Universidade Federal Rural de Pernambuco
Departamento de Física

Pós-graduação em Física Aplicada

ASPECTOS ESTATÍSTICOS DA DISTRIBUIÇÃO ESPACIAL DE PALAVRAS EM LINGUAGEM ESCRITA

Maelyson Rolim Fonseca dos Santos

DISSERTAÇÃO DE MESTRADO

Recife
13 de agosto de 2014

Universidade Federal Rural de Pernambuco
Departamento de Física

Maelyson Rolim Fonseca dos Santos

ASPECTOS ESTATÍSTICOS DA DISTRIBUIÇÃO ESPACIAL DE PALAVRAS EM LINGUAGEM ESCRITA

Trabalho apresentado ao Programa de Pós-graduação em Física Aplicada do Departamento de Física da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do grau de Mestre em Física.

Orientador: *Prof. Dr. Pedro Hugo de Figueirêdo*

Recife
13 de agosto de 2014

Para Bruna, Fagna e Maurício.

AGRADECIMENTOS

Sou grato ao professor Pedro Hugo pela orientação e pela disposição ao longo de madrugadas, feriados e pontos facultativos. Porém não somente pela qualidade no exercício da docência mas principalmente devido às ótimas conversas sempre temperadas com uma boa dose de humor. Obrigado por me ensinar a rir ao encontrar no *arXiv* aquele resultado *inédito* feito por um grego ou chinês. Não importando se o tema fosse ciência, livros, cinema, música, política ou qualquer outro assunto que nos passasse pela cabeça, o fato é que, entre tapiocas e pastéis, o aprendizado foi enorme.

Durante a graduação ouvi que *"É preciso muita coragem para decidir fazer Física, mas é preciso muita sorte para ter uma família que lhe apóie."*. Se a coragem eu fui juntando ao longo dos anos de formação, a sorte eu encontrei desde cedo com painho e mainha. Dona Fagna e Seu Maurício saibam que faltam-me palavras e sobram lágrimas para agradecer todo o Amor que recebo. Afinal, só mesmo lá em casa pra mesa do café ser espaço para falar de política, ciência e tudo mais que um garoto agitado possa querer inventar. Cada conquista que obtenho é mérito de vocês.

São seis os tios que me acompanham e eles merecem meus eternos agradecimentos. Obrigado Tias Corrinha, Cristina e Liduina, Tios Eraldo, Fagno e Sérgio.

Nem meus esquecimentos e a falta de tempo conseguiram abalar o fato de quão felizardo eu sou por ter Bruna Peralva ao meu lado. Muito obrigado por ficar firme em meios as minhas mais agoniadas balançadas de cabeça. Jamais serei capaz de recompensar tanto carinho.

Obrigado Dona Romana, Seu Giko e Laís!

Como não ser grato ao irmão genial que a vida me deu? Valeu Zé! Valeu Cássio!

Obrigado Pedro e Tati, é uma honra ter vocês na discagem rápida do celular. Novamente obrigado por ter contado com vocês quando a estrada fez a curva mais díficil.

Obrigado Victor Pessôa pela inspiração perene.

Five Physics Boys pode ser um nome horrível mas nem isso diminui a admiração, o carinho e o respeito — com boas doses de infantilidades, é claro — que tenho para com Adson, Daniel, Mário e Milton. Obrigado e que sigam anos de risadas de nossas boas histórias.

Aos amigos da graduação: Tiago Araújo, Hugo, Brenda, Raphael, Cinthia, Ademar, Poli, Allan, Saulo, Chico, Pablo, obrigado pela parceria e pelos dias compartilhados.

E quem foi que disse que a vida não pode ser feita de noites de sexta-feira? Obrigado Abel, Aline, Amanda, Bela, Biel, Camila, Débora, Elis, Emmanoel, Héber, Heloísa, Guiga (meu pirraia), Gustavo, Jampa, Joana, Jônathas, Karine, Lídia, Malu, Messias, Nathan, Paula, Paulão, Pri, Rafa, Raiana, Thiago, Thiago e todo mundo que esqueci de colocar aqui, por sempre me estimularem a questionar o que me impede de pensar.

Conexões com São Paulo, Goiás e o Espírito Santo animam qualquer dia. Obrigado Clarinha, Bruna e Eunice pela companhia.

Valeu @aiade, @keerolz, @joaocarrara,@isaacpalma1, @marceliniia e @filipe_machado!

Obrigado Leo. Como dizem lá pras bandas da Área II: *"As palavras convencem, o exemplo arrasta"* :D

Obrigado Ceça e Tereza. É uma alegria dividir conquistas com quem me guarda no coração e nas preces.

Sou grato aos Professores Ailton, Anderson, Wictor e Ramon pelo conhecimento compartilhado mas sobretudo pelas conversas e cafés...

Obrigado grande Neto por conseguir solucionar todos os tipos de problemas burocráticos que consegui arrumar durante o mestrado.

Foi um prazer dividir os dias e as dores de cabeça do mestrado com os companheiros Aguinaldo, Augusto PM, Augusto Rubin, Chico *Buttiker*, Cosmo, Danilo, Ivelton, Izabelly, James, Magda, Marilia, Raphael e Tiago.

Valeu Paul Zaloom e Mark Rits!

Obrigado Eunice, Edmilson, Lindaci, Emanuel, John, Cristiane Pífano, Marcelo Leite e Ernesto Raposo. Seus grandiosos exemplos seguem me inspirando.

As melhores horas de trabalho foram aquelas com trilha sonora. Obrigado Kings Kaleidoscope, Mahmundi, Palavrantiga, Emicida, Tanlan, Fresno, Gungor, Tiago Cavaco, Mombojó, Medulla, Mumford & Sons, Eddie, Nação Zumbi, Projota, Topaz, Criolo, Deolinda, Faringes da Paixão, Jon Foreman, Banda Sinfônica da Cidade do Recife, Xangai, Alceu Valença, Caetano, Mutemath, Arcade Fire e tantos outros que me fizeram imergir na *Vida Secreta de Maelyson Rolim.*

Obrigado aos podcasters do Braincast, Diversitá, MacMagazine no Ar, 45 Minutos, RapaduraCast, Cinema em Cena, AntiCast, BTCast e Fronteiras da Ciência. As idas e vindas nos ônibus de Recife são enriquecedoras ouvindo vocês.

Obrigado Miguel Gonçalves por *José e Pilar*.

Sou grato a CAPES pelo fomento.

Justo González escreveu: *"Segundo os filósofos gregos, o que faz com que o mundo seja inteligível é que tanto a mente quando o mundo participam do mesmo Logos. É graças a esse Logos que sabemos que dois e dois são quatro. E é também graças a ele que em todo universo dois e dois são quatro. Sem o Logos, dois e dois não seria quatro, e minha mente não poderia sabê-lo."*. Assim, ao *Logos* seja toda minha gratidão. Obrigado *Verbo* por ter me feito viver nesse momento da história e nesse local do Universo!

ἐν ἀρχῇ ἦν ὁ λόγος
No princípio era o Verbo
—JOÃO (Evangelho)

RESUMO

A investigação do processo de evolução e caracterização das diversas linguagens humanas tem sido um dos campos mais ativos de pesquisa nas últimas décadas. Embora a busca por padrões linguísticos que possam estabelecer uma filogenia das línguas seja bem mais antiga, a caracterização estatística da linguagem escrita, comumente denominada *linguística quantitativa*, possui uma tradição mais recente que se apoia nos trabalhos desenvolvidos por George Zipf e Claude Shannon, escritos no final da década de 1940. Nesta dissertação investigamos aspectos frequencistas e espaciais da distribuição de verbetes em textos e o papel destas quantidades sobre a informação contida em linguagem escrita. Num primeiro momento exploramos a relação de escala entre o vocabulário V e o tamanho dos textos T, denominada Lei de Heaps, que segundo nossos resultados é típica para cada língua. Estabelecemos empiricamente, uma relação funcional entre a frequência máxima k_{max} e o número total de palavras do texto T. Num segundo momento analisamos características morfológicas dos símbolos obtendo a distribuição de tamanho $P(l)$ dos verbetes e a partir desta a sua respectiva entropia, concluimos que este procedimento nos permite categorizar diferentes grupos linguísticos. Por fim introduzimos dois modelos capazes de fornecer comportamentos limitantes universais, para a relação entre a intermitência σ e a frequência k dos verbetes. Os modelos foram concebidos de forma a descrever o comportamento de verbetes correlacionados e não correlacionados, reproduzindo diversas propriedades de textos como a fração de verbetes correlacionada f e a entropia estrutural $\overline{H}$. Ao longo de nossa abordagem, todos os nossos resultados teóricos foram comparados com aqueles obtidos de um *corpus* composto por 500 textos, que incluem artigos da *wikipedia* e obras literárias de diversas épocas, em 10 idiomas distribuídos em 3 famílias linguísticas: germânica (alemão, dinamarquês, inglês e sueco), latina (espanhol, italiano, francês e português) e urálica (finlandês e húngaro).

Palavras-chave: entropia, linguística quantitativa, mecânica estatística

ABSTRACT

The investigation of the process of evolution and characterization of different human languages has been one of the most active research fields in recent decades. Although the search for linguistic patterns that can establish a phylogeny of languages is much older, the statistical characterization of the written language, commonly called quantitative linguistic, has a newer tradition that relies on the work developed by Claude Shannon and George Zipf, written at the end of the 1940s. In this work we investigate some statistical aspects of the frequencies and positions for words in texts and the function of this quantities into the information contained in written language. Initially we explored the scaling relationship between the vocabulary V and the text sizes T, called Heaps' Law, which according to our results is typical for each language. We establish, empirically, a functional relationship between maximum frequency k_{max} and the total number of words in the text. Secondly we analyze morphological features of symbols, obtaining the word sizes distribution and from its respective entropy. We conclude that this procedure allows us to categorize different linguistic groups. Finally we introduce two models able to provide universal limiting behaviors to the relationship between standard deviation σ and frequency k. The models were designed to describe the behavior of correlated and uncorrelated words, reproducing various properties of texts as the fraction f of correlated words and the structural entropy $\overline{H}$. All our theoretical results were compared with those obtained from 500 texts that include wikipedia articles and literary works from various epochs in 10 languages distributed in three linguistic families: germanic (german, danish, swedish and english), romanic (spanish, italian, french and portuguese) and uralic (finnish and hungarian).

