. "A formal derivation of Heaps' Law". *Information Sciences* 170 (2-4): 263–72.
- Li, Wentian. 1992. "Random texts exhibit Zipf's-law-like word frequency distribution". *IEEE Transactions on information theory* 38 (6): 1842–45.

- Mandelbrot, Benoit. 1953. *Contribution à la théorie mathématique des jeux de communication*. Publications de l'Institut de Statistique de l'Université de Paris. Institut Henri Poincaré.
- . 1966. "Information Theory and Psycholinguistics: A Theory of Word Frequencies, Readings in Mathematical Social Sciences". MIT Press, MA, USA.
- Mandelbrot, Benoît. 1963. "The Variation of Certain Speculative Prices". *The Journal of Business* 36 (4): 394–419.
- Mandelbrot, Benoît B. 1953. "An informational theory of the statistical structure of languages". Em *Communication Theory*, editado por Betterworth W. Jackson, 486–502.
- Manin, Dmitrii Y. 2008. "Zipf's law and avoidance of excessive synonymy". *Cognitive Science* 32 (7): 1075–98.

- McComas, William F. 2014. "Law (Scientific Law or Principle)". Em *The Language of Science Education: An Expanded Glossary of Key Terms and Concepts in Science Teaching and Learning*, editado por William F. McComas, 58–58. Rotterdam: SensePublishers.
https://doi.org/10.1007/978-94-6209-497-0_51.
- Mehri, Ali, e Maryam Jamaati. 2017. "Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations". *Physics Letters A* 381 (31): 2470–77.
<https://doi.org/10.1016/j.physleta.2017.05.061>.
- Menzerath, Paul. 1954. *Die Architektonik des deutschen Wortschatzes*. Phonetische Studien. F. Dümmler.
- Miller, George A. 1957. "Some effects of intermittent silence". *The American journal of psychology* 70 (2): 311–14.

- Mitzenmacher, Michael. 2004. "A brief history of generative models for power law and lognormal distributions". *Internet mathematics* 1 (2): 226–51.
- Moura Jr, Newton J, e Marcelo B Ribeiro. 2006. "Zipf law for Brazilian cities". *Physica A: Statistical Mechanics and its Applications* 367: 441–48.
- Nicolis, Grégoire, Cathy Nicolis, e John S. Nicolis. 1989. "Chaotic dynamics, Markov partitions, and Zipf's law". *Journal of Statistical Physics* 54 (3–4): 915–24.
- Parker-Rhodes, AF, e T Joyce. 1956. "A theory of word-frequency distribution". *Nature* 178 (4545): 1308–8.
- Piantadosi, Steven T, Harry Tily, e Edward Gibson. 2011. "Word lengths are optimized for efficient communication". *Proceedings of the National Academy of Sciences* 108 (9): 3526–29.
- Popescu, Ioan-Iovitz, e Gabriel Altmann. 2008. "Hapax Legomena and Language Typology". *Journal of Quantitative Linguistics* 15 (4): 370–78. <https://doi.org/10.1080/09296170802326699>.

- Rosen, Kristin M. 2005. "Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison". *Journal of Phonetics* 33 (4): 411–26.
- Salge, Christoph, Nihat Ay, Daniel Polani, e Mikhail Prokopenko. 2015. "Zipf's law: balancing signal usage cost and communication efficiency". *PLoS one* 10 (10): e0139475.
- Sayli, Omer. 2002. "Duration analysis and modeling for Turkish text-to-speech synthesis". Mathesis, Bogazici Universitesi.
- Shaw, Jason A, e Shigeto Kawahara. 2019. "Effects of surprisal and entropy on vowel duration in Japanese". *Language and speech* 62 (1): 80–114.
- Sigurd, Bengt, Mats Eeg-Olofsson, e Joost Van Weijer. 2004. "Word length, sentence length and frequency–Zipf revisited". *Studia linguistica* 58 (1): 37–52.
- Simon, Herbert A. 1955. "On a class of skew distribution functions". *Biometrika* 42 (3/4): 425–40.
- Solla Price, Derek John de. 1965. "Networks of Scientific Papers". *Science* 149 (3683): 510–15.

- Stachowski, Kamil. 2020. "Piotrowski-Altmann law: State of the art". *Glottology* 11 (1).
<https://doi.org/10.1515/glot-2020-2002>.
- Strauss, Udo, Peter Grzybek, e Gabriel Altmann. 2007. "Word length and word frequency". Em *Contributions to the science of text and language*, 277–94. Springer.
- Teahan, William J, Yingying Wen, Rodger McNab, e Ian H Witten. 2000. "A compression-based algorithm for Chinese word segmentation". *Computational Linguistics* 26 (3): 375–93.
- Tomaschek, Fabian, Martijn Wieling, Denis Arnold, e R Harald Baayen. 2017. "Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience". Em *14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013), Lyon, France, 25-29 August 2013*, 1302–6. International Speech Communications Association.

- Torre, Iván G., Łukasz Dębowski, e Antoni Hernández-Fernández. 2021. "Can Menzerath's law be a criterion of complexity in communication?" Editado por Diego Raphael Amancio. *PLOS ONE* 16 (8): e0256133. <https://doi.org/10.1371/journal.pone.0256133>.
- Torre, Iván G., Bartolo Luque, Lucas Lacasa, Christopher T. Kello, e Antoni Hernández-Fernández. 2019. "On the physical origin of linguistic laws and lognormality in speech". *Royal Society Open Science* 6 (8): 191023. <https://doi.org/10.1098/rsos.191023>.
- Tweedie, Fiona J, e R Harald Baayen. 1998. "How variable may a constant be? Measures of lexical richness in perspective". *Computers and the Humanities* 32 (5): 323–52.
- Wichmann, Søren. 2005. "On the power-law distribution of language family sizes". *Journal of Linguistics* 41 (1): 117–31. <https://doi.org/10.1017/s0022222670400307x>.
- Wikipedia. 2021a. "Martinsches Gesetz". *Wikipedia*.
https://de.wikipedia.org/wiki/Martinsches_Gesetz.

———. 2021b. "Textblockgesetz". *Wikipedia*.

<https://de.wikipedia.org/w/index.php?title=Textblockgesetz&oldid=218036550>.

Wimmer, Gejza, e Gabriel Altmann. 2005. "Unified derivation of some linguistic laws (Die vereinheitlichte Ableitung linguistischer Gesetze)." Em *Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An International Handbook*, editado por Reinhard Köhler, Gabriel Altmann, e Rajmund G Piotrowski. Berlin, New York: Walter de Gruyter.

Yule, G. Udny. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.

Zipf, George Kingsley. 1935. *The Psycho-biology of language: an introduction to dynamic philology*. The MIT Press.

———. 1945. "The Meaning-Frequency Relationship of Words". *The Journal of General Psychology* 33 (2): 251–56. <https://doi.org/10.1080/00221309.1945.10544509>.

———. 1949. *Human behavior and the principle of least effort*. Addison-Wesley.