Proposta recebida em September 2019 e aceite para publicação em April 2020.

Análise da Lei de Menzerath no Português Brasileiro

Menzerath's law analysis in Brazilian Portuguese

Leonardo Araujo
Universidade Federal de São João del Rei
leolca@ufsj.edu.br

Aline Benevides
Universidade de São Paulo
benevides.aline12@gmail.com

Marcos Pereira
Universidade Federal de São João del Rei
marcos.vinicius@ufsj.edu.br

Resumo

Sob a ótica da Linguística Quantitativa, este trabalho revisita a Lei de Menzerath, aplicando-a aos dados do português brasileiro, a partir das seguintes unidades de análise: palavras, sílabas e fonemas. Os dados foram extraídos do Corpus ABG. Análises estatísticas foram realizadas nos modelos propostos, as quais demonstraram uma relação de decrescimento entre o comprimento médio das palavras (em sílabas) e o comprimento médio das sílabas (em fonemas); resultados esses que corroboram a Lei de Menzerath. Além disso, constatou-se, de maneira geral, que melhores medições ou a existência de variáveis não consideradas no modelo poderão ser utilizadas para melhorá-lo.

Palavras chave

lei de Menzerath, linguística quantitativa, análise estatística

Abstract

Under the perspective of Quantitative Linguistics, this paper revisits the Menzerath's Law, applying it to data from Brazilian Portuguese, using the following unities of analysis: words, syllables and phonemes. The data was extracted from the ABG Corpus. Statistical analyses are performed on the proposed models, corroborating the existence of a decay relationship between the mean length of words (in syllables) and the average length of syllables (in phonemes); what corroborates the Menzerath Law. It is noticed that better measures or variables not considered in the model might be used to improve it.

Keywords

Menzerath's law, quantitative linguistics, statistical analysis

1. Introdução

A linguística é o campo das ciências que estuda a linguagem, analisando a sua forma, utilização, significado e contexto. Ela se estabeleceu, de forma consistente, a partir do século XX, através da publicação do Curso de Linguística Geral, de Ferdinand de Saussure, em 1916, dando início a Linguística Moderna. Neste período, a primeira orientação teórica, denominada de Linguística Estruturalista, emerge com o objetivo de estabelecer e de descrever os sistemas linguísticos, valendo-se, para tanto, da noção de valor, a partir de distinções teóricas como *língua* vs. *fala*, *forma* vs. *substância*. Os estruturalistas, por serem avessos ao estudo do sentido, de caráter mental e, portanto, pertencente à psicologia individual, dedicam-se ao estudo da forma. Nesse sentido, a língua passa a ser analisada do ponto de vista sistêmico por meio de relações de oposição (noção de valor). Sob orientação de Leonard Bloomfield, o estruturalismo americano estabelece que a análise das estruturas e das categorias gramaticais devem ser realizadas a partir de dados de sentenças ou de textos, e não mais extraídos de experiências prévias (Ilari, 2003).

A Linguística Quantitativa ganha abrangência, nesse sentido, ao conjugar métodos estatísticos e computacionais, a partir da análise de corpora linguísticos, a fim de caracterizar a linguagem, sua evolução e estrutura. O seu principal propósito é estabelecer leis que modelem a linguagem ou a comunicação e produzir formulações para uma teoria geral da linguagem (Altmann & Schwibbe, 1989; Köhler, 2005). Atribui-se o status de leis científicas àquelas formulações que podem ser derivadas a partir de axiomas, criando uma estrutura firme de rede nomológica.

Sob o auspício da tradição das ciências teoréticas, a Linguística Quantitativa busca estabelecer leis e formulações capazes de inter-

DOI: 10.21814/lm.12.1.300

relacioná-las, através de proposições e derivações lógicas. É sob essa perspectiva que o presente trabalho se baseia. Propomo-nos, neste artigo, a testar a predição da Lei de Menzerath a partir da análise de um corpus linguístico do português brasileiro, o Corpus ABG (Benevides & Guide, 2017). A Lei de Menzerath é conhecida por predizer, em termos gerais, que "um som é mais curto quão maior o todo em que ele ocorre (lei da quantidade)" e que "quantos mais sons possuir uma sílaba, menor serão seus comprimentos relativos"(Menzerath, 1954, p. 100).

Este artigo estrutura-se da seguinte maneira: na Seção 1, apresenta-se uma contextualização da área de Linguística Quantitativa (Seção 1.1), da Lei de Menzerath-Altmann (Seção 1.2) e de sua formulação matemática (Seção 1.3); a Seção 2 apresenta a abordagem deste trabalho, destacando as unidades linguísticas e o corpus com os dados do português brasileiro que foram utilizados para realizar as análises desta pesquisa; a Seção 2.4 apresenta os resultados gerados a partir da análise dos dados do corpus e os resultados de ajustes dos modelos matemáticos; na Seção 2.5, busca-se explorar os resultados e contrastá-los com as teorias linguísticas; e, por fim, conclui-se o trabalho na Seção 3.

1.1. Linguística Quantitativa

Os trabalhos em Linguística Quantitativa buscam explicações provenientes de leis estocástico-linguísticas que venham a estabelecer uma teoria geral da linguagem. Uma rápida revisão a respeito das principais leis estatísticas na linguística pode ser encontrada em Altmann & Gerlach (2016) ou também na enciclopédia *on-line* destinada à Linguística Quantitativa, Glottopedia (2019). A lei mais conhecida é a Lei de Zipf (Zipf, 1935, 1949; Ferrer-i-Cancho & Solé, 2002; Mitzenmacher, 2004; Ferrer-i-Cancho, 2006). Zipf observou, por meio da quantidade de ocorrências de palavras em um corpus, a existência de uma proporção inversa entre a frequência da palavra e o seu ranque, quando ordenadas por frequência. O mesmo tipo de relação também é verificada em outros fenômenos da natureza, por exemplo: na magnitude de terremotos (Abe & Suzuki, 2005), na população de cidades (Gabaix, 1999) e no número de requisições de páginas na internet (Adamic & Huberman, 2002). Há, ainda, trabalhos que buscam relacionar diferentes leis da Linguística Quantitativa, como o trabalho de Lü et al. (2010) que buscou relacionar a Lei de Zipf com a Lei de Heaps (Grzybek, 2007; Lü et al., 2010; Heaps, 1978; Herdan, 1960), a qual descreve o crescimento sublinear do vocabulário com o comprimento do corpus.

Tendo em vista que as bases ontológicas são essenciais para a construção de uma verdadeira teoria da linguagem e da comunicação sob o prisma da Linguística Quantitativa, transcreve-se abaixo um trecho do capítulo inicial de Altmann & Schwibbe (1989):

> *Leis são hipóteses bem fundamentadas e confirmadas. Uma generalização empírica nunca poderá se tornar uma lei, a menos que sejamos capazes de derivar a teoria de uma hipótese correspondente a ela. Nas ciências empíricas, esta é a forma mais comum de pesquisa: observações são feitas sob o pano de fundo de uma 'teoria' ainda embrionária, vaga e não formalizada, levando a generalizações empíricas, para a qual uma teoria correspondente é construída. Sem o estabelecimento de leis, um conjunto de afirmações dificilmente poderá ser chamado de teoria. Por esta razão, hoje não podemos falar na existência de uma teoria da linguagem, teoria gramatical, e assim por diante. A maioria dos conceitos linguísticos, embora bem complicados, consiste em uma gama de generalizações empíricas.* (Altmann & Schwibbe, 1989, p. 1)

1.2. O todo sem a parte não é todo

É notório que todas as línguas, apesar de terem uma representação fonológica das unidades linguísticas, se manifestam de forma que há grande variação em suas realizações, seja entre grupos ou indivíduos, ou mesmo quando se analisa diferentes realizações de um mesmo indivíduo. Os falantes não são meros usuários passivos, mas são parte integrante e ativa na dinâmica de uma língua. O uso acarreta mudanças, sendo que diversos fatores contribuem para tal. Fatores cognitivos, culturais, sociais e históricos, por exemplo, levam a mudanças constantes que, ao longo do tempo, podem provocar o surgimento de novos dialetos e idiomas (Bybee, 2015). Além dos fatores extralinguísticos, fatores internos à língua, como fonêmicos, morfológicos, sintáticos e semânticos, também contribuem para as constantes mudanças. Estas podem ser visualizadas de forma sincrônica ou diacrônica e sempre evidenciam a existência da ordem imanente. Quer a análise de uma língua seja feita no nível de sentenças, palavras, morfemas, sílabas ou fonemas, ordem e desordem são forças inerentes ao

processo linguístico.[1] Embora muito tenha sido investigado sobre a estrutura e a forma da língua e sua variação no tempo, ainda não existem certezas rígidas a respeito da aquisição e do processamento linguístico. Uma das hipóteses seria a existência de processos primários que guiariam a estruturação e o uso da língua, processos esses que atuariam em níveis mais altos e que poderiam ser compreendidos como generalizações (ou abstrações) de vários processos linguísticos, como leis fonéticas e gramaticais.

Em consonância com essa hipótese, vários estudos buscam encontrar e analisar quais seriam os motores atuantes em níveis hierárquicos mais altos. A Lei de Menzerath, por exemplo, é uma das leis mais conhecidas e corroboradas pela Linguística Quantitativa. Ela foi inicialmente elaborada por Menzerath (1928), sendo matematicamente formulada por Altmann (1980) e, posteriormente, confirmada pelos trabalhos de Hřebíček (1995); Andres (2010), dentre outros. Menzerath (1928, p. 104) propõe que "um som é mais curto quão maior o todo em que ele ocorre (lei da quantidade)" e "quantos mais sons possuir uma sílaba, menor serão seus comprimentos relativos". Tal postulação foi realizada a partir de uma análise do léxico da língua alemã, em que Menzerath (1954, p. 101) concluiu que "o número relativo de sons em uma sílaba decresce quando o número de sílabas em uma palavra aumenta", cunhando a frase "quão maior o todo, menores as partes!".