Keywords: entropy, quantitative linguistics, statistical mechanics

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO 1

INTRODUÇÃO

Não tenhas pressa e não percas tempo.
—JOSÉ SARAMAGO

1.1 PRELIMINARES

Mais notadamente nas três últimas décadas, a física tem participado de uma intensa colaboração com as demais áreas do conhecimento. Tal processo de intercâmbio que vai das ciências sociais aplicadas como a economia até as ciências biológicas, passando pela psicofísica, pela sociologia, pela linguística e pelo urbanismo, tem permitido avanços não apenas na descrição e previsão dos fenômenos destas áreas, como também tem ajudado na compreensão de problemas fundamentais do escopo da física, com o desenvolvimento de uma série de ferramentas teóricas e experimentais. Esses sistemas complexos, como têm sido denominados, são usualmente constituídos por um número considerável de elementos que interagem por meio de mecanismos não triviais, compartilhando características fundamentais como heterogeneidade, adaptabilidade, frustração e memória [1, 2]. Uma dessas profícuas relações tem ocorrido com o campo da linguística.

A linguagem é uma das mais importantes características da espécie humana foi esta particularidade que nos possibilitou interpretar e moldar o mundo de acordo com muitas de nossas necessidades, estabelecendo uma série de mecanismos complexos de comunicação. Estima-se que existam hoje cerca de 7103 línguas vivas desse total aproximadamente 12.75% encontram-se em vias de extinção [3]. Dentro dessa diversidade destacam-se as línguas escritas, muitas vezes denominadas *linguagem natural*, que acredita-se terem surgido por volta de 3500 A.C. [4]. Sua criação possibilitou que a própria linguagem pudesse ser registrada e discutida e o seu estudo é o principal foco dos filólogos, linguistas e filósofos da linguagem [5]. Por outro lado a escrita também permitiu a representação concreta de idéias abstratas, o que atualmente é um grande campo de estudo da psicologia e da neurologia [6, 7, 8, 9].

Uma terceira e importante característica da linguagem escrita é que seu registro permite uma análise quantitativa de suas propriedades estruturais. Essa é uma das particularidades que tem atraído a atenção de físicos, estatísticos, matemáticos e teóricos da informação, seja pela similaridade com problemas fundamentais de suas áreas ou pela possibilidade de aplicação de técnicas oriundas delas [10, 11, 12].

Embora a busca por padrões linguísticos que possam estabelecer uma filogenia das línguas seja bem mais antiga, a caracterização estatística da linguagem escrita, comumente denominada *linguística quantitativa*, possui uma tradição mais recente que se apoia nos trabalhos desenvolvidos por George Zipf [13, 14] e Claude Shannon [15], escritos no final da década de 1940, com contribuições posteriores de B. Mandelbrot [16].

O campo da linguística quantitativa tornou-se uma grande área de investigação por parte da comunidade de físicos a partir da primeira década do século XXI, com o estabelecimento de diversos modelos [17, 18] e relações de tamanho na tentativa de explicar leis empíricas já conhecidas pelos linguistas. Dentre essas leis, descatam-se a lei de Zipf, que estabelece uma relação de escala para o número n de verbetes que aparecem k vezes num dado texto:

$$n(k) \sim k^{-\beta} \tag{1.1}$$

onde $\beta \approx 2$ é o exponte de Zipf, e a lei de Heaps [19] que relaciona o número total V de verbetes num texto com o número total de palavras T:

$$V \sim T^{\lambda} \tag{1.2}$$

onde λ é o expoente de Heaps.

A despeito da grande diversidade de tópicos, podemos classificar os temas em linguística quantitativa em três grandes categorias. O primeiro grande tema diz respeito a cognição, características universais e dinâmica da evolução da linguagem escrita. Para esse estudo são empregados métodos quantitativos para investigação de fenômenos oriundos da psicologia, da neurologia e da teoria da informação, como por exemplo: a relação entre o comprimento das palavras e a eficiência da comunicação [20], aspectos da evolução sintática e do vocabulário e o processamento de dados envolvendo as linguagens naturais [21, 22, 23, 24]. Também são estudados os efeitos da tradução na complexidade de textos [25, 26] e aplicações em outras áreas como a biologia molecular e a genética [27]. Algumas abordagens visam a utilização de princípios fundamentais como o da extremização da entropia na tentativa de explicar estas relações empíricas bem como estabelecer medidas para a complexidade [28, 29].

A deteção e extração de palavras-chave e autoria formam o segunda grande categoria em linguística quantitativa. Paralelamente ao desenvolvimento de métodos e teorias, o aperfeiçoamento de recursos computacionais de alto desempenho tem permitido um grande avanço no tratamento, processamento e interpretação de uma grande quantidade de informação gerada pela humanidade e em particular da linguagem escrita [30, 31]. A produção deste grande volume de dados tornou necessária a elaboração de técnicas para a extração de conteúdos específicos tanto em linguagens naturais, quanto em estruturas biológicas. Especificamente em textos, o problema da extração de palavras-chave tem uma longa tradição na área de linguística quantitativa. Podemos tomar como marco inicial o trabalho proposto por H. Luhn [32], onde uma métrica baseada na frequência dos verbetes foi estabelecida para construção automática de resumos. Mais recentemente diversas propostas, fundamentadas em termos da distribuição espacial dos símbolos, têm sido formuladas de modo a tornar mais eficiente os métodos de deteção de palavras-chave [33, 34, 17, 35, 36, 37, 38].

A terceira grande área, estudo da diversidade linguística, consiste na formulação de modelos que expliquem de que modo se dá a dinâmica e o surgimento das distribuições espaciais de línguas [39, 40].

1.2 SOBRE A ESTRUTURA DESTA DISSERTAÇÃO

Usualmente a principal metodologia para quantificação da informação contida num texto tem como base o cálculo da entropia de Shannon H [15] para as frequências dos verbetes que o compõe. Embora seja robusto e indique uma certa diversidade do vocabulário, este parâmetro não captura a informação associada a estrutura (disposição espacial dos termos). Recentemente alguns trabalhos têm sugerido que a utilização de métodos que incluam o contexto da palavra, ou seja, o estudo da vizinhaça de um determinado termo, pode revelar aspectos importantes sobre a informação contida numa mensagem escrita [20].

Dentre os nossos objetivos está investigar o papel da distribuição espacial dos verbetes sobre a informação contida nos textos, bem como estabelecer aspectos estatísticos associados a frequência dos verbetes e de seus comprimentos, identificando suas possíveis características universais e/ou particulares para diferentes grupos linguísticos. A estrutura desta dissertação está disposta como se segue.

No Capítulo 2 apresentamos os fundamentos necessários para o estudo estatístico da linguagem escrita, discutimos as conexões entre entropia e informação e introduzimos a Lei de Zipf e algumas estimadores utilizadas para deteção de palavras-chave.

No Capítulo 3, realizamos um estudo estatístico a partir da frequência dos verbetes de um texto. Analisamos o coeficiente de Zipf para todas as dez línguas que compõem o *corpus* (conjunto de textos escritos em uma determinada língua e que serve como base de análise). Em seguida, discutimos a dependência do número de verbetes e da frequência máxima de ocorrência com relação ao número total de palavras de um texto. Ademais analisamos a distribuição de probabilidade dos tamanhos dos verbetes bem como a entropia de Shannon associada a essa distribuição e discutimos de que modo esse parâmetro pode ser utilizado para caracterização de grupos linguísticos.

No Capítulo 4 abordamos o problema de como descrever a distribuição espacial de palavras em um texto. Investigamos inicialmente a relação entre o desvio padrão σ e a frequência k de ocorrência de um verbete, em seguida propomos dois modelos capazes de descrever os comportamentos limitantes para essa relação. Também analisamos os efeitos das correlações espaciais sobre a informação relativa a estrutura.

Por fim, no Capítulo 5, serão apresentadas as conclusões deste trabalho e expostas suas perspectivas.

1.3 SOBRE O CORPUS

Como amostras de linguagem escrita foram selecionados 500 textos em dez línguas. O *corpus* é dividido em dois conjuntos: 250 obras literárias (Apêndice A) e 250 artigos da *Wikipedia* — um projeto de enciclopédia coletiva universal e multilíngue estabelecido na Internet — (Apêndice B). Tal seleção se encontra numa região intermediária de tamanho de *corpus*. Na literatura relativa ao tema são encontrados trabalhos com abordagens tendo como base apenas um texto [17] assim como conjuntos compostos por milhões de livros [31, 22].

Os idiomas foram selecionandos a partir de três grupos linguísticos: família germânica

(alemão, dinamarquês, inglês e sueco), família latina (espanhol, francês, italiano e português) e família urálica (finlandês e húngaro). Segundo Steven Roger Fischer [41]: *"As famílias são grupos de línguas geneticamente relacionadas. Ou seja, que dividem um ancestral comum, demonstrado por meio de correspondências sistemáticas, em forma e significado, não atribuíveis a mudanças ou apropriações."*

Como caracteres válidos foram considerados todas as letras do alfabeto com suas possíveis acentuações e os algarismos de 0 a 9. Palavras compostas tiveram o símbolo "-" computado como válido. Em nossa análise assumimos que uma palavra é o conjunto contínuo de símbolos válidos compreendidos entre dois espaços em branco. Para a preparação do nosso corpus, foi retirada toda a pontuação de cada texto e em seguida todas as letras foram transformadas em maiúsculas, conforme procedimento comum à linguistica quantitativa [42, 23].