Antes de Menzerath, outros pesquisadores já demonstraram algumas observações que vão ao encontro da formulação de Menzerath. Jespersen (1904), por exemplo, analisou o comprimento das sílabas do francês, em especial a duração da vogal *a* em palavras como *pâtisserie*, *pâte* e *pâté*, e verificou que a vogal era sistematicamente mais curta em palavras mais longas. Outros autores também observaram semelhante tendência de redução da duração das vogais (Meyer, 1904; Roudet, 1910). Essas observações foram sistematizadas e estruturadas por Menzerath (1954), em uma análise sobre a estrutura morfológica do alemão. De forma semelhante ao *Princípio do Esforço Mínimo* formulado por Zipf (1935, 1949), Menzerath (1954) utilizou essa mesma proposição filosófica, chamando-a de *Princípio da Economia Cognitiva*, o que se manifestaria como um "fluxo constante de informação linguística" (Fenk & Fenk-Oczlon, 2013).

Conforme salientado, a proposta de Menzerath (1954) ganhou formulação matemática com o trabalho de Altmann (1980), que buscou descrever a relação entre os constituintes e os construtos. A sua validade foi observada, inicialmente, no indonésio e no inglês com Altmann (1980), seguida pelos trabalhos de Gerlach (1982), Hřebíček (1995), Polikarpov (2000a) para as línguas alemã, turca e russa, respectivamente. Além desses trabalhos, verificou-se, mais recentemente, a aplicação da Lei de Menzerath em uma análise paralela de um mesmo texto em 50 línguas diferentes (Coloma, 2015). Mais tarde, a mesma relação foi mostrada válida em música (Boroda & Altmann, 1991), cromossomos e genes (Wilde & Schwibbe, 1989; Ferrer-I-Cancho & Forns, 2009; Li, 2011; Nikolaou, 2014), proteínas (Shahzad et al., 2015) e, ainda, na compressão de dados, sob a ótica da teoria da informação (Gustison et al., 2016; Ferrer-i-Cancho, 2017).

1.3. Formulação Matemática da Lei de Menzerath

A formulação matemática, apresentada por Altmann (1980), propõe uma taxa de decrescimento constante do comprimento do componente, ou seja, $(1/y)\mathrm{d}y/\mathrm{d}x = -c$, onde y é o comprimento do constituinte e x o comprimento do construto. A formulação usual se restringe aos componentes de unidades imediatamente vizinhas, como oração e sentença ou palavras e sílabas. Entretanto, é possível estabelecermos relações compostas entre unidades que não sejam imediatamente vizinhas hierárquicas, por exemplo, palavras e sentenças, fonemas e palavras. Essa formulação pode ser necessária em línguas que apresentam palavras não silábicas,[2] como o russo (Grzybek & Altmann, 2002). Um refinamento no modelo é feito, considerando, além do decrescimento constante, um membro adicional inversamente proporcional ao comprimento do construto, como apresentado na Equação (1).

$$\frac{\mathrm{d}y/\mathrm{d}x}{y} = -c + \frac{b}{x} \tag{1}$$

[1] A linguagem pode ser vista como um sistema complexo e adaptativo, consistindo de múltiplos agentes que interagem entre si. O comportamento de cada agente depende de suas experiências passadas e também do ambiente e do contexto em que está inserido. O comportamento de um agente é fruto de diversos fatores, desde restrições perceptuais até motivações sociais. Como resultado da desordem criada nesse processo complexo, há a emergência de regularidades e de padrões (Beckner et al., 2010).

[2] Algumas línguas possuem palavras não silábicas, usualmente constituídas por uma ou duas consoantes (não silábicas) e sem a presença de vogais. O russo, por exemplo, possui as preposições *k*, *v* e *s* que funcionam como proclíticos para as palavras seguintes, contribuindo, assim, para o seu comprimento.

A taxa relativa de mudança no comprimento do constituinte ($^{(\mathrm{d}y/\mathrm{d}x)}/_{y}$) é uma soma de duas parcelas: a primeira, inversamente proporcional ao comprimento do construto ($^{b}/_{x}$), e a segunda, um fator constante ($-c$). Essa equação diferencial em (1) pode ser solucionada pela integração direta,[3] resultando em

$$y = Ax^b e^{-cx}, \tag{2}$$

onde o termo e^{-cx} e a constante A são sempre maiores do que zero. A curva dada pela eq. (2) é convexa crescente quando $b > 1$, uma curva côncava crescente quando $0 < b < 1$ e uma curva convexa decrescente quando $b < 0$ (Altmann, 1980). A eq. (2) pode assumir diferentes formas para $b = 0$, $b \neq 0$, $c = 0$ e $c \neq 0$, conforme a Tabela 1.

$b = 0$		$y = Ae^{-cx}$	modelo I	(3)
$b \neq 0$	$c = 0$	$y = Ax^b$	modelo II	(4)
	$c \neq 0$	$y = Ax^b e^{-cx}$	modelo III	(5)

Tabela 1: Soluções para diferentes possibilidades das constantes b e c (Altmann, 1980).

Cada uma das soluções apresentadas na Tabela 1 pode ser linearizada aplicando o logaritmo natural a ambos os lados, como nas eqs. (6) a (8):

$$\log y = \log A - cx \qquad \text{I} \tag{6}$$

$$\log y = \log A + b \log x \qquad \text{II} \tag{7}$$

$$\log y = \log A + b \log x - cx \qquad \text{III} \tag{8}$$

Note que, em cada uma das eqs. (6) a (8), tem-se uma relação linear entre as formas transformadas dessas variáveis:

$$\begin{aligned} y' = \log y &= \log A - cx \\ &= \beta_0 + \beta_2 x \end{aligned} \tag{9}$$

$$\begin{aligned} y' = \log y &= \log A + b \log x \\ &= \eta_0 + \beta_1 x' \end{aligned} \tag{10}$$

$$\begin{aligned} y' = \log y &= \log A + b \log x - cx \\ &= \beta_0 + \beta_1 x' + \beta_2 x \end{aligned} \tag{11}$$

onde $\beta_0 \triangleq \log A$, $\beta_1 \triangleq b$, $x' \triangleq \log x$ e $\beta_2 \triangleq -c$.

[3] Verifica-se a seguir que a Equação (2) é de fato solução:

$$\begin{aligned} {}^{\mathrm{d}y}/_{\mathrm{d}x} &= Abx^{b-1}e^{-cx} - cAx^b e^{-cx} \\ &= Ax^b e^{-cx}\left({}^{b}/_{x} - c\right) \\ &= y\left({}^{b}/_{x} - c\right), \end{aligned}$$

ou seja, obtém-se a Equação (1).

É possível, então, utilizar um modelo linear para relacionar a versão transformada da variável independente (variável explicativa), chamada de x', com a versão transformada da variável dependente (variável de resposta), chamada de y'. A essência desse modelo pode ser expressa, de forma geral, por

$$E[Y'|x] = \beta_0 + \beta_1 x' + \beta_2 x, \tag{12}$$

onde $E[\cdot]$ representa o valor esperado e $Y'|x$ indica a busca de possíveis valores de Y' (em que $Y' = \log Y$), o que restringe x a um único valor (consequentemente, x' também estará restrito). O parâmetro β_0 é o intercepto, β_1 e β_2 são as constantes de proporcionalidade em relação a cada um dos fatores (x' e x, respectivamente). Dado um conjunto de observações, é possível ajustar o modelo,[4] ou seja, encontrar os parâmetros β_0, β_1 e β_2 que melhor (sob algum critério a ser definido) explicam os dados. Para realizar o ajuste do modelo, deve-se ter um conjunto de dados, relacionando y e x.

Para uma regressão linear simples, a principal hipótese nula é $H_0 : \beta_1 = 0$ e $\beta_2 = 0$, ou seja, a média populacional de Y' é β_0 para todo valor de x, implicando que x não possui efeito em Y. A hipótese alternativa, dessa maneira, será H_1: $\beta_1 \neq 0$ e/ou $\beta_2 \neq 0$, implicando que mudanças em x acarretam mudanças em Y. Em alguns casos, é razoável considerar uma hipótese nula diferente, por exemplo, quando comparado com um padrão de referência. Nesse caso, a hipótese nula usualmente considerada é $H_0 : \beta_1 = 1$, com a hipótese alternativa $H_1 : \beta_1 \neq 1$. Neste trabalho, assume-se a seguinte hipótese nula: o comprimento médio das sílabas (em termos do número de fonemas que as constituem) é constante, para qualquer comprimento de palavra (em número de sílabas), ou seja, H_0: $\beta_1 = 0$ e $\beta_2 = 0$.

Para que o modelo I, dado pela Equação (3), descreva uma relação de decrescimento entre componentes e construtos, devemos ter $c > 0$. No caso do modelo II, dado pela Equação (4), devemos ter $b < 0$. Já o modelo III, dado pela Equação (5), terá derivada dada por

$$\begin{aligned} {}^{\mathrm{d}y}/_{\mathrm{d}x} &= Abx^{b-1}e^{-cx} - Acx^b e^{-cx} \\ &= (b - cx)Ae^{-cx}x^{b-1}. \end{aligned} \tag{13}$$

[4] Se o ajuste for realizado através do método dos mínimos quadrados, o critério escolhido será minimizar o erro quadrático médio; se o modelo for ajustado usando o método da máxima verossimilhança, busca-se maximar a probabilidade de que os dados observados sejam provenientes do modelo encontrado, ou seja, maximizar a verossimilhança.

Considerando que o comprimento do construto x não pode assumir valores negativos, para que a Equação (13) seja negativa, e assim exista uma relação de decrescimento, será necessário ter $(b - cx) < 0$, que poderá ocorrer quando $b < c$ e $c > 0$, considerando $x \geq 1$ ($x_{\min} = 1$), ou quando $b < cx_{\sup}$ e $c < 0$, onde $x_{\sup}$ é o limite superior para os valores que o comprimento do construto pode assumir.

Em algumas contextos, podemos analisar uma condição em que há um elemento hierarquicamente intermediário. Nesses casos, podemos obter uma relação decrescente entre construto e constituinte ou observar uma relação crescente entre eles (Altmann & Schwibbe, 1989; Prün, 1994; Grzybek & Stadlober, 2007). Suponha, então, que z seja nosso elemento intermediário entre y e x. Vamos analisar cada um dos modelos a seguir.