Nos próximos capítulos faz-se necessário observar a diferença conceitual entre os termos "verbete" e "palavra". O primeiro refere-se a um conjunto distinto de símbolos válidos, enquanto o segundo é aplicado para as várias ocorrências desse mesmo conjunto. Assim nos versos *"A cidade não para, a cidade só cresce"*[1], contamos seis verbetes (A, CIDADE, NÃO, PARA, SÓ, CRESCE) e oito palavras.

No Apêndice C são apresentados os valores obtidos a partir desse *corpus* para um conjunto de 11 parâmetros discutidos ao longo da dissertação.

[1] A Cidade — Chico Science e a Nação Zumbi

CAPÍTULO 2

FUNDAMENTOS

Quem entender a linguagem entende Deus
cujo filho é Verbo. Morre quem entender.
—ADÉLIA PRADO (Antes do Nome)

2.1 CONCEITOS PRELIMINARES

Nos séculos XVIII e XIX, o continente europeu foi palco de diversas mudanças sociais e econômicas, motivadas sobretudo pela emergência do modo de produção capitalista. Esse período histórico é marcado pela Revolução Industrial que teve como um de seus principais pilares a indústria têxtil. Essa relação é de suma importância para entendermos a enfâse dada ao estudo da termodinâmica pelos cientistas dessa época. Como afirma J. D. Bernal citado por Rocha et al.: *"Na realidade, quanto mais estreitas são as relações entre a ciência, a técnica, a economia e a política do período, mais claramente se mostram a formação de um processo único de transformação da cultura. Tal período é de capital importância para o progresso da humanidade."* [43].

Foi nesse contexto que Sadi Carnot pesquisou, culminando com a apresentação dos seus resultados na obra *Reflexões sobre a potência motriz do fogo*, cuja introdução escrita por Javier Odin Ordónez apresenta a motivação do engenheiro militar Carnot para estudar termodinâmica: *"Seu interesse era fundar uma ciência geral para tratar os problemas da transformação calor potência motriz em sua maior universalidade, utilizando as leis obtidas em sua análise para justificar aspectos relevantes que eram conhecidos na ciência do calor da época."* [43].

Dentre os trabalhos que ajudaram a estabelecer a termodinâmica e que foram fundamentados na herança de Carnot podemos destacar os desenvolvidos de Rudolf J. Clausius e William Thomson (Lord Kelvin). Clausius, em 1865, propôs a existência de uma grandeza S tal que:

$$dS = \frac{đQ}{T}, \tag{2.1}$$

onde $đQ$ seria a quantidade de calor fornecida a um sistema à temperatura T [44]. Em suas palavras: *"Nós podemos definir S como o conteúdo transformacional de um corpo, assim como nós definimos a magnitude U como seu conteúdo térmico e ergonal. Porém, como eu considero ser melhor utilizar termos em línguas antigas para magnitudes importantes, de modo que eles podem ser adotados sem alterações em todas as línguas modernas, proponho chamar a magnitude S de entropia do corpo, a partir da palavra grega* $\tau\rho o\pi\grave{\eta}$*, transformação. Eu intencionalmente escolhi a palavra entropia de modo a ser o mais semelhante possível à palavra energia; as duas magnitudes a serem denotadas por*

estas palavras têm tão próximos seus significados físicos, que uma certa semelhança na designação parece ser desejável." [1][45].

Duas décadas antes dos trabalhos de Clausius, Helmholtz formalizou o que ficou conhecido como primeira lei da termodinâmica: Para um processo termodinâmico há uma função de estado, chamada energia interna (U), tal que:

$$dU = TdS + YdX + \sum_j \mu'_j dN_j, \tag{2.2}$$

onde Yé uma força mecânica generalizada, X denota um deslocamento generalizado, μ'_j é uma força química e dN_j é um deslocamento químico [46].

Proposto inicialmente para quantificar o calor dissipado por um sistema físico durante uma transformação termodinâmica, o conceito de entropia foi extendido para fenômenos fora do equilíbrio por Ludwig Boltzmann. Em sua formulação, a entropia de um sistema físico é função do número (Ω) de estados, associados a uma mesma energia configuracional, acessíveis ao sistema tal que:

$$S = -k_B \ln \Omega, \tag{2.3}$$

onde a constante k_B que define a unidade da entropia é conhecida como constante de Boltzmann.

Como exemplo do cálculo da entropia S, vamos considerar o seguinte sistema de dois níveis: R bolas distribuídas em dois compartimentos distintos com R_1 bolas no primeiro compartimento e $R_2 = R - R_1$ bolas no segundo compartimento. Para o caso em que as bolas são distinguíveis, o número possível de combinações de bolas retiradas será:

$$\Omega = \frac{R!}{R_1!R_2!} \tag{2.4}$$

e portanto:

$$S = -k_B \ln \frac{R!}{R_1!R_2!}. \tag{2.5}$$

Considerando R grande o suficiente, podemos utilizar os primeiros termos da série assintótica de Stirling:

$$\ln n! = n \ln n - n + \mathcal{O}(\ln n) \tag{2.6}$$

para escrever:

$$S \approx -k_B R \left(\frac{R_1}{R} \ln \frac{R_1}{R} + \frac{R_2}{R} \ln \frac{R_2}{R} \right). \tag{2.7}$$

[1] "*We might call S the transformational content of the body, just as we termed the magnitude U its thermal and ergonal content. But as I hold it to be better to borrow terms for important magnitudes from the ancient languages, so that they may be adopted unchanged in all modern languages, I propose to call the magnitude S the entropy of the body, from the Greek word τροπὴ, transformation. I have intentionally formed the word entropy so as to be as similar as possible to the word energy; for the two magnitudes to be denoted by these words are so nearly allied in their physical meanings, that a certain similarity in designation appears to be desirable.*"

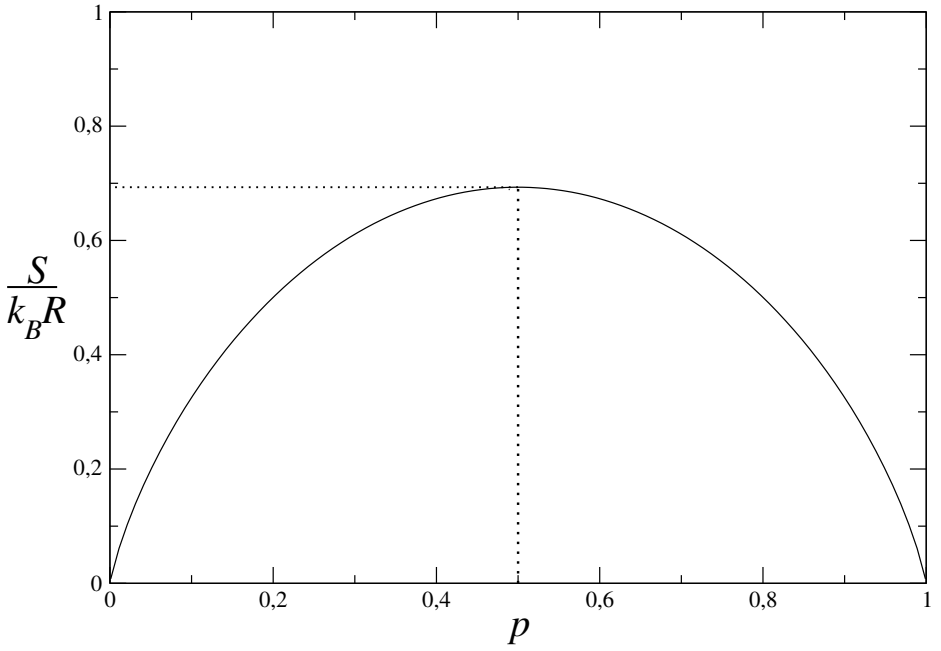


Figura 2.1 Gráfico da entropia S, para o modelo da urna de Ehrenfest com dois compartimentos, como função da razão $p = R_1/R$.

Na Figura 2.1 observamos que o valor mínimo da entropia ($S = 0$) é obtido quando $R_1 = 0$ ou $R_2 = 0$ enquanto obteremos o valor máximo ($S = k_B \ln 2$) quando $R_1/R = R_2/R = 0,5$. Portanto, a entropia é minimizada quando todas as bolas ocuparem o mesmo compartimento e maximizada quando as bolas forem igualmente distribuídas entre os dois compartimentos. Definindo $f_i = R_i/R$, podemos estender o modelo para K compartimentos:

$$S = -k \ln \frac{R!}{R_1!R_2! \cdots R_K!} \approx -kR \sum_{i=1}^{K} f_i \ln f_i. \tag{2.8}$$

Paul e Tatyana Ehrenfest foram os responsáveis por associar de modo definitivo o conceito de entropia à ideia de ordem (e desordem) de uma sistema. Porém, como alerta Feynman: *"A ordem não é ordem no sentido que nós gostamos do arranjo, mas no sentido que o número de maneiras diferentes que podemos arrumá-lo, tal que ele ainda pareça o mesmo no exterior."* [47].

Os trabalhos de Claude Shannon, publicados a partir da década de 40 do século passado, lançaram as bases da teoria da informação e ampliaram os conceitos de entropia, ordem e informação. No artigo *A Mathematical Theory of Communication* [15], Shannon demonstrou que, dada uma distribuição de probabilidade $p(r_i)(i = 1, 2, ..., N)$, a grandeza:

$$H = -\sum_{i=1}^{N} p(r_i) \log_2 p(r_i) \tag{2.9}$$

associada ao evento r_i, safistaz à seguinte série de propriedades coerentes com a definição

de incerteza associada com a distribuição:

- H é uma função contínua de $p(r_i)$;
- Quando todas as probabilidades são iguais (por exemplo: $p(r_i) = N^{-1}$), H cresce como função de N;
- Quando cada evento $p(r_i)$ pode ser tomado como a probabilidade conjunta de ocorrência de dois eventos independentes r_i^I e r_i^{II}, a função H divide-se em duas contribuições correspondentes a cada conjunto de sub-eventos: $H = H_I + H_{II}$.

Associado a essa formulação, Shannon apresentou a definição do conceito de informação tendo como unidade o *bit*. Podemos entender informação como sendo o ganho (ou perda) de certeza durante um processo onde nosso conhecimento acerca de um conjunto de eventos varia. Suponhamos que num lançamento de uma moeda saibamos num primeiro momento que a probabilidade de sair cara ou coroa são iguais ($p(cara) = p(coroa) = 0.5$). Assim da equação 2.9, temos: $H_0 = 1$ *bit*. Suponha agora que descubramos que se trata de um moeda desonesta e que na verdade: $p(cara) = 0.3$ e $p(coroa) = 0.7$, assim: $H \approx 0,88$ *bits*. Observe que entropia diminuiu assim como nossa incerteza acerca do resultado do lançamento. A diferença entre as entropias é a informação ganha e que podemos quantificar:

$$I = H_0 - H. \tag{2.10}$$

Nesse caso: $I \approx 0,12$ *bits*.

O termo *bit* (originalmente *binary digits*) definido por Claude Shannon como *"uma unidade de medida da informação"* foi cunhado por John W. Turkey. Em termos da equação 2.10, um *bit* pode ser entendido como o total de informação ganha quando o resultado de um evento com dois resultados igualmente prováveis ($p_1 = p_2 = 0.5 \rightarrow H_0 =$ 1 bit) se torna conhecido ($p_1 = 1$ e $p_2 = 0 \rightarrow H = 0$ bit). Ao longo desta dissertação será utilizada a seguinte formulação da equação 2.9:

$$H = -\sum_{i=1}^{N} p(r_i) \ln p(r_i) \tag{2.11}$$

seguindo o que se tornou padrão na literatura sobre o tema. É importante salientar que a proposta de Shannon inspirou a definição de uma série de entropias e medidas de informacão [48].

Sobre a importância desse trabalho pioneiro, James Gleick escreveu: *"A teoria de Shannon construiu uma ponte entre a informação e a incerteza; entre a informação e a entropia; e entre a informação e o caos. Levou aos CDS e aos aparelhos de fax, aos computadores e ao ciberespaço, à Lei de Moore e a todas as empresas pontocom do mundo. Assim nasceu o processamento de informações, junto com o armazenamento de informações e o acesso à informação."* [10].

2.2 LINGUÍSTICA QUANTITATIVA

Os trabalhos de George Kingsley Zipf (1902-1950) são tidos como o ponto inicial da relação moderna entre linguística e estudos estatísticos. Zipf foi um linguista e filólogo americano cujos trabalhos publicados nas décadas de 30 e 40 do século XX fomentaram os primeiros estudos das propriedades estatísticas de textos literários. Em sua obra *Human behavior and the principle of least effort* [13] que tinha como principal objetivo apresentar *"alguns princípios fundamentais que parecem governar aspectos importantes de nosso comportamento, como indivíduos e como membros de grupos sociais"*, Zipf propôs estudar textos literários, ou palavras mais precisamente. É importante destacar que o interesse naquele momento tinha pouca ou nenhuma relação com os estudos físicos, o principal ponto era que o estudo das palavras oferecesse elementos para a compreensão do processo da fala e, a partir daí, para a compreensão da personalidade e de todo o campo da dinâmica social.

Em seus trabalhos, Zipf faz uso do princípio do mínimo esforço para explicar como se dá o processo de comunicação. Em suas palavras: "[uma pessoa sempre irá] *esforçar-se para resolver seus problemas de maneira a minimizar o trabalho total que ele deve gastar em resolver os seus problemas imediatos e de seus prováveis problemas futuros*" [13].

Zipf apresentou um modelo estruturado a partir da dinâmica — economia, em seus termos — entre o orador/escritor e o ouvinte/leitor. Essa economia seria regida pela disputa entre uma força de unificação e uma força de diversificação. A força de unificação seria guiada pelo orador e buscaria reduzir o tamanho do vocabulário a uma única palavra, unificando todos os significados. Por outro lado, a força de diversificação, guiada pelo ouvinte, tenderia a aumentar o vocabulário até alcançar a relação um para um entre palavras e significados. A força de unificação atuaria no sentido de diminuir o número de palavras diferentes para 1, aumentando assim o número de ocorrências dessa palavra e a força de diversificação agiria no sentido oposto tendendo a aumentar o número de palavras diferentes, enquanto diminuiria o número de ocorrências médio para a 1. Portanto, número de palavras distintas e frequência seriam parâmetros quantitativos dessa dinâmica do vocabulário.

Na década passada, Ferrer i Cancho e Solé [18] apresentaram uma proposta capaz de reproduzir o processo de comunicação partindo do princípio do menor esforço sugerido por Zipf. Nesse modelo foi considerado um conjunto de n sinais (que fariam o papel das palavras), $S = \{s_1, s_2, ..., s_n\}$, e um conjunto de m objetos (que fariam o papel dos significados), $R = \{r_1, r_2, ..., r_m\}$. A interação entre sinais e objetos era modelada por uma matriz binária $A = \{a_{ij}\}$, onde $1 \leq i \leq n$ e $1 \leq j \leq m$, assim se a palavra s_i pudesse assumir o significado r_j então $a_{ij} = 1$, caso contrário $a_{ij} = 0$.

O trabalho de Ferrer i Cancho e Solé mostrou que a lei de Zipf é o resultado do arranjo não trivial de associações palavra-significado adotado para suprir as necessidades do ouvinte e falante.

Como exemplo de apliacação, utilizando o livro *Ulisses* de James Joyce e definindo *rank* (categoria) $r = 1$ para a palavra mais frequente, *rank* $r = 2$ para a segunda palavra mais frequente e assim sucessivamente, Zipf observou que multiplicando o *rank* de um verbete pelo seu número de ocorrências (k) obtinha-se um valor constante C. Assim,

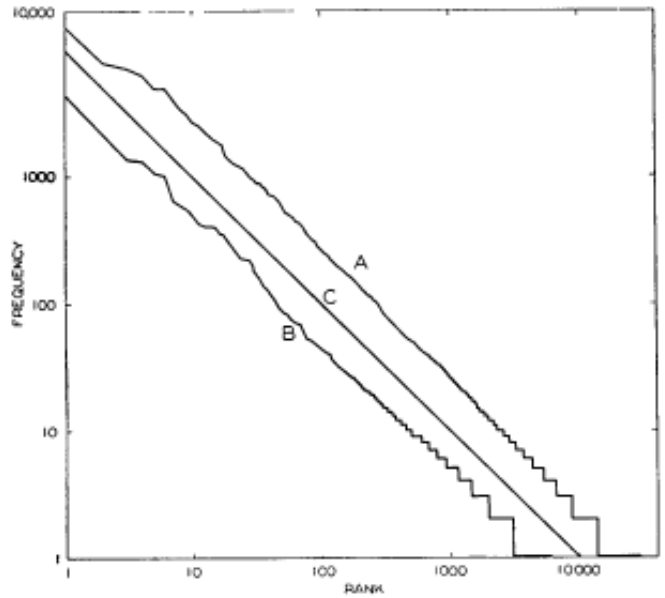


Figura 2.2 Gráfico da frequência como função do rank para verbetes extraídos do livro Ulisses (curva A), de uma coleção de textos de jornais norte-americanos (curva B) e a função r^{-1} (curva C). Figura extraída de *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* [13].

a relação posteriormente conhecida como Lei de Zipf, foi originalmente apresentada na seguinte forma:

$$r \times k = C. \tag{2.12}$$

Segundo Zipf, os dados apresentados em 1949 [13] apresentavam uma evidência clara da existência de um balanço (economia) de vocabulário. Nesse trabalho, Zipf apresentou a Figura 2.2 que exibe o gráfico onde os dados do livro *Ulisses* (curva A), de uma seleção de jornais (curva B) e a função $k = r^{-1}$ (curva C) são comparados. Observamos que exceto para pequenos valores de ocorrências as curvas seguem uma lei de potência de modo que podemos escrever de forma mais concisa:

$$k(r) \sim r^{-z}, \tag{2.13}$$

ou ainda, usando $f(r)$, como sendo a razão entre $k(r)$ e o número total de palavras do texto:

$$f(r) \sim r^{-z}. \tag{2.14}$$

Em 1936, no trabalho *The Psycho-biology of Language* [14], Zipf propôs que o número de palavras n que ocorrem exatamente k vezes é descrito por:

$$n(k) \sim k^{-\beta} \tag{2.15}$$

onde o expoente $\beta = 2$. No presente trabalho será utilizada esta última formulação da Lei de Zipf (2.15) que é denominada frequencista.

O *rank* de um verbete que ocorre k vezes em um texto pode ser entendido como o número de verbetes que ocorrem pelo menos k vezes. Podemos escrever:

$$r = \sum_{k'=k}^{\infty} n(k') \approx \int_{k}^{\infty} n(k')dk'. \tag{2.16}$$

Com as equações (2.13) e (2.15), temos:

$$k^{-\frac{1}{z}} \approx \int_{k}^{\infty} k'^{-\beta}dk'. \tag{2.17}$$

De onde:

$$z = \frac{1}{\beta - 1}. \tag{2.18}$$

Assim obtemos $z = 1$ quando $\beta = 2$ retomando os resultados descritos por Zipf na primeira metade do século XX.

2.3 DETEÇÃO DE PALAVRAS-CHAVE

Toda análise apresentada até aqui fez uso de uma abordagem frequencista dos textos literários, assim não é estranho que o primeiro trabalho sobre extração de palavras-chave tenha a frequência dos verbetes como cerne de sua proposta. Dez anos após Shannon publicar os trabalhos onde introduziu o conceito de informação, H. P. Luhn [32] propôs um método para criação automática de resumos de artigos.