Para o modelo I, teremos:

$$y = Ae^{-cz}, \tag{14}$$

$$z = A'e^{-c'x}, \tag{15}$$

e, assim, podemos escrever y em função de x,

$$y = Ae^{-c(A'e^{-c'x})}. \tag{16}$$

Para que exista uma relação de decrescimento entre y e z, devemos ter $c > 0$. Da mesma forma, analisando z e x, devemos ter $c' > 0$. Para que também exista uma relação de decrescimento entre y e x, devemos analisar a derivada de Equação (16).

$$\mathrm{d}y/\mathrm{d}x = AA'cc'e^{-cA'e^{-c'x}-c'x}. \tag{17}$$

A Equação (17) será positiva para $c > 0$ e $c' > 0$ e, portanto, haverá uma relação crescente entre os vizinhos indiretos. Caso exista uma relação crescente apenas entre um dos vizinhos diretos, $c < 0$ ou $c' < 0$, haverá uma relação decrescente entre os vizinhos indiretos. Se existir uma relação crescente para ambos os vizinhos diretos, a relação também será crescente para os vizinhos indiretos.

Analisando agora o modelo II, teremos:

$$y = Az^b, \tag{18}$$

$$z = A'x^{b'}, \tag{19}$$

dessa maneira, a relação entre y e z será da forma

$$y = A(A')^b x^{b'b} = A''x^{b''}, \tag{20}$$

onde definimos $A'' = A(A')^b$ e $b'' = b'b$.

Para este modelo composto, poderemos ter uma relação crescente ou decrescente entre os vizinhos indiretos, dependendo do valor das constantes b e b'. Se os vizinhos diretos possuírem o mesmo tipo de relação entre eles, crescente ou decrescente, ou seja, $b > 0$ e $b' > 0$ ou $b < 0$ e $b' < 0$, respectivamente, teremos $b'' > 0$ e, por conseguinte, existirá uma relação crescente entre y e x, vizinhos indiretos. Se a relação entre os vizinhos diretos não for a mesma, teremos um expoente positivo e outro negativo, de forma que obteremos $b'' < 0$, uma relação decrescente entre vizinhos indiretos.

Para o caso mais geral, dado pelo modelo III (Equação (5)), teremos as seguintes relações entre vizinhos hierárquicos:

$$y = Az^b e^{-cz}, \tag{21}$$

$$z = A'x^{b'}e^{-c'x}. \tag{22}$$

A partir das eqs. (21) e (22), podemos estabelecer uma relação entre vizinhos indiretos y e x, nos mesmos moldes da Equação (5):

$$\begin{aligned} y &= A(A'x^{b'}e^{-c'x})^b e^{-c(A'x^{b'}e^{-c'x})} \\ &= A(A')^b x^{b'b} e^{-(c'bx+cA'x^{b'}e^{-c'x})} \\ &= A''x^{b''}e^{-(c''x+c'''x^{b'}e^{-c'x})}, \end{aligned} \tag{23}$$

onde definimos $A'' = A(A')^b$, $b'' = b'b$, $c'' = c'b$, $c''' = cA'$. O uso do modelo mais geral acaba levando a uma relação intrincada entre as variáveis. A derivada de y será dada por

$$\begin{aligned} \mathrm{d}y/\mathrm{d}x = -(&A(c'x - b')e^{A'cx^{b'}(-e^{-c'x})-c'x} \\ &(A'x^{b'}e^{-c'x})^b(be^{c'x} - A'cx^{b'}))/x. \end{aligned} \tag{24}$$

A relação entre x e y será decrescente se,

- $c'x - b' > 0$ e $be^{c'x} - A'cx^{b'} > 0$, ou
- $c'x - b' < 0$ e $be^{c'x} - A'cx^{b'} < 0$.

Em geral, os trabalhos que analisam a Lei de Menzerath utilizam os modelos simplificados e limitam suas análises a um mesmo sistema hierárquico, ainda que possa haver alguma sobreposição e interação entre sistemas hierárquicos distintos (Pike, 1967). Os modelos analisados buscam descrever a relação entre construtos e constituintes, sendo regidos por determinadas formulações matemáticas que utilizam constantes a serem determinadas ao ajustar o modelo aos dados observados. Entretanto, ainda não é claro qual é a relação entre as constantes e a interpretação delas sob a ótica da teoria da linguagem (Prün, 1994; Köhler, 1989; Altmann &

Schwibbe, 1989). Existem, entretanto, regiões de valores que aparentemente possuem relação com o nível linguístico de análise, havendo a formação de grupos quando observamos os parâmetros dos modelos (Cramer, 2005).

É importante ressaltar que, do ponto de vista linguístico, quando analisamos tais construtos, constituintes e suas relações, estamos diante de conjuntos e relações, de certa forma, difusos, sobretudo quando hierarquias lexicais e fonológicas são analisadas concomitantemente. As unidades linguísticas e a forma como seus sistemas hierárquicos se relacionam variam de uma língua para outra. Os modelos aqui propostos podem se adequar melhor ou pior a cada caso, não podendo ser considerados como uma forma de abarcar todas as nuances de uma língua.

2. Abordagem deste trabalho

Para realizar a análise da Lei de Menzerath em uma língua, deve-se, inicialmente, definir sob qual nível será realizada a análise. Este trabalho atém-se às unidades: palavras, sílabas e fonemas. Em uma análise fonológica, cada uma dessas unidades encontra-se em um nível hierárquico distinto da língua. Ainda que seja possível realizar análises no nível morfológico (Altmann, 1980; Gerlach, 1982; Polikarpov, 2000b; Krott, 1996), ou mesmo considerando a taxa de elocução (Menzerath, 1954), iremos nos ater a palavras, sílabas e fonemas, unidades essas decorrentes do nível fonológico, que estão acessíveis no Corpus ABG e que são escopo deste trabalho.

Embora uma análise simplista conceba a palavra como uma unidade situada entre dois espaços em branco, do ponto de vista linguístico, o conceito de palavra ainda não é claramente definido. Assumimos, aqui, a definição de palavra empregada por Cristófaro-Silva (2011, p. 169), apresentada no *Dicionário de Fonética e Fonologia*, a qual consiste em uma "unidade linguística que agrega som e significado em uma unidade. Pode ser compreendido [*sic*] como a menor unidade de significado em uma língua". Para Coulmas (2002, p. 38), "palavras são unidades na fronteira entre morfologia e sintaxe, possuindo importante função ao levar concomitantemente informação semântica e sintática, estando assim sujeitas à variação. Em algumas línguas, palavras parecem ser mais bem definidas e estáveis que em outras. A estrutura que constitui as palavras depende das características tipológicas das línguas".

Segundo Martinet (1978), as palavras podem ser segmentadas, do ponto de vista morfológico, em morfemas, que consistem na menor unidade linguística portadora de significado, e, do ponto de vista fonológico, em sílabas e em fonemas, sendo estes as menores unidades distintivas de significado. Os fonemas, dependendo do contexto em que ocorrem, podem ser expressos de diferentes formas, sendo estas chamadas de fones. Uma elocução qualquer pode ser descrita por uma sequência de fones que convencionalmente são representados graficamente através de um alfabeto fonético.

De modo semelhante à palavra, a definição de sílaba ainda é alvo de inúmeros debates na literatura linguística. Steriade (2002, p. 1) define-a como "uma sequência de segmentos agrupados em torno de uma vogal obrigatória ou de um elemento vocálico (silábico)", mas tal definição não é consensual. De forma que, para alguns linguistas, a sílaba é elemento central na teoria fonológica (Selkirk, 1982), enquanto para outros é um construto teórico desnecessário (Köhler, 1966). Não adentraremos aqui nesse embate teórico, tendo em vista que o corpus utilizado neste estudo já apresenta silabificação proveniente de um silabificador automático pautado na escala de sonoridade (Selkirk, 1984). Independente do método de silabificação empregado, seja por meio do dicionário, seja pelo silabificador automático, não se espera grandes discrepâncias que venham a prejudicar a análise realizada (Marchand et al., 2009).

No âmbito da escrita, as unidades da língua (fonemas, sílabas e palavras) usualmente são representadas por meio do seu sistema ortográfico, o qual requer a dissecação do fluxo da fala em partes distinguíveis. Para Coulmas (2002, p. 151), "todo sistema de escrita mapeia um sistema linguístico, incorpora e exibe visivelmente a dissecação das unidades da língua e, portanto, realiza uma análise linguística".

Embora os corpora linguísticos, em geral, apresentem apenas a representação ortográfica da palavra, o Corpus ABG, base de dados desta pesquisa, foi selecionado por já dispor da transcrição fonêmica das palavras, bem como de outras informações fonológicas. Deve-se salientar que, por questões metodológicas, foram utilizados caracteres do teclado para representar os fonemas, o que não significa que eles sejam grafemas em si. São, na verdade, apenas símbolos diferentes do próprio IPA (*International Phonetic Alphabet*), mas que têm a mesma função: representar e expressar os sons das línguas.

2.1. Tokenização

O primeiro problema com o qual se deve lidar ao trabalhar com um corpus escrito é o da *tokenização*. *Tokens* são realizações de uma determinada unidade, isto é, toda ocorrência de uma sequência idêntica de caracteres (unidades) representa uma realização de um tipo (entidade abstrata) na forma de um *token* (instância concreta). Estabelecer, a partir de uma sequência de caracteres, a sua divisão, eliminando os caracteres irrelevantes, como os de pontuação, não é uma tarefa trivial. A simples remoção da pontuação e a utilização dos espaços em branco para delimitar os *tokens* geram alguns resultados indesejados, como em palavras com apóstrofe e/ou com hífen. Por exemplo, como devemos lidar com as sequências: "*ex-presidente*", "*dona-de-casa*", "*arqui-inimigo*" e "*copo-d'água*"? Deveríamos separá-los e gerar os *tokens*: "*ex*", "*presidente*", "*dona*", "*de*", "*casa*", "*arqui*", "*inimigo*", "*copo*", "*d*" e "*água*"? Ou devemos tratar a sequência como um único *token*: "*ex-presidente*", "*dona-de-casa*", "*arqui-inimigo*" e "*copo-d'água*"? Observe que, ao utilizar o hífen e o apóstrofe como caracteres que delimitam a fronteira de palavra, geram-se palavras malformadas, como *arqui* e *d*. Para os casos de hífen, de forma geral, uma estratégia aparentemente mais apropriada seria assumi-las como palavras compostas e, conforme a análise linguística a ser realizada, considerá-las ou não no corpus em investigação - abordagem adotada neste estudo, uma vez que essa foi a abordagem utilizada para criar o Corpus ABG (Benevides & Guide, 2017).

2.2. Frequência de Ocorrência

Vários estudos mostram que o efeito de frequência é ubíquo nas áreas ligadas à cognição e ao comportamento humano (Sikström, 2002; Nosofsky, 1988; Ellis, 2015), sobretudo, na aquisição de linguagem (Ambridge et al., 2015; Ellis, 2002), no processamento de linguagem (Ellis, 2002) e nas mudanças linguísticas (Bybee, 2010). Especificamente, a frequência de ocorrência pode ser utilizada como um instrumento de análise quantitativa de vários aspectos linguísticos.[5] Segundo Bybee (2001), frequência de ocorrência consiste na quantidade de vezes que uma unidade, em geral uma palavra, ocorre em determinado corpus ou texto. Para calculá-la, deve-se contabilizar quantas vezes cada uma delas aparece em uma dada amostra. Ela vem sendo incorporada na descrição e na análise de diversos estudos linguísticos, tanto no estudo de língua adulta, como de aquisição de linguagem (Bybee, 1995, 2001; Pierrehumbert, 2003; Jarosz et al., 2016).

Na Linguística Quantitativa, diversas leis examinam o comportamento da frequência de ocorrência de tipos e sua relação com propriedades linguísticas. A mais conhecida é a Lei de Zipf, já mencionada na Seção 1. Zipf (1935, 1949) propôs, no âmago de sua teoria, o *Princípio do Esforço Mínimo*, sendo este responsável por explicar as observações que obteve sobre a frequência de ocorrência de palavras.

As unidades linguísticas, como fonemas, letras, sílabas, morfemas e palavras, podem ser estudadas através de suas frequências de ocorrência. Tal análise é possível até mesmo em estruturas que ocupam nível hierárquico mais alto. A frequência de ocorrência pode ser utilizada para inferir, por exemplo, sobre a distribuição subjacente da fonte que produz uma sequência de símbolos observada. Busca-se, assim, deduzir e quantificar a informação produzida por uma fonte. Esta também é usada em psicolinguística para explicar o fenômeno de recuperação lexical, considerado um dos processos centrais do processamento da linguagem, o qual consiste na transformação de um conceito abstrato em uma realização concreta, a elocução da palavra (Gleason & Ratner, 1998; Marantz, 2015). A frequência de ocorrência também é utilizada em outras áreas como no estudo do aprendizado, da organização e do desenvolvimento de uma língua (Phillips, 2006; Lieberman et al., 2007; Bybee, 2007, 2015).

2.3. Base de dados para o presente trabalho

Este trabalho verificou a aplicação da Lei de Menzerath em um corpus do português, analisando a relação entre o número de sílabas das palavras e o comprimento médio das sílabas, em fonemas. Para tanto, fez-se necessária uma base de dados que possuísse o número de sílabas e o número de fonemas para cada palavra, além de sua frequência de ocorrência no corpus. Utilizou-se como base de dados o Corpus ABG (Benevides & Guide, 2017).[6] Este é um corpus linguístico do português brasileiro (PB) que pode ser utilizado como fonte de extração de dados fonológicos, contendo, para cada palavra, sua frequência de ocorrência (oral e escrita), transcrição fonológica,

[5] Aprendizado, memorização, percepção, recuperação lexical, regularização, redução fonética, dentro muitos outros. Estes são apenas alguns exemplos da relevância e da atuação da frequência de ocorrência.

[6] O Corpus ABG está disponível no sítio `https://github.com/SauronGuide/corpusABG`.

codificação da estrutura da sílaba,[7] além de categoria morfológica, lema e acentuação. Mais informações sobre a criação, a estruturação e a utilização do corpus podem ser obtidas em Benevides & Guide (2017).

2.4. Lei de Menzerath no português brasileiro

São poucos os trabalhos que analisam a Lei de Menzerath no português. Por exemplo, Coloma (2015) faz um paralelo com 50 línguas distintas, dentre elas o português. O foco principal deste trabalho é comparar o ajuste de dois modelos distintos: o tradicional, de Menzerath-Altmann (Altmann, 1980), e o modelo hiperbólico proposto por Milička (2014). Utilizou, para isso, o conto "O Vento Norte e o Sol" de Esopo, em suas diversas traduções contidas no *Handbook of the International Phonetic Association*. Este livro é um guia sobre como utilizar o alfabeto fonético IPA e assim apresenta, além dos textos traduzidos, suas transcrições fonéticas. A versão traduzida para o português possui apenas 8 orações, 98 palavras e 380 fonemas. Ao analisar a relação entre fonemas por palavra e palavras por oração nas 50 línguas, Coloma (2015) concluiu que ambos os modelos apresentam igualmente bem a correlação negativa visualizada entre número de fonemas por palavra e palavras por oração.

Rothe-Neves et al. (2018), por sua vez, analisou a relação entre a duração média das sílabas em elocuções e o número de sílabas nas sentenças, a partir da elocução de 20 sujeitos para 40 sentenças, que variam de 2 a 29 sílabas. O trabalho buscou investigar a tendência geral de compressão da duração dos sons da fala em função do número de sons em uma sentença, independente do seu conteúdo linguístico. Para tanto, Rothe-Neves et al. (2018) ajustaram o modelo II (Equação (4)) para os dados de cada um dos sujeitos e também para o conjunto de todos os dados. Suas observações corroboraram a hipótese de independência entre os parâmetros A e b (Milička, 2014; Rothe-Neves et al., 2018), em oposição à argumentação de que tais parâmetros seriam dependentes da língua (Cramer, 2005; Kelih, 2010; Kułacka, 2010); e, sobretudo, corroboraram a Lei de Menzerath ao observar uma tendência de encurtamento na duração das sílabas conforme o alongamento dos enunciados.

Embora ambos os trabalhos sugiram que a Lei de Menzerath descreve adequadamente o comportamento de encurtamento do constituinte em relação ao construto, eles analisam dados em níveis hierárquicos linguísticos distintos, os quais também são diferentes daqueles que são abordados neste trabalho, e, principalmente, possuem uma base de dados de pequeno volume, o que compromete o estabelecimento de conclusões robustas. Diante dessa carência, este trabalho apresenta os resultados da aplicação da Lei de Menzerath em dados do português brasileiro, a partir do Corpus ABG, que "contabiliza 3.616.625 ocorrências de palavras e 92.602 tipos de palavras, sendo que 1.938.805 ocorrências são provenientes dos corpora de fala e 1.676.820 ocorrências dos corpora escritos" (Benevides & Guide, 2017, p. 1).

O corpus foi submetido ao *script* de tratamento e de criação do banco de dados desta pesquisa. Como o corpus já apresenta dados de transcrição fonológica e de estrutura silábica, o número de fonemas e o número de sílabas das palavras foram extraídos diretamente do corpus. Ao final, foi criada uma base de dados com informações de frequência de ocorrência, número de fonemas e número de sílabas para cada uma das palavras do Corpus ABG.

Os dados obtidos, a partir da relação entre o número de fonemas e o número de sílabas das palavras do PB, são expostos nas Figuras 1 e 2. Como o número de sílabas e o número de palavras são valores inteiros, o comprimento médio das sílabas só poderá assumir alguns valores fracionários possíveis. Têm-se, assim, apenas alguns níveis discretos de comprimento médio, o que resultou em uma grande sobreposição dos dados. Os gráficos de densidade expostos na Figura 2 apresentam, portanto, uma interpolação dos valores. Para representar graficamente a sobreposição de dados, nas Figuras 1 e 3, utilizou-se o tamanho dos círculos para representar o número de palavras encontradas com determinado número de sílabas e de fonemas. A Figura 1 apresenta também a distribuição marginal do número de fonemas e do número de sílabas no Corpus ABG.

O ajuste dos mínimos quadrados[8] aponta para uma relação decrescente entre o número de sílabas das palavras e o tamanho médio das sílabas, em número de fonemas. Os resultados estatísticos exibidos na Tabela 4 evidenciam que

[7] A codificação utilizada consistiu em: C para consoante, V para vogal, G para glide e S para as fricativas alveolares em posição de coda em final de palavra. Esta codificação foi empregada em decorrência da invisibilidade de tal segmento às regras acentuais (Bisol, 1994; Lee, 1995).

[8] O método dos mínimos quadrados aproxima um dado modelo (isto é, encontra o conjunto de parâmetros), buscando a solução que minimize a soma dos quadrados do resíduo.

palavra	frequência	transcrição fonética*	número de fones	tipo silábico	número de sílabas
de	125749	de	2	CV	1
que	116882	ke	2	CV	1
a	102779	a	1	V	1
o	91246	o	1	V	1
e	87868	e	1	V	1
é	61550	3	1	V	1
eu	46558	eW	2	VG†	1
do	46538	do	2	CV	1
não	43919	nAW	3	CVG	1
da	40205	da	2	CV	1
em	37053	EJ	2	VG	1
um	35188	U	1	V	1
você	29544	vo-se	4	CV-CV	2
na	29447	na	2	CV	1
com	29013	kO	2	CV	1
uma	28659	u-ma	3	V-CV	2
no	28427	no	2	CV	1
né	25291	n5	2	CV	1
assim	24666	a-sI	3	V-CV	2

* transcrição fonológica adotada no Corpus ABG Benevides & Guide (2017)
† Glide (ou semivogal).