Para determinar quais sentenças de um artigo deveriam constar no resumo criado automaticamente, Luhn propôs que seria necessário definir uma medida do conteúdo de informação para que todas as sentenças pudessem ser comparadas e classificadas. A frequência de ocorrência de verbetes foi primeiramente escolhida como medida útil para esse conteúdo de informação pois *"um escritor normalmente repete certas palavras à medida que avança ou varia seus argumentos e como ele elabora sobre um aspecto de um assunto."* [32].

Na proposta de Luhn, os verbetes mais frequentes (comuns) ou menos frequentes (raras) seriam excluídos e então os verbetes restantes seriam considerados como palavras-chave. Na Figura 2.3, original do trabalho de 1958, Luhn sugere a definição de pontos de corte C e D entre os quais estariam concentradas as palavras-chave.

Como será discutido nas seções seguintes, a proposta de Luhn seleciona como palavras-chave termos não importantes quanto ao conteúdo do texto utilizado para geração automática do resumo e, por outro lado, descarta inúmeras palavras de grande relevância para o texto avaliado.

Suponha que um texto tenha todas as suas palavras completamente embalharadas gerando uma cadeia de verbetes distribuídos aleatoriamente. Embora perdêssemos toda a correlação espacial manteriamos a frequência com que cada palavra aparece no texto em questão. Se utilizássemos a análise discutida até aqui, não haveria diferença entre os resultados obtidos a partir texto original ou com sua versão embaralhada. Assim sendo, a estrutura organizacional das palavras surge com um ponto de interesse no estudo da informação contida na linguagem escrita.

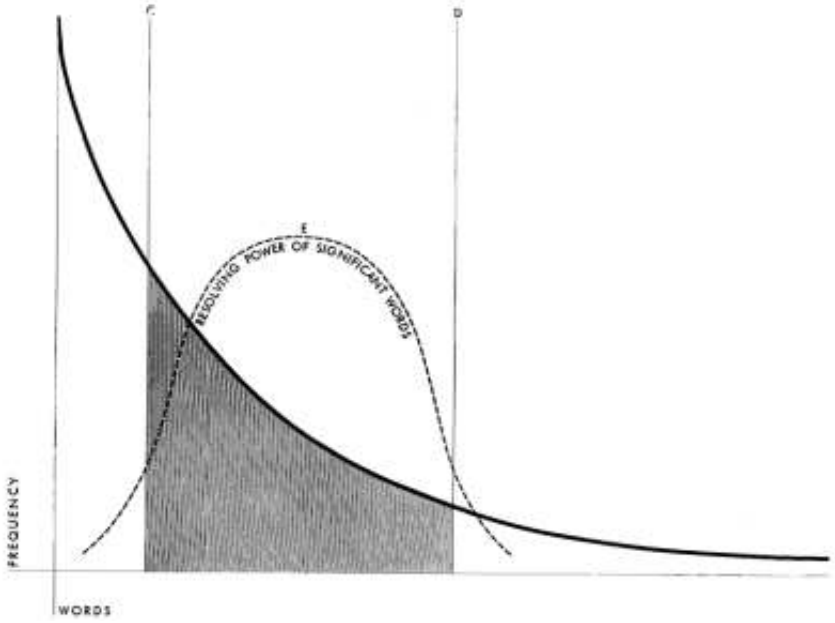


Figura 2.3 Diagrama onde a abscissa representa verbetes dispostos em ordem de frequência. A área hachurada destaca o conjunto de palavras relevantes segundo o método proposto por H. P. Luhn. Figura retirada de *The Automatic Creation of Literature Abstracts* [32].

No estudo da distribuição espacial de símbolos, inicialmente foram descobertas correlações de longo alcance em sequências de nucleotídeos [49] e em seguida foram apresentadas correlações de longo alcance em textos [50] e correlações de longo alcance entre as letras e sentenças [51].

W. Eberling e A. Neiman [51] investigaram as contribuições de símbolos, palavras e sentenças na estrutura dos textos. Três obras (*Bíblia, Contos dos Irmãos Grimm e Moby Dick*) foram comparadas com versões embaralhadas de três formas. Num primeiro nível foram embaralhados todos os símbolos do texto (26 letras do alfabeto inglês, o espaço em branco, a vírgula, o ponto final, os parantêses e o símbolo númerico #), no segundo nível foram embaralhadas todas as palavras e por fim foram embaralhadas todas as sentenças (definidas como conteúdo entre dois pontos finais). Segundo os autores, as correlações de longo alcance seriam perdidas ao embaralhar o texto no nível de frases e palavras indicando portanto que a organização geral da linguagem está mais relacionada com a distribuição e ordenação das palavras.

Em 2002, M. A. Montemurro e P. A. Pury [52] apresentaram um trabalho baseado na palavra como unidade fundamental da comunicação. Para analisar a estrutura fractal de registros escritos da linguagem humana, textos foram mapeados como séries temporais. Como resultado foi observado que além das correlações de curto alcance resultantes de regras sintáticas que atuam no nível da frase, as estruturas de longo alcance emergiam em grandes amostras da linguagem escrita dando origem a correlações de longo alcance no uso das palavras.

Os trabalhos realizados nos ínicio desse século mantiveram um problema em aberto:

como formular critérios eficientes para extração de verbetes relevantes de um texto a partir de suas propriedades estatísticas. Em 2002, Ortuño e colaboradores [33] propuseram que a informação espacial de uma palavra, ou seja, a forma como ela é distribuída ao longo do texto (independentemente da sua frequência relativa) seria uma boa forma de atacar um dos principais problemas da mineração de dados que é a extração de palavras-chave de textos para os quais não há informação disponível *a priori*.

A abordagem proposta por Ortuño consiste em determinar a distribuição $p(x)$ de distâncias entre sucessivas ocorrências de um verbete. Para isso, todas as palavras do texto de interesse devem ser numeradas em ordem de aparição para, então, extrair as posições correspondentes a um determinado verbete. Tendo obtido o conjunto de distâncias entre sucessivas ocorrências de um verbete $\{x_i\}$, definimos $p(x)$ como a frequência relativa de ocorrência de uma determinada separação x, e $P_1(x) = \sum_{x'=1}^{x} p(x')$ como a sua distribuição aculumada. Esta abordagem é semelhante ao método adotado no estudo das estatísticas de níveis do espectro de sistemas quânticos desordenados, segundo a teoria de matrizes aleatórias [53]. Para eliminar a dependência com frequência, é conveniente normalizar as separações para cada verbete. Para tal devemos montar o conjunto $\{x_i\}$ em unidades da distância média ($\bar{x}$), definindo assim:

$$s = \frac{x}{\bar{x}}. \tag{2.19}$$

Se as palavras estiverem distribuídas de forma aleatória, a distribuição acumulada para cada verbete no limite de altíssimas frequências segue uma distribuição de Poisson:

$$P_1(s) = 1 - \exp(-s). \tag{2.20}$$

Na Figura 2.4 é apresentada a distribuição acumulada $P_1(s)$ para quatro verbetes diferentes da versão inglesa de *Dom Quixote*: "Quijote", "Sancho", "the"e "and". Essa escolha foi feita pois as dois primeiros verbetes são relevantes para o texto considerado, ao contrário dos dois últimos termos. A linha contínua corresponde à distribuição de Poisson. É possível observar que as curvas dos verbetes não-relevantes se aproximam da curva de Poisson, o que pode ser entendido como se tais verbetes fossem distribuídos aleatoriamente ao longo do texto. O comportamento das outras dois verbetes pode ser entendido se imaginarmos que as palavras relevantes normalmente aparecerão em um contexto muito específico, concentradas em algumas regiões do texto.

Devido o cálculo das distribuições acumuladas de todos os verbetes de um texto demandar muito tempo computacional, Ortuño propôs que o desvio padrão:

$$\sigma = \sqrt{\bar{s^2} - \bar{s}^2} \tag{2.21}$$

seria um estimador que carregaria a mesma informação que $P_1(s)$ com menor custo computacional. Esse estimador se mostra eficaz pois o desvio padrão cresce rapidamente conforme aumenta a heterogeneidade da distribuição dos espaçamentos. Sendo para uma distribuição de Poisson $\sigma = 1$. Seguindo a analogia com estatística de níveis de energia, todas as ocorrências de um verbete podem ser consideradas como um "nível de energia"ϵ_i dentro de um "espectro de energia", formado por todas as ocorrências do verbete. O valor do nível de energia ϵ_i é dado simplesmente pela posição do verbete no texto.

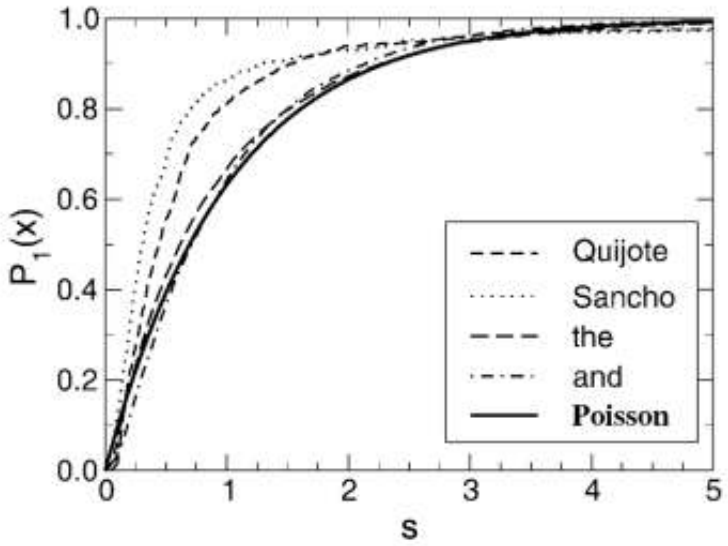


Figura 2.4 Gráfico da distribuição acumulada P_1 para quatro verbetes de *The Quijote* (versão inglesa) como uma função da separação normalizada s. A linha contínua corresponde à distribuição de Poisson. Figura extraída de Ortuño et al. [33].