Tabela 2: Exemplificação de alguns dados contidos no Corpus ABG.

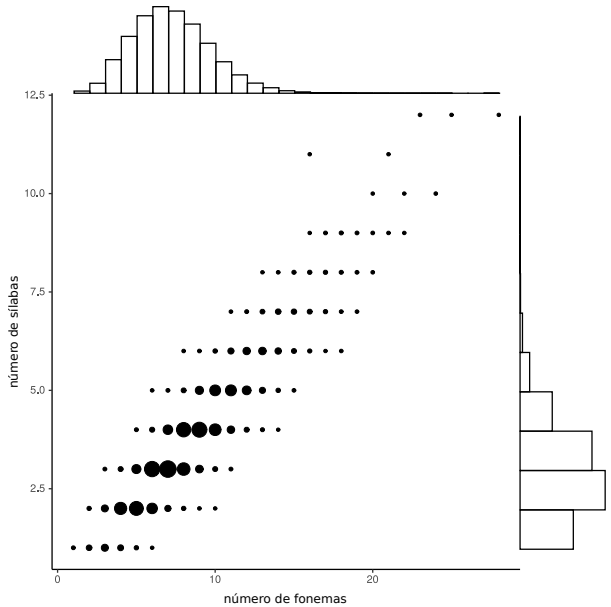
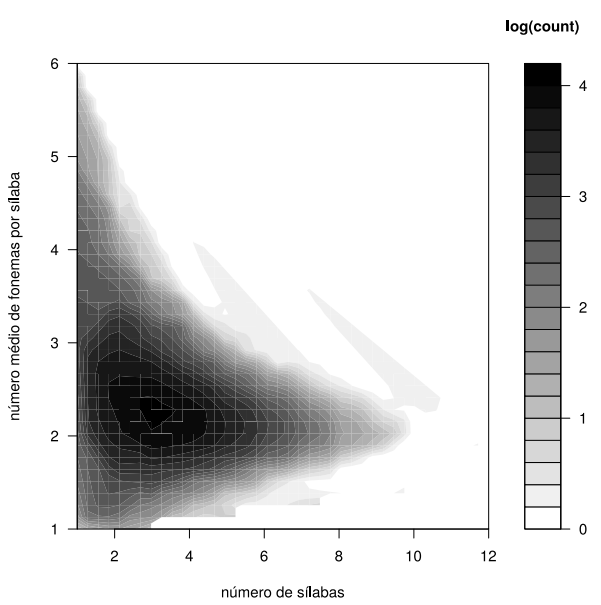


Figura 1: Relação entre o número de sílabas e o número de fonemas nas palavras do PB. A figura apresenta os círculos com tamanho proporcional ao número de palavras com cada uma das relações encontradas entre número de fonemas e sílabas. Nas laterais são apresentados os histogramas do número de sílabas e do número de fonemas.

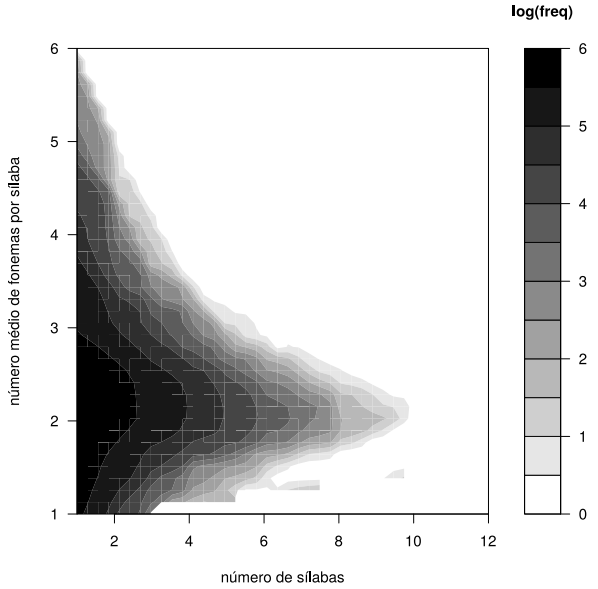
é possível descartar com segurança a hipótese nula: a hipótese de que não existe relação entre o número de sílaba e o tamanho médio das sílabas.[9] Em outros termos, observa-se que existe uma tendência a se utilizar sílabas menores em palavras com mais sílabas. Os modelos obtidos aqui também foram comparados utilizando ANOVA, sendo que os resultados apresentados na Tabela 3 evidenciam que o modelo III é mais explicativo para os dados observados.

Pelo gráfico dos resíduos, exibido nas Figuras 4a, 4c e 4e (veja Apêndice A), constata-se que, em média, o modelo é adequado, visto que os resíduos encontram-se centrados em zero, o que pode ser observado pela linha média dos resíduos ao longo do eixo da variável independente (número de sílabas). Note também que o resíduo não é sistematicamente grande ou pequeno em diferentes regiões, o resíduo está simetricamente distribuído em torno do zero, não sendo assim possível estabelecer alguma predição sobre os resíduos a partir da variável independente. Conclui-se, dessa maneira, que não há in-

[9] O coeficiente de determinação aparentemente é baixo, quando comparado aos valores observados na literatura. Entretanto, devemos observar que utilizamos aqui um volume de dados muito maior que o usual e, ainda mais, o modelo é determinado para todos os dados da amostra, enquanto na literatura obtém-se valores altos de R^2 tão somente por ajustarem o modelo às médias.



(a) Considerando a frequência de tipo das palavras.



(b) Considerando a frequência de ocorrência de cada palavra.

Figura 2: Relação entre o número de sílabas e o número médio de fonemas por sílabas observada nas palavras do Corpus ABG.[10]

formação explanatória do modelo sendo perdida e, consequentemente, sendo observada através dos resíduos.

Observa-se, porém, que a variância do resíduo cresce com a variável independente, indicando a presença de heterocedasticidade.[11] Isto pode invalidar os testes estatísticos de significância, pois estes pressupõem que os erros na modelagem são descorrelacionados e uniformes. A ausência de homoscedasticidade pode também indicar a existência de não-linearidade nos dados. Heteroscedasticidade usualmente ocorre quando há uma grande diferença no tamanho das observações. No caso em questão, quanto maior o número de sílabas de uma palavra, maior a variabilidade das sílabas usadas para construir essa palavra. Quando a palavra é pequena, a variabilidade é menor. A existência de heteroscedasticidade implica que o teorema de Gauss-Markov[12] não é válido para o caso em questão, de forma que é possível que o estimador linear de mínimos quadrados ordinário[13] utilizado não seja a melhor escolha, em termos de prover a menor variância dentre todos estimadores não polarizados. Com isso, observa-se que não há correlação entre o valor ajustado pelo modelo e o resíduo, como era esperado, e, ainda, que há heteroscedasticidade nos modelos propostos. Essencialmente, qualquer modelo é impreciso, desta forma, dentre as inúmeras opções de modelos que poderiam ser propostas, busca-se sempre o mais simples, que seja capaz de explicar os dados observados (princípio da Navalha de Occam). Talvez os modelos propostos possam ser melhorados se for acrescentada uma nova variável, por eles não abarcada, ou ainda, uma melhor estimativa das variáveis envolvidas poderia ser suficiente e, com isso, o efeito da heteroscedasticidade poderia ser diminuído, mantendo as demais qualidades dos modelos utilizados. Trabalhos como o de van Heuven et al. (2014) mostram a importância de

[10] Conforme exposto na Figura 2, é possível encontrar no Corpus ABG, e no português, palavras simples com 6, 7 e até mesmo 8 sílabas, como *pro.gres.si.va.men.te*, *pro.ble.ma.ti.za.re.mos* e *tra.di.ci.o.na.lís.si.mo*, ainda que elas sejam poucas. Há, ainda, palavras com extensão igual ou superior a elas, que são, em geral, palavras compostas, como *ex-pro.cu.ra.dor-ge.ral*, *la.ti.no-a.me.ri.ca.na* e *pre.si.dên.cia-e.xe.cu.ti.va*.

[11] Diz-se que há heterocedasticidade em um conjunto de variáveis aleatórias quando existe subpopulações com diferentes variabilidades. Se a variabilidade de uma variável não se mantém igual ao longo da extensão de uma segunda variável que a prediz, então, diz-se que há heterocedasticidade. Algumas possíveis causas da heterocedasticidade são: a própria natureza de algumas variáveis que apresentam tendência à heterocedasticidade, a existência de valores extremos e as falhas na especificação do modelo. A existência de heterocedasticidade pode comprometer alguns testes de significância em uma análise de regressão.

[12] O teorema de Gauss-Markov estipula que, em um modelo de regressão linear, sob certas condições, o estimador linear não polarizado com menor variância é dado pelo estimador dos mínimos quadrados ordinário, se existir.

[13] O estimador de mínimos quadrados ordinário é aquele que utiliza a soma dos quadrados das diferenças entre a variável dependente observada e o valor predito pelo modelo linear, ou seja, $\sum_x (y(x) - \hat{y}(x))^2$, onde x é a variável independente, y a variável dependente e $\hat{y}$ o valor predito pelo modelo.

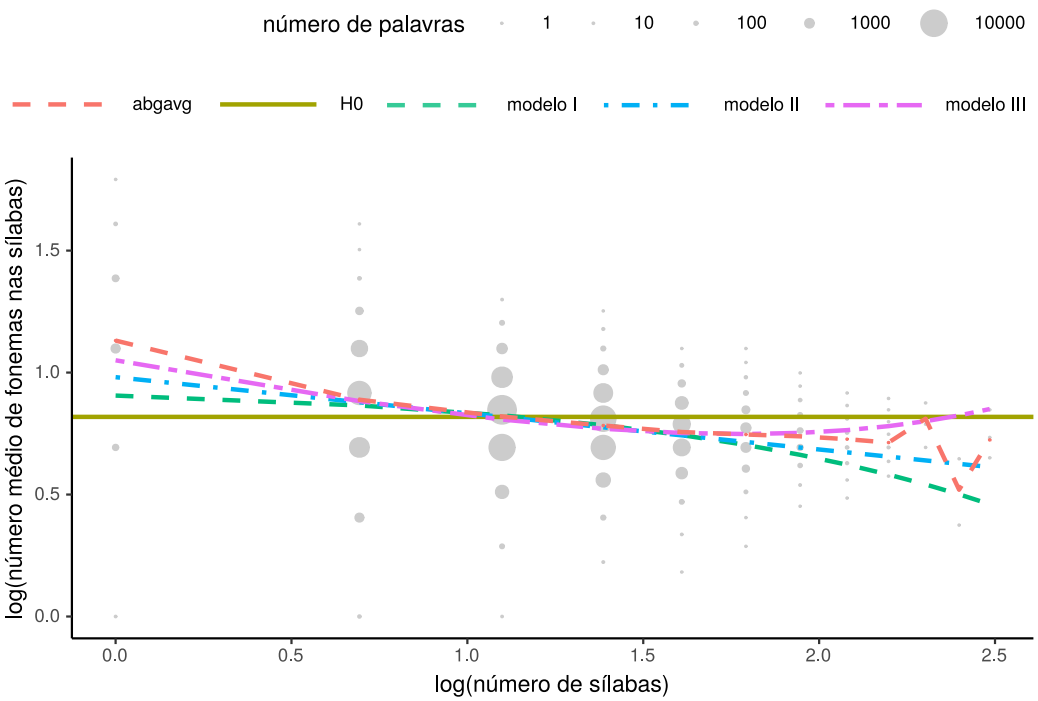


Figura 3: Relação entre o número de sílabas e o número médio de fonemas nas palavras do PB. São traçados os modelos I, II e III obtidos pelo ajuste de modelo linear a partir da base de dados ABG e a hipótese nula como referência.

	Res.Df	RSS	Df	Soma Qd.	F	Pr(>F)
I	92602	2096,3				
III	92601	2020,5	1	75,829	3475,3	$< 2{,}2 \times 10^{-16}$ *
II	92602	2043,7				
III	92601	2020,5	1	23,207	1063,6	$< 2{,}2 \times 10^{-16}$ *
Nota:						$^{*}p < 0{,}001$

Tabela 3: Análise de Variância (ANOVA) comparando modelo I com modelo III e comparando modelo II com modelo III para os dados do Corpus ABG.

uma boa estimação de variáveis como frequência de ocorrência, comprimento e similaridade de palavras, uma vez que a melhor estimação dessas variáveis é capaz de agregar maior explicação sobre a variância dos dados do que a inclusão de novas variáveis.

O gráfico Q-Q normal (*normal quantile-quantile plot*), apresentado nas Figuras 4b, 4d e 4f (veja Apêndice A), revela um desvio em direção a uma distribuição com cauda longa, o que pode ser constatado pela observação de valores mais extremos do que o esperado, caso os dados fossem provenientes de uma distribuição normal. Avaliar a normalidade dos dados é importante, pois muitos procedimentos de inferência estatística presumem que as amostras, resultantes de um conjunto fixo de valores de uma variável dependente, provenham de uma distribuição normal. Violações severas dessa suposição podem levar a resultados errados em valor de p e de intervalo de confiança. Embora, no caso em questão, constate-se a não normalidade dos resíduos, ainda assim a normalidade é uma suposição razoável.

2.5. A Lei de Menzerath frente aos dados do português

No presente trabalho, conforme apresentado na seção 2.4, buscou-se verificar a veracidade da Lei de Menzerath frente aos dados de um corpus linguístico do português brasileiro. A Lei prediz que o número de fonemas nas sílabas diminui à medida que o número de sílabas da palavra aumenta. A fim de testar essa predição, utilizou-se como fonte de dados o Corpus ABG.

Observa-se, de maneira geral, que o número de fonemas diminui à medida que o número de sílabas aumenta, o que corrobora a Lei de Menzerath. Tal proporcionalidade é claramente vi-

	Variável dependente:		
	log(comprimento médio em fonemas)		
	(I)	(II)	(III)
número de sílabas	0,041*		-0,054*
	(0,0004)		(0,002)
log(número de sílabas)		−0,148*	-0,319*
		(0,001)	(0,005)
Constante	0,946*	0,981*	0,996*
	(0,002)	(0,002)	(0,002)
Observações	92.604	92.604	92.604
R^2	0,091	0,114	0,124
R^2 Ajustado	0,091	0,114	0,124
Erro padrão residual	0,150 (df = 92.602)	0,149 (df = 92.602)	0,148 (df = 92.601)
Estatística F	9.281,210* (df = 1; 92.602)	11.904,500* (df = 1; 92.602)	6.552,366* (df = 2; 92.601)
Nota:			*$p < 0,01$

Tabela 4: Resultados do ajuste dos modelos lineares para os dados do Corpus ABG.

sualizada na Figura 2a, a qual expressa uma maior densidade de tipos de palavras de até 4 sílabas com até 3 fonemas (por sílaba). É interessante notar que a redução do número de fonemas por sílaba conduz, consequentemente, a predileção por sílabas mais simples, em geral, CV, V e CVC. Essa preferência é esperada, tendo em vista que, segundo Crystal (1988), as duas primeiras sílabas são universais nas línguas naturais. No português, as três figuram entre as mais frequentes, com índices, respectivamente, de 192.532, 26.907 e 24.055 em termos de frequência de tipo,[14] segundo dados do Corpus ABG (Benevides & Guide, 2017).

Além de nos mostrar que, quanto maior a palavra, menor a quantidade de fonemas por sílaba, as distribuições expostas nas Figuras 2 e 3, a partir da Lei de Menzerath, demonstram que, quanto maior a estrutura da palavra, menor a quantidade de tipos de palavra. Isto é, as estruturas de palavras mais frequentes da língua tendem a ter estruturas silábicas mais simples (CV). Tal afirmativa é visualizada no Corpus ABG, o qual apresenta as seguintes estruturas mais frequentes: CV-CV-CV (5.243 tipos), CV-CV (3.553), CV-CV-CV-CV (3.229), CV-CV-CVS (1.671) e V-CV-CV-CV (1.625). Note que a única estrutura com ramificação de rima ocupa a quarta posição e ainda assim é preenchida por *s*, marcador, em geral, de plural, o qual é tido por diversas propostas como invisível a alguns fenômenos fonológicos, como o acento (Massini-Cagliari, 1992; Bisol, 1994). Tal correlação poderia ser estendida, no caso do português, a frequência das palavras, tendo em vista que, segundo Araújo et al. (2007), os pentassílabos figuram entre as palavras com índice de frequência mais raro da língua, com 41,5%, em comparação a 21% dos trissílabos, por exemplo. A raridade de frequência das palavras tende a aumentar conforme aumenta as suas extensões, como demonstra a Figura 2b.

Diante disso, destaca-se, no presente artigo, a aplicabilidade da Lei de Menzerath para o corpus do português, o que permite tecer análises quantitativas relevantes com relação à estrutura da palavra e das sílabas. A Linguística Quantitativa, dessa maneira, permite realizar descrições essenciais às análises fonológicas em geral.

[14] Assume-se, aqui, a concepção de frequência de tipo de (Bybee, 2001, p. 10), segundo a qual "frequência de tipo refere-se à frequência de dicionário de um padrão específico"; neste caso, a quantidade de tipos de palavras que possui semelhante estrutura silábica.

3. Conclusão

O presente estudo não fornece resposta a todas as questões em aberto que permeiam a Lei de Menzerath-Altmann na linguagem. Buscou-se aqui analisar os modelos frente aos dados do português brasileiro; a partir dos quais verificamos a existência de uma tendência à diminuição do número de constituintes na composição de construtos maiores, corroborando assim a Lei de Menzerath. No contexto de uma teoria geral da comunicação, argumenta-se que mecanismos de percepção e de cognição devam ser considerados para explicar essa observação. Zipf (1935, 1949) usa o *Princípio do Esforço Mínimo* como fundamento para suas observações. Mais tarde, Köhler (1989) utiliza um modelo de processamento de linguagem para fundamentar as observações feitas através da Lei de Menzerath. O processamento de linguagem é sequencial, ao menos no nível mais baixo, uma vez que a fala se realiza sequencialmente ao longo do tempo, através da sucessão linear de perturbações acústicas com características que se modificam no decorrer do tempo. Além disso, Köhler (1989) utiliza como argumento a capacidade finita de processamento e memória em tarefas cognitivas. Esse raciocínio leva à conclusão de que construtos mais complexos e mais longos necessitam utilizar-se de constituintes menores e mais simples. Desta forma, não apenas se observa a tendência à utilização de constituintes mais simples na composição de construtos mais complexos, como também uma menor variabilidade. Já na construção de construtos mais simples, a variabilidade dos constituintes é maior. Tal argumentação entra em consonância com o limite da memória de curta duração (Miller, 1956), segundo o qual o número de objetos que em média uma pessoa é capaz de manter na memória de trabalho é de 7 ± 2, a chamada Lei de Miller. As tarefas analisadas por Miller (1956) mostram que há queda na performance à medida que o número de estímulos diferentes (variando apenas um dos atributos) aumenta para além de cinco ou seis. A capacidade de memória imediata, o presente psicológico, é dito ter uma duração de 1,5 a 3 segundos, operando principalmente no nível de processamento de sentenças. Além disso, para palavras em que a relação entre o número de sílaba e o número de fonemas se mantém constante, a duração temporal é fator determinante para a recuperação de palavras, apresentando melhor desempenho aquelas de curta duração (Baddeley et al., 1975). Essa análise sugere a necessidade de se analisar também a duração de palavras e de sentenças em elocuções para verificar se a Lei de Menzerath também se faz presente.

Nos modelos aqui contemplados, através da análise dos resíduos, é possível verificar que não são satisfeitas as considerações de heteroscedasticidade e de normalidade dos resíduos. Isto indica que possivelmente existem fatores que não foram abarcados pelo modelo proposto, ou que é necessário obter medidas mais acuradas das variáveis envolvidas, ou ainda que é necessário modificar o modelo, utilizando algum termo de outra sorte para conseguir um melhor ajuste. A frequência de ocorrência de palavras pode ser utilizada para ponderar a relação entre o número de sílabas em palavras e o comprimento médio das sílabas em fonemas. A influência dessa nova variável pode alterar a relação observada entre comprimento médio das sílabas e o comprimento das palavras. Essa variável, ainda não considerada, pode inclusive ser fator causador da heterocedasticidade observada. Outros fatores importantes também poderiam ser considerados, como a frequência de ocorrência de sequências de palavras, a complexidade articulatória, a duração e a distintividade linguística de fonemas, sílabas e palavras. Afinal, fatores relacionados à produção e à percepção também possuem papel importante no uso da linguagem.

Deve-se ter em mente a separação entre os três processos centrais que foram abarcados no presente trabalho: a seleção do modelo, a estimação dos parâmetros e a predição de resultados a partir do modelo e dos parâmetros dos dados. O modelo correto nunca será conhecido, mas, como ponderou Box (1976), "alguns são úteis". A análise de um problema não deve consistir na aplicação dos três passos descritos uma única vez, mas deve retornar ao passo anterior sempre que se verificar falsas suposições que foram rejeitadas nas etapas seguintes. Durante a construção e a aplicação de um modelo, muitas vezes adotam-se generalizações; esse processo pode ser perigoso, pois incertezas estão inseridas em distintas etapas.

A partir do estudo aqui depreendido sobre a relação entre construto e constituinte na comunicação restritos ao âmbito fonológico, constatase que as observações quantitativas e o tratamento matemático tornam-se necessários para a descrição de fenômenos que não podem ser representados por um arquétipo de uma determinada categoria, nem explicado por regras simbólicas estruturais, ou seja, é uma abordagem necessária para estudar a variabilidade e a imprecisão nas línguas naturais. Prefere-se lidar aqui com tendências e com preferências a lidar com relações estáveis de estruturas bem definidas. As relações dinâmicas e a variabilidade revelam mais sobre o fenômeno do que o funcionamento rígido de um sistema estrutural bem definido, uma vez que os modelos tesos e austeros culminam em uma descrição imprecisa e inconsistente. As expressões quantitativas permitem uma melhor adequação à realidade que aquelas qualitativas, permitem ainda uma análise mais fina em diferentes resoluções, uma gradação contínua de formatos representacionais que, sob outra ótica, seriam discretos e muitas vezes impossibilitariam o estabelecimento de inter-relações entre diferentes níveis de análise. A análise quantitativa da linguagem busca explorar os fundamentos, estabelecer explicações e construir uma teoria consistente com hipóteses refutáveis.

Referências

Abe, Sumiyoshi & Norikazu Suzuki. 2005. Scale-free statistics of time interval between successive earthquakes. *Physica A: Statistical Mechanics and its Applications* 350(2–4). 588–596. 10.1016/j.physa.2004.10.040.

Adamic, Lada A. & Bernardo A. Huberman. 2002. Zipf's law and the internet. *Glottometrics* 3. 143–150.

Altmann, Eduardo G. & Martin Gerlach. 2016. Statistical laws in linguistics. Em M. Degli Esposti, E.G. Altmann & F. Pachet (eds.), *Creativity and Universality in Language*, 7–26. Springer. 10.1007/978-3-319-24403-7_2.

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. *Glottometrika* 2(2). 1–10.

Altmann, Gabriel & Michael H. Schwibbe. 1989. *Das Menzerathsche Gesetz in informationsverarbeitenden systemen*. Hildesheim: Olms.

Ambridge, Ben, Evan Kidd, Caroline F. Rowland & Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language* 42(2). 239–273. 10.1017/s030500091400049x.

Andres, Jan. 2010. On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics* 17(2). 101–122. 10.1080/09296171003643189.

Araújo, Gabriel A. de, Zwinglio O. Guimarães Filho, Leonardo Oliveira & Viaro M. Eduardo. 2007. As proparoxítonas e o sistema acentual do português. Em *O acento em português: abordagens fonológicas*, 37–60. Parábola.

Baddeley, Alan D., Neil Thomson & Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* 14(6). 575–589. 10.1016/s0022-5371(75)80045-4.

Beckner, Clay, Nick C. Ellis, Richard Blythe, John Holland, Joan Bybee, Jinyun Ke, Morten H. Christiansen, Diane Larsen-Freeman, William Croft & Tom Schoenemann. 2010. Language is a complex adaptive system: Position paper. Em Nick C. Ellis & Diane Larsen-Freeman (eds.), *Language as a Complex Adaptive System*, 1–26. University of Michigan: Wiley-Blackwell.

Benevides, Aline De Lima & Bruno Ferrari Guide. 2017. Corpus ABG. *Texto Livre: Linguagem e Tecnologia* 10(1). 139–163. 10.17851/1983-3652.10.1.139-163.

Bisol, Leda. 1994. O acento e o pé métrico. *Letras de Hoje* 29(4). 25–36.

Boroda, Mojsej G. & Gabriel Altmann. 1991. Menzerath's law in musical texts. *Musikometrica* 3. 1–13.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71(356). 791–799. 10.1080/01621459.1976.10480949.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455. 10.1080/01690969508407111.

Bybee, Joan. 2001. *Phonology and language use* Cambridge Studies in Linguistics. Cambridge University Press. 10.1017/CBO9780511612886.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. USA: Oxford University Press. 10.1093/acprof:oso/9780195301571.001.0001.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge University Press. 10.1017/cbo9780511750526.

Bybee, Joan. 2015. *Language change* Cambridge Textbooks in Linguistics. Cambridge University Press. 10.1017/CBO9781139096768.

Coloma, Germán. 2015. The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics* 28. 139–159.

Coulmas, Florian. 2002. *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press. 10.1017/CBO9781139164597.

Cramer, Irene. 2005. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics* 12(1). 41–52. 10.1080/09296170500055301.

Cristófaro-Silva, Thaís. 2011. *Dicionário de fonética e fonologia*. Contexto.

Crystal, David. 1988. *Dicionário de linguística e fonética*. J. Zahar Editor.

Ellis, Nick C. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition* 24(2). 143–188. 10.1017/s0272263102002024.

Ellis, Nick C. 2015. Cognitive and social aspects of learning from usage. Em *Usage-Based Perspectives on Second Language Learning*, 49–74. De Gruyter. 10.1515/9783110378528-005.

Fenk, August & Gertraud Fenk-Oczlon. 2013. Menzerath's Law and the constant flow of linguistic information. Em Reinhard Köhler & Burghard B. Rieger (eds.), *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics*, Springer. 10.1007/978-94-011-1769-2_2.

Ferrer-i-Cancho, Ramon. 2006. On the universality of Zipf's law for word frequencies. Em P. Grzybek & R. Köhler (eds.), *Exact methods in the study of language and text. In honor of Gabriel Altmann*, 131–140. Gruyter.

Ferrer-i-Cancho, Ramon. 2017. The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics* 39. 38–71.

Ferrer-I-Cancho, Ramon & Núria Forns. 2009. The self-organization of genomes. *Complexity* 15. 34–36. 10.1002/cplx.20296.

Ferrer-i-Cancho, Ramon & Ricard V. Solé. 2002. Zipf's law and random texts. *Advances in Complex Systems* 5(1). 1–6. 10.1142/S0219525902000468.

Gabaix, Xavier. 1999. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* 114(3). 739–767. 10.1162/003355399556133.

Gerlach, Rainer. 1982. Zur Überprüfung des menzerath'schen gesetzes im bereich der morphologie. *Glottometrika* 4. 95–102.

Gleason, Jean Berko & Nan Bernstein Ratner. 1998. *Psycholinguistics*. Fort Worth: Harcourt Brace College Publishers.

Glottopedia. 2019. The free encyclopedia of linguistics. http://www.glottopedia.org/.

Grzybek, Peter. 2007. *Contributions to the science of text and language: Word*

length studies and related issues Text, Speech and Language Technology. Springer. 10.1007/978-1-4020-4068-9.

Grzybek, Peter & Gabriel Altmann. 2002. Oscillation in the frequency-length relationship. *Glottometrics* 2. 97–107.

Grzybek, Peter & Ernst Stadlober. 2007. Do we have problems with arens' law? a new look at the sentence-word relation. Em Peter Grzybek & Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text*, 205–218. Walter de Gruyter. 10.1515/9783110894219.205.

Gustison, Morgan L., Stuart Semple, Ramon Ferrer i Cancho & Thore J. Bergman. 2016. Gelada vocal sequences follow menzerath's linguistic law. *Proceedings of the National Academy of Sciences of the USA* 113(19). E2750–E2758. 10.1073/pnas.1522072113.

Heaps, Harold Stanley. 1978. *Information retrieval, computational and theoretical aspects* Library and information science. USA: Academic Press.

Herdan, Gustav. 1960. *Type-token mathematics*, vol. 4 Janua linguaarum, studia memoriae Nicolai van Wijk dedicata. Series maior. Mouton & Cie.

van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology* 67(6). 1176–1190. 10.1080/17470218.2013.850521.

Hřebíček, Luděk. 1995. *Text levels: Language constructs, constituents and the menzerath-altmann law* Quantitative linguistics. Wissenschaftlicher Verlag Trier.

Ilari, Rodolfo. 2003. *A lingüística e o ensino da língua portuguesa*. São Paulo: Martins Fontes.

Jarosz, Gaja, Shira Calamaro & Jason Zentz. 2016. Input frequency and the acquisition of syllable structure in polish. *Language Acquisition* 24(4). 361–399. 10.1080/10489223.2016.1179743.

Jespersen, Otto. 1904. *Lehrbuch der phonetik*. Leipzig: Teubner.

Kelih, Emmerich. 2010. Parameter interpretation of the menzerath law: evidence from serbian. Em Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and Language: Structures, Functions, Interrelations: Quantitative Perspectives*, 71–79. Praesens.

Köhler, Konrad. J. 1966. Is the syllable a phonological universal? *Journal of Linguistics* 2(2). 207–208. 10.1017/S0022226700001493.