Na Figura 2.5 apresentamos um típico diagrama de níveis para dois verbetes do livro, escrito por Machado de Assis, *Quincas Borba* ($T = 57756$). Embora os verbetes *"Borba"* e *"este"* tenham número de ocorrências quase idênticos, $k_{BORBA} = 97$ e $k_{ESTE} = 92$, tais verbetes apresentam diferentes graus de relevância. Tal diferenciação é bem quantificada pelo desvio padrão, temos $\sigma_{BORBA} = 3,02$ e $\sigma_{ESTE} = 0,82$. Uma analogia didática é imaginar que verbetes relevantes se atraem e enquanto verbetes não relevantes se repelem [34]. O fenômeno da agloremaração é bem quantificado pelo desvio padrão, para os casos de atração $\sigma > 1$ e para os casos de repulsão $\sigma < 1$.

A Tabela 2.1 apresenta os verbetes com os dez maiores valores do parâmetro σ para o livro Quincas Borba de Machado de Assis (detalhes sobre o *corpus* utilizado nesta dissertação estão disponíveis no Apêndice A). Destaca-se que os nove verbetes que apresentam o maior valor para o desvio padrão sejam de fato relevantes no livro em análise. Na Figura 2.6 é apresentada a relação entre os valores do desvio padrão e o número de ocorrências para a obra supracitada.

Estabelecido o desvio padrão como estimador para extração de palavras chave, Ortuño e colaboradores estendeu a abordagem para sequências de DNA para identificar "palavras relevantes" compostas por sequências de nucleotídeos [33]. O estudo do intervalo de recorrência entre verbetes também foi utilizado para classificação textual [54], identificação autoral [54, 55] e no estudo de partes funcionais do genoma [56].

Hongding Zhou e Gary W. Slater [35] apontaram que a proposta de Ortuño possui algumas limitações como, por exemplo, a dificuldade para identificar verbetes com baixa frequência como sendo releventes. Ainda nesse regime, esse estimador não seria estável no sentido de que pode ser fortemente afetada pela mudança de uma única posição de

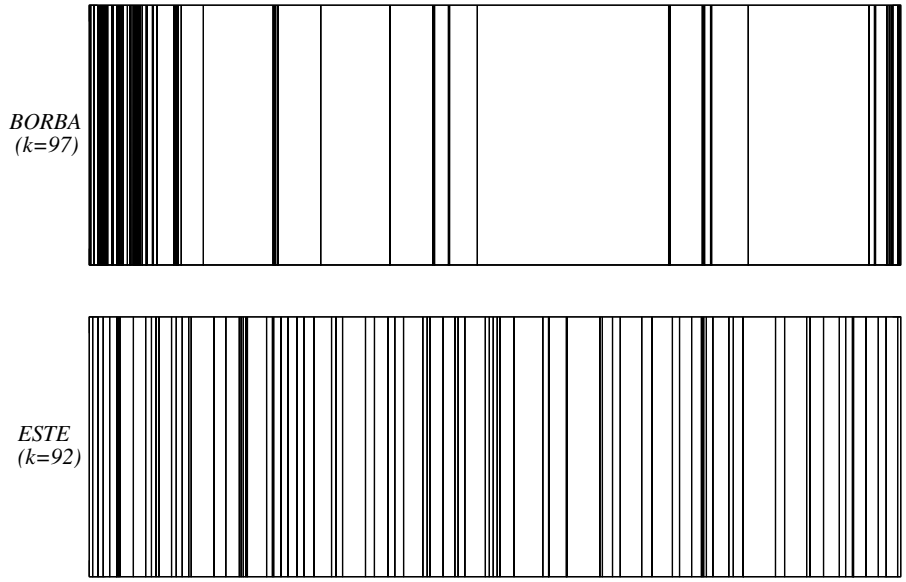


Figura 2.5 Espectro das posições absolutas para os verbetes BORBA ($k = 97$; $\sigma = 3,02$) e ESTE ($k = 92$; $\sigma = 0,82$). A distribuicão é computada a partir início do livro Quincas Borba ($T = 57756$). Para construir a imagem, definimos uma linha vertical fina (de altura arbitrária) na posição de cada ocorrência do verbete.

	Verbete	Frequência	σ
1	batatas	19	3,83
2	Fernanda	90	3,68
3	Maria	225	3,63
4	Comadre	15	3,51
5	Freitas	32	3,10
6	senhoria	14	3,09
7	Marquês	18	3,04
8	Borba	97	3,02
9	Quincas	96	3,00
10	tu	41	2,89

Tabela 2.1 Verbetes, da obra Quincas Borba, com os maiores valores de desvio padrão (σ).

ocorrência. Como forma de melhorar o estimador exposta anteriormente, esses autores propuseram incluir a informação que está contida nas extremidades dos textos analisados, ou seja, incluir a primeira e a última palavra que compõe o texto na contagem das distâncias entre as sucessivas ocorrências de uma palavra. O principal argumento seria que as palavras comuns (baixa relevância) sendo distribuídas aleatoriamente, devem ser encontradas perto das duas extremidades do texto. Enquanto as palavras relevantes são muito improváveis de serem encontradas perto de ambas as extremidades.

Ainda baseado no trabalho de Ortuño e colaboradores, P. Carpena et. al [34] pro-

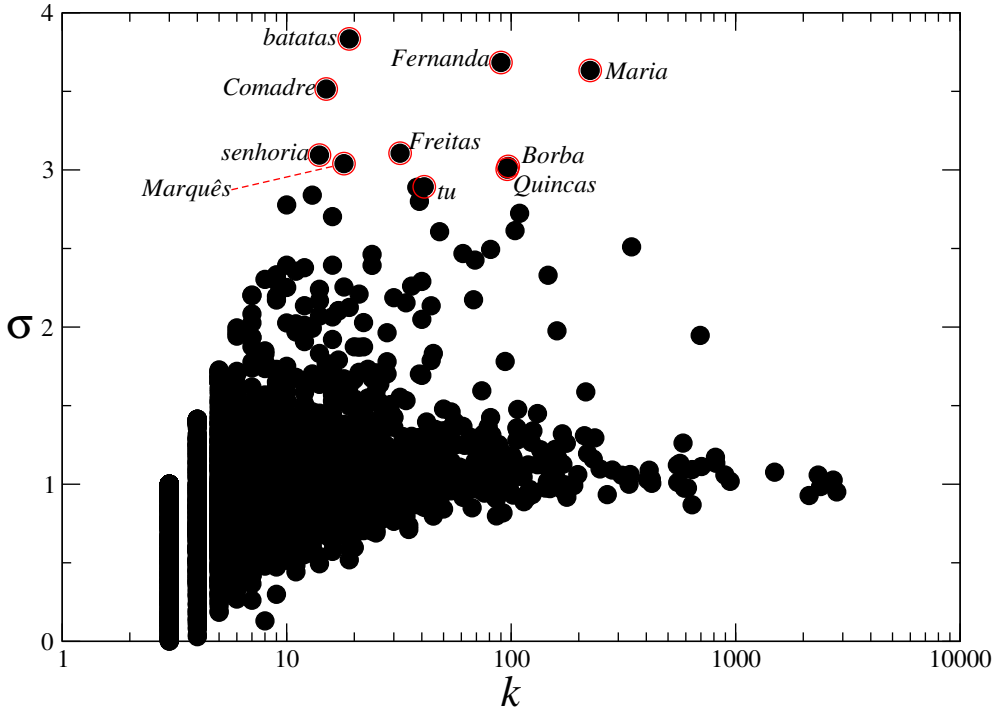


Figura 2.6 Gráfico da relação entre o desvio padrão σ e número de ocorrências k para todos os verbetes do livro Quincas Borba. Os círculos vermelhos destacam os dez verbetes associados aos maiores valores de σ.

puseram uma abordagem que combinava a frequência de cada verbete com a informação fornecida pelo agrupamento da palavra (estrutura espacial ao longo do texto). Os autores apontaram que o comportamento poissoniano seria visto somente para uma distribuição contínua de distâncias, como por exemplo no caso dos níveis de energia, mas não para as palavras, onde as distâncias são números inteiros. O equivalente discreto da distribuição de Poisson é a distribuição geométrica:

$$P_{geo}(d) = p(1-p)^{d-1}, \tag{2.22}$$

onde $p = k/T$ é a probabilidade de encontrar tal verbete no texto, k é a frequência de ocorrências desse verbete, d é a distância entre sucessivas ocorrências de um dado verbete e T é o número total de palavras do texto.