Köhler, Reinhard. 1989. Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. Em Gabriel Altmann & Michael H. Schwibbe (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Hildesheim: Olms.

Köhler, Reinhard. 2005. Gegenstand und Arbeitsweise der Quantitativen Linguistik. Em Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik. Ein internationales Handbuch*, 1–16. de Gruyter.

Krott, Andrea. 1996. Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics* 3(1). 29–37. 10.1080/09296179608590061.

Kułacka, Agnieszka. 2010. The coefficients in the formula for the menzerath-altmann law. *Journal of Quantitative Linguistics* 17(4). 257–268. 10.1080/09296174.2010.512160.

Lee, Seung Hwa. 1995. *Morfologia e fonologia lexical do português do Brasil*. Campinas: Universidade Estadual de Campinas. Tese de Doutoramento.

Li, Wentian. 2011. Menzerath's law at the geneexon level in the human genome. *Complexity* 17(4). 49–53. 10.1002/cplx.20398.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163). 713–716. 10.1038/nature06137.

Lü, Linyuan, Zi-Ke Zhang & Tao Zhou. 2010. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE* 5(12). e14139. 10.1371/journal.pone.0014139.

Marantz, Alec. 2015. Morphology. Em Gregory Hickok & Steven L. Small (eds.), *Neurobiology of Language*, chap. 13, 153–163. Academic Press.

Marchand, Yannick, Connie R. Adsett & Robert I. Damper. 2009. Automatic syllabification in english: A comparison of different algorithms. *Language and Speech* 52(1). 1–27. 10.1177/0023830908099881.

Martinet, André. 1978. *Elementos de lingüística geral*. Martins Fontes.

Massini-Cagliari, Gladis. 1992. *Acento e ritmo*. São Paulo: Contexto.

Menzerath, Paul. 1928. Über einige phonetische Probleme. Em *Actes du premier Congres International de Linguistes*, 104–105.

Menzerath, Paul. 1954. *Die architektonik des deutschen Wortschatzes* Phonetische Studien. F. Dümmler.

Meyer, Ernst Alfred. 1904. Zur Vokaldauer im Deutschen. Em Adolf Gotthard (ed.), *Nordiska studier tillegnade Adolf Noreen på hans 50-årsdag den 13 Mars 1904*, Wentworth Press.

Milička, Jiří. 2014. Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics* 21(2). 85–99. 10.1080/09296174.2014.882187.

Miller, George A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63(2). 81–97. 10.1037/h0043158.

Mitzenmacher, Michael. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1(2). 226–251. 10.1080/15427951.2004.10129088.

Nikolaou, Christoforos. 2014. Menzerath-altmann law in mammalian exons reflects the dynamics of gene structure evolution. *Computational Biology and Chemistry* 53, Part A. 134–143. 10.1016/j.compbiolchem.2014.08.018.

Nosofsky, Robert M. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(1). 54–65. 10.1037/0278-7393.14.1.54.

Phillips, Betty. 2006. *Word frequency and lexical diffusion*. Basingstoke England New York: Palgrave Macmillan.

Pierrehumbert, Janet. 2003. Probabilistic phonology: Discrimination and robustness. Em Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probability Theory in Linguistics*, 177–228. Cambridge MA: The MIT Press.

Pike, Kenneth Lee. 1967. *Language in relation to a unified theory of the structure of human behavior* Janua Linguarum. Series Maior. De Gruyter.

Polikarpov, Anatoliy A. 2000a. Menzerath's law for morphemic structures of words: A hypothesis for the evolutionary mechanism of its arising and its testing. Em *Qualico: Quantitative Linguistics Conference*, .

Polikarpov, Anatoliy Anatolyevich. 2000b. Menzerath's law for morphemic structures of words: A hypothesis for the evolutionary mechanism of its arising and its testing. Em *Qualico: Quantitative Linguistics Conference*, .

Prün, Claudia. 1994. Validity of menzerath-altmann's law: Graphic representation of language, information processing systems and synergetic linguistics. *Journal of Quantitative Linguistics* 1(2). 148–155. 10.1080/09296179408590009.

Rothe-Neves, Rui, Bárbara Marques Bernardo & Robert Espesser. 2018. Shortening tendency for syllable duration in brazilian portuguese utterances. *Journal of Quantitative Linguistics* 25(2). 156–167. 10.1080/09296174.2017.1360172.

Roudet, Léonce. 1910. *Éléments de phonétique générale*. Welter.

Selkirk, Elisabeth. 1982. The syllable. Em Harry van der Hulst & Norval Smith (eds.), *The structure of phonological representations*, Foris.

Selkirk, Elisabeth. 1984. On the major class features and syllable theory. Em Mark Aronoff & Richard T. Oehrle (eds.), *Language Sound and Structure*, The MIT Press.

Shahzad, Khuram, Jay E. Mittenthal & Gustavo Caetano-Anollés. 2015. The organization of domains in proteins obeys Menzerath-Altmann's law of language. *BMC Systems Biology* 9. 44. 10.1186/s12918-015-0192-9.

Sikström, Sverker. 2002. Habituation during encoding of episodic memory. Em *Connectionist Models of Cognition and Perception*, 107–117. 10.1142/9789812777256_0009.

Steriade, Donca. 2002. The syllable. Em William Frawley (ed.), *International Encyclopedia of Linguistics*, Oxford University Press.

Wilde, Joachim & Michael H. Schwibbe. 1989. Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. Em Gabriel Altmann & Michael H. Schwibbe (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, 92–107. Olms.

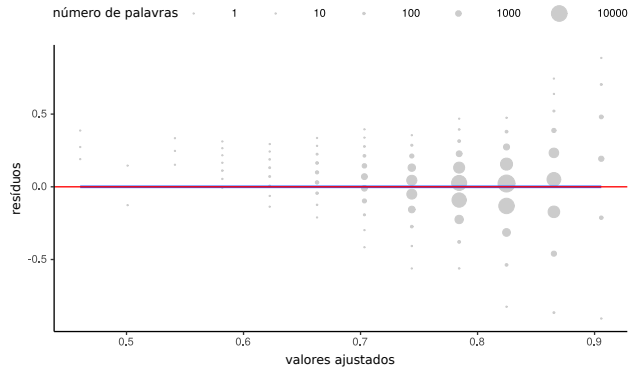
Zipf, George Kingsley. 1935. *The psycho-biology of language: an introduction to dynamic philology*. The MIT Press.

Zipf, George Kingsley. 1949. *Human behaviour and the principle of least effort: An introduction to human ecology*. Hafner Pub. Co.

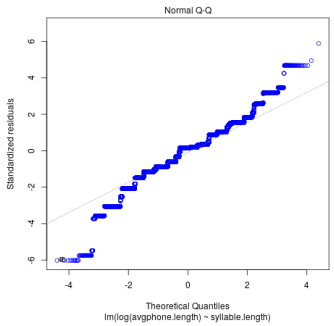
A. Dados, scripts e outros resultados

Os dados, os *scripts* desenvolvidos para a realização deste trabalhos e todos os resultados estão disponíveis em repositórios no GitHub.

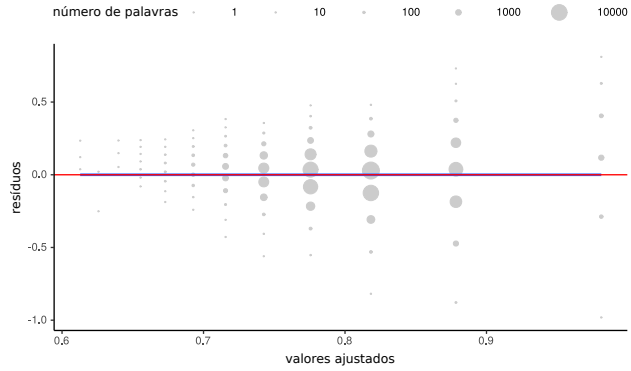
1. Corpus ABG: https://github.com/SauronGuide/corpusABG.
2. *Scripts* voltados para a área de Linguística Quantitativa e Computacional: https://github.com/leolca/clscripts.
3. *Notebook* com anotações, códigos e resultados gerados para este trabalho: https://github.com/leolca/clscripts/blob/master/menzerath_abg.ipynb.



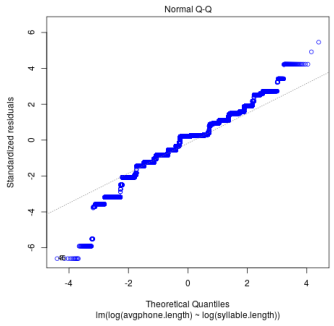
(a) Gráfico dos resíduos para o modelo I.



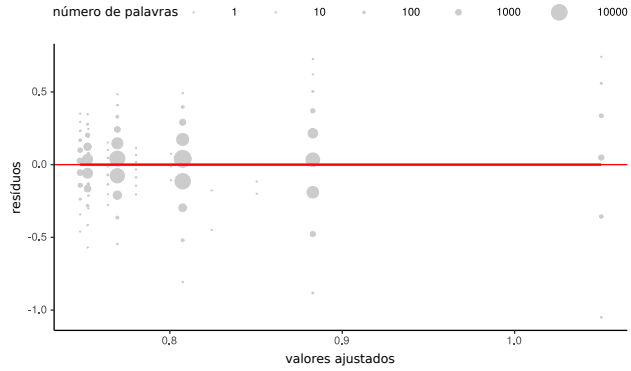
(b) Gráfico QQ para o modelo I.



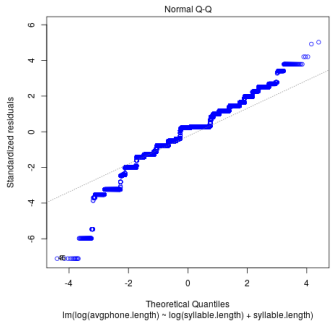
(c) Gráfico dos resíduos para o modelo II.



(d) Gráfico QQ para o modelo II.



(e) Gráfico dos resíduos para o modelo III.



(f) Gráfico QQ para o modelo III.

Figura 4: Análise dos resíduos para os modelos utilizando os dados do Corpus ABG.