P. Carpena e colaboradores sugeriram que a distribuição geométrica seria então um bom modelo para descrever o comportamento de palavras irrelevantes. Como para o caso geométrico

$$\sigma_{geo} = \sqrt{1-p}, \tag{2.23}$$

os autores sugeriram definir a seguinte medida de agrupamento:

$$\sigma_{nor} = \frac{\sigma}{\sigma_{geo}} = \frac{\sigma}{\sqrt{1-p}}. \tag{2.24}$$

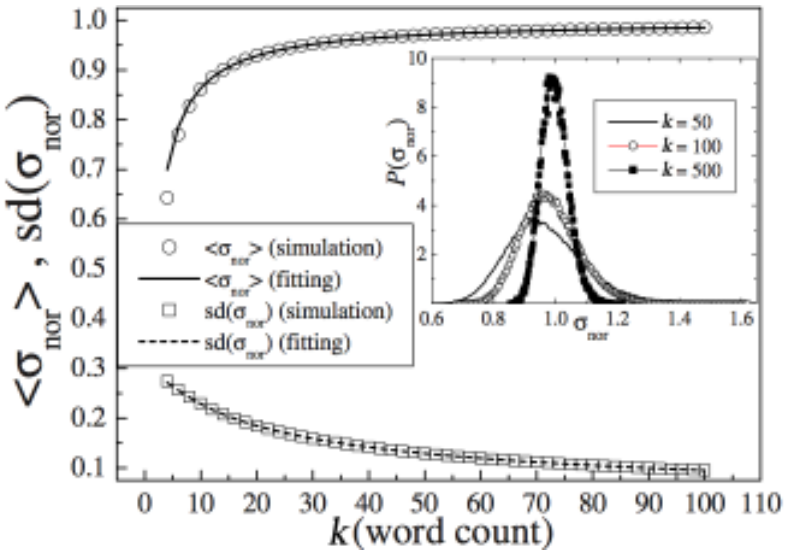


Figura 2.7 Gráfico do valor médio $\langle\sigma_{nor}\rangle$ e desvio padrão $sd(\sigma_{nor})$ da distribuição $P(\sigma_{nor})$ como função de frequência k obtidos por simulação de textos aleatórios. Subgráfico: $P(\sigma_{nor})$ para três valores de k. Figura adaptada de P. Carpena et. al [34].

Para os verbetes, a forte dependência com a frequência k de σ_{nor} é evidenciada na Figura 2.7, bem como a dependência da distribuição $P(\sigma_{nor})$. Para pequenos valores de k, a distribuição $P(\sigma_{nor})$ é larga, enquanto a medida que a frequência cresce a distribuição assume a forma de uma gaussiana estreita centrada em $\sigma_{nor} = 1$ como exposto no subgráfico.

Simulando textos aleatórios como sequências binárias aleatórias (0000100011100001...), onde o "1" aparece com probabilidade p e representa um certo verbete do texto e o símbolo "0" assume o papel dos demais verbetes com probabilidade $1 - p$, P. Carpena e colaboradores obtiveram computacionalmente as seguintes relações entre a média e o desvio padrão de σ_{nor} e a frequência:

$$\langle\sigma_{nor}\rangle = \frac{2k-1}{2k+2}, \tag{2.25}$$

$$sd(\sigma_{nor}) = \frac{1}{\sqrt{k}(1+2.8k^{-0.865})}. \tag{2.26}$$

2.4 ANÁLISES DE ENTROPIA

Tendo por base o conceito de entropia de informação introduzido por Claude Shannon, M. A. Montemurro e D. H. Zanette [57] propuseram uma abordagem para estudar a distribuição de verbetes de acordo com sua função linguística. Dividindo, em P partes, o texto a ser analisado a partir das partições naturais como sentenças, parágrafos, seções ou capítulos, podemos definir N_i como sendo o número total de palavras contidas na i-ésima partição e $k_i(w)$ como o número de ocorrências do verbete w nessa partição. Então a

fração

$$f_i(w) = \frac{k_i(w)}{N_i} \quad (i = 1, 2, ..., P) \tag{2.27}$$

é a frequência relativa do verbete w na partição i. Portanto é possível definir a seguinte medida de probabilidade sobre todas as partições:

$$p_i(w) = \frac{f_i(w)}{\sum_{j=1}^{P} f_j(w)}. \tag{2.28}$$

A entropia associada a distribuição $p_i(w)$ será:

$$S(w) = -\frac{1}{\ln P} \sum_{i=1}^{P} p_i(w) \ln(p_i(w)). \tag{2.29}$$

onde $k = 1/\ln P$ foi escolhido para que o valor máximo de S seja igual a 1. Assim, para uma palavra uniformente distribuída ($p_i = 1/P$) obtemos $S = 1$, por outro lado para um verbete utilizado em uma partição específica ($p_j = 1$ e $p_i = 0$ para $i \neq j$) temos $S = 0$. Dessa forma, verbetes com uso gramatical frequente como preposições, advérbios, adjetivos, conjunções e pronomes apresentariam altos valores de entropia enquanto palavras-chave apresentariam baixos valores de entropia.

Essas mesmas ideias foram utilizadas por J. P. Herrera e P. A. Pury [17] para detetar a relevância de palavras. Para tal foi proposta uma medida de entropia, E_{nor}, que foi normalizada para livrar-se da dependência da frequência:

$$E_{nor}(w) = k(w)\frac{2 \ln P}{P - 1}(1 - S(w)). \tag{2.30}$$

Na Figura 2.8 são apresentados os valores de Entropia normalizada para a obra *"On the Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life"*, com $P = 16$ que é o número de capítulos do livro de Charles Darwin. Em destaque, os círculos vermelhos apontam os verbetes que constituem o glossário da obra.

Um estimador alternativo foi proposto por Ali Mehri e Amir H. Darooneh [37] e consistiria na extração de palavras-chave utilizando medidas não-extensivas. Podemos compreender a mecânica estatística não-extensiva como uma generalização da mecânica estatística de Boltzmann-Gibbs. Na formulação em questão, a entropia de Tsallis [58] é dada por:

$$S_q = -k \sum_{i=1}^{N} p_i^q \ln_q p_i, \tag{2.31}$$

onde p_i é a probabilidade de ocorrência do i-ésimo microestado, N é o número total de microestados, q é um parâmetro que caracteriza a não-extensividade do sistema e $\ln_q$ é uma generalização da função logaritmo:

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q}, \tag{2.32}$$

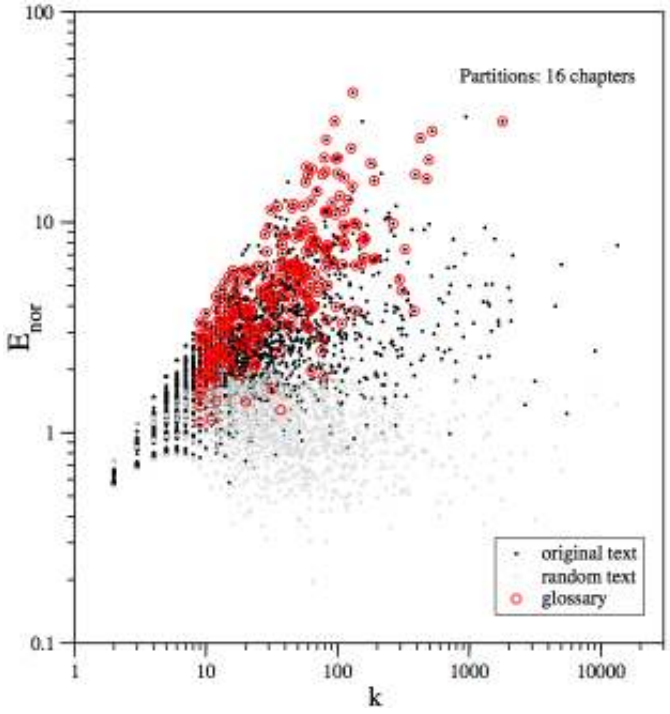


Figura 2.8 Gráfico da entropia normalizada E_{nor} como função do número de ocorrências k para cada verbete no livro *"On the Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life"*. Os pontos cinzas correspondem à versão aleatória do texto e os círculos abertos vermelhos às palavras correspondentes ao glossário da obra. Figura adaptada de Herrera e colaboradores [17].

com $q = 1$ correspondendo ao caso extensivo (Boltzmann-Gibbs).

Segundo o método proposto, primeiramente é computada a distribuição de distâncias para cada verbete. Para tal é proposto que seja utilizada condição de contorno periódica. Em seguida, a distribuição acumulada é calculada e então são obtidas regressões q-exponenciais. A. Mehri e A. H. Darooneh apontam que o valor de q seria então um bom estimador para extração de palavras-chave. Sendo os valores altos de q associados às palavras mais relevantes da obra analisada. Os autores também propõe usar tal método para a construção de glossários bem como sumarização e classificação de documentos.

Inspirados por um sistema bosônico unidimensional, A. Mehri e A. H. Darooneh [59] propuseram considerar um texto com comprimento T como um sistema que é constituído por P seções para o qual são apresentadas vários estimadores entrópicos com condições de contorno periódicas. Computando a distribuição de distâncias entre sucessivas ocorrências de um verbete num texto, podemos definir $p(d) = M_d/M$ onde M_d é o número de distâncias de tamanho d. Assim, os autores definem uma entropia:

$$E_B = -\sum_{d=1}^{d_{max}} p(d) \ln p(d) \tag{2.33}$$

onde d_{max} é a maior distância entre sucessivas ocorrências de um dado verbete.

Mais recentemente, C. Carretero-Campos e colaboradores [38], visando aumentar a acurácia na deteção de palavras-chave em textos pequenos, propuseram melhorias nos estimadores apresentadas até aqui. Enquanto a busca por novos estimadores e por melhorias nos métodos descritos nesse capítulo persiste [36], diversas aplicações têm sido apresentadas. Cabendo o destaque para a análise feita por Marcelo A. Montemurro e Damián H. Zanette [60] do manuscrito *Voynich*, um pergaminho medieval escrito numa língua desconhecida que tem-se mantido até agora como um mistério para linguistas e criptologistas.

Os métodos discutidos nesse Capítulo fornecem um conjunto de abordagens que podem ser utilizadas no estudo de aspectos mais fundamentais da distribuição de frequências e da ordenação espacial de palavras em linguagem escrita. No Capítulo 3, procedemos uma investigação do caráter frequencista dos verbetes e dos seus comprimentos e sua possível implicação na categorização de idiomas.

CAPÍTULO 3

ESTATÍSTICA DE FREQUÊNCIA EM LINGUAGEM ESCRITA

As palavras compridas não são palavras difíceis.
Difíceis são as palavras curtas. Há muito mais sutileza metafísica na palavra "dane-se!" do que na palavra "degeneração".
—G. K. CHESTERTON (Ortodoxia)

A contagem de frequências é a maneira mais elementar de estudar propriedades estatísticas de linguagem escrita. Nesse capítulo, partindo das ideias de George K. Zipf, analisamos inicialmente a relação entre o número de ocorrências e a frequência de um verbete, apresentando o coeficiente de Zipf para todas as dez línguas que compõem o *corpus* utilizado. Em seguida, discutimos a dependência do número de verbetes e da frequência máxima de ocorrência com relação ao número total de palavras de um texto. Tal análise tem como principal objetivo validar o conjunto de textos adotados nesta dissertação como um *corpus* que apresenta propriedades estatísticas em concordância com os parâmetros descritivos de amostras de linguagem escrita. Por fim, analisamos a distribuição de probabilidades dos tamanhos dos verbetes bem como a entropia de Shannon associada a essa distribuição e discutimos de que modo essa entropia pode ser utilizada para caracterização de grupos linguísticos.

3.1 CARACTERÍSTICAS GERAIS

Conforme exposto no Capítulo 2, George K. Zipf [14] propôs que o número n de verbetes que ocorrem exatamente k vezes num texto é descrito por uma lei de potência:

$$n(k) \sim k^{-\beta}, \tag{3.1}$$

sendo o expoente $\beta = 2$. A Lei de Zipf aponta a ausência de uma escala de frequência característica associada as ocorrências do verbetes. Do ponto de vista matemático, essa lei pode ser formalizada utilizando funções harmônicas [61].

A Lei de Zipf é uma característica bem estabelecida da linguagem escrita [62]. Sendo, portanto, esse estudo ideal para uma caracterização inicial do *corpus*. Para isso nesta dissertação foram computadas as frequências dos verbetes de todos os textos literários (Apêndice A). O conjunto de artigos retirados da *Wikipedia* não foi utilizado nessa análise devido ao pequeno número de verbetes que constituem cada texto tornando-os pouco representativos do ponto de vista estatístico. Para o conjunto de textos selecionados, foi

calculado o número de verbetes com frequência entre k' e a frequência máxima k_{max}:

$$n(k \geq k') \sim \sum_{k'}^{k_{max}} k^{-\beta}. \tag{3.2}$$

Portanto a distribuição acumulada segue uma lei de potência na forma:

$$n(k \geq k') \sim k^{-\beta+1}. \tag{3.3}$$

Na Figura 3.1 são apresentados, em escala duplo-logarítmica, os dados extraídos dos maiores textos de cada língua, a linha tracejada corresponde a uma lei de potência cujo expoente é $\beta = 2$. Para cada idioma foi calculado o expoente médio da regressão realizada a partir das frequências dos verbetes que compõem os vinte e cinco textos selecionados. Os valores médios e seus respectivos desvios são apresentados na Tabela 3.1. Para o conjunto de todas as dez línguas foi obtido $\beta = 1.95 \pm 0.08$ como sendo o valor médio do expoente. Portanto, o resultado extraído dos textos literários demonstrou estar em conformidade com outros valores apresentados na literatura para outros textos [61, 63].

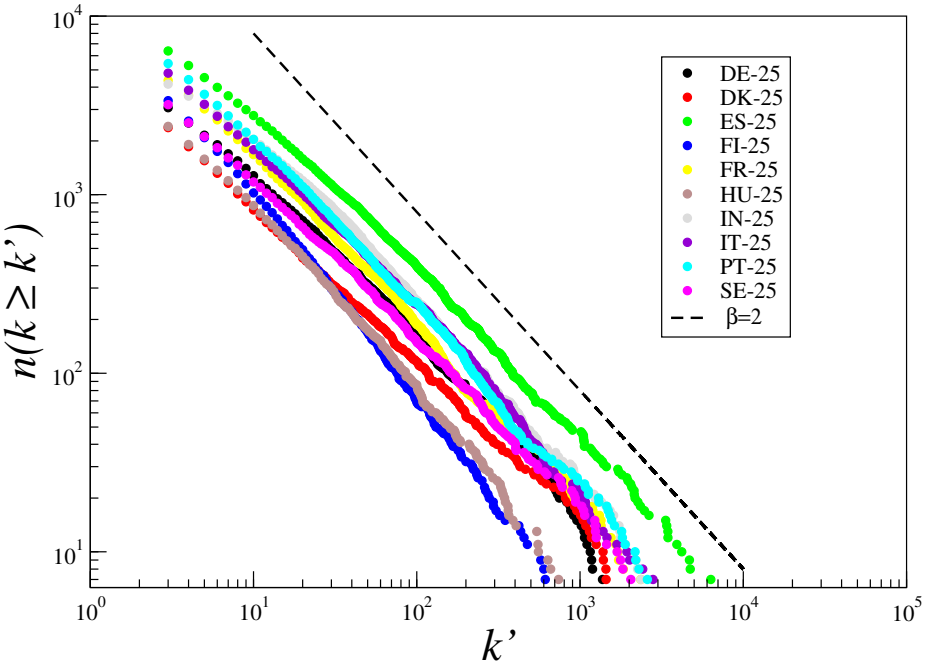


Figura 3.1 Gráfico, em escala duplo-logarítmica, do número de verbetes n com frequência k maior que k' para os maiores textos de cada uma das dez línguas que compõe o *corpus*. A linha tracejada corresponde a uma lei de potência cujo expoente é $\beta = 2$.

Seguindo o estudo das frequências, analisamos o número de verbetes, ou vocabulário, V em função do número total de palavras T de um texto. Para obtermos um espectro maior de tamanhos, acrecentamos ao *corpus* utilizado anteriormente uma seleção, por língua, de vinte e cinco artigos retirados da *Wikipedia* (Apêndice B).

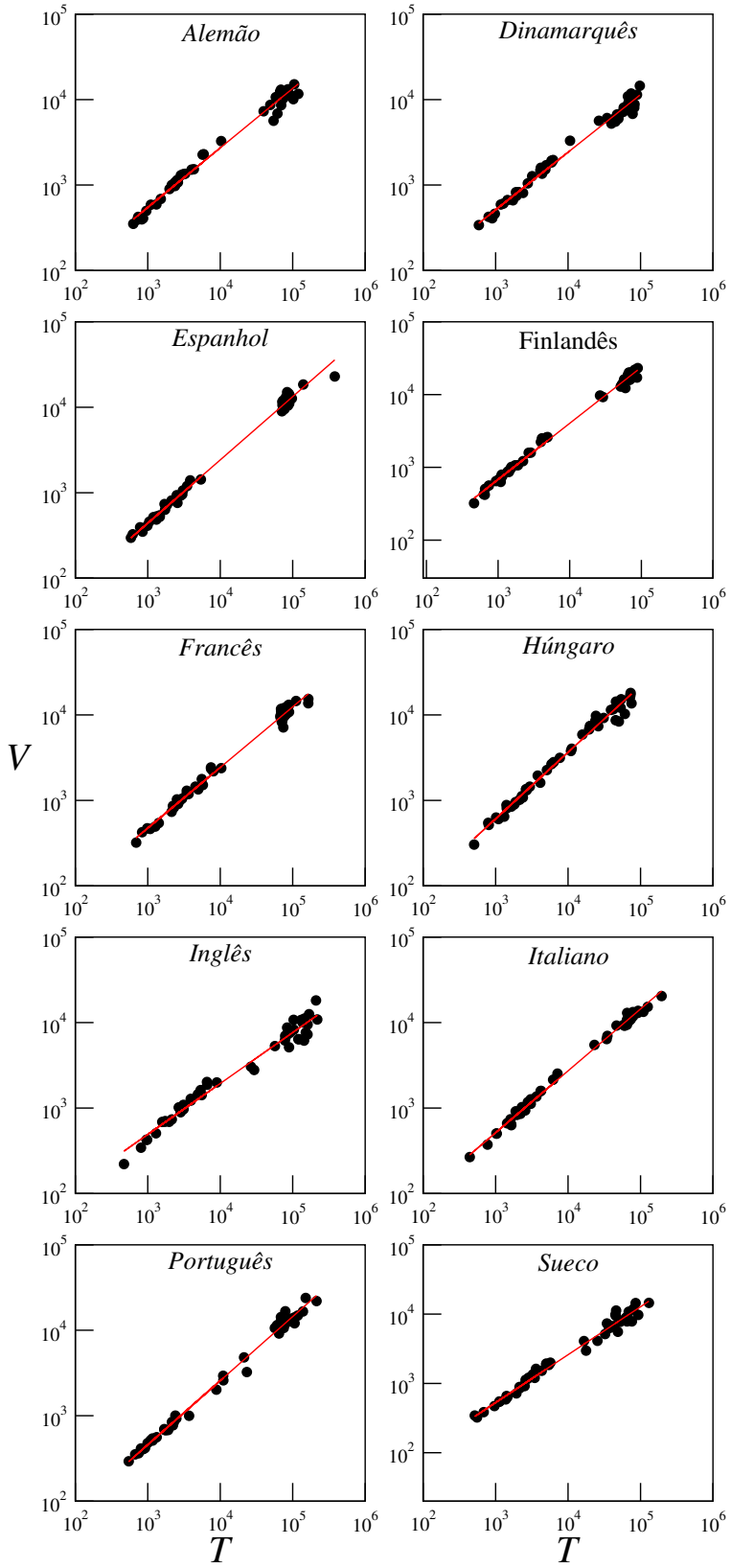


Figura 3.2 Gráfico em escala duplo-logarítmica da relação entre o número de verbetes (vocabulário) V e o número total de palavras T para o *corpus* de textos literários e artigos da *Wikipedia*. As linhas vermelhas apresentam as regressões realizadas a partir dos dados extraídos do *corpus*.

Para todos os idiomas expostos na Figura 3.2, observamos que é possível descrever a relação entre o vocabulário V e o tamanho T do texto segundo uma lei de potência:

$$V(T) \sim T^{\lambda}. \tag{3·4}$$

Essa abordagem foi proposta em 1958 por G. Herdan [64] e foi sistematizada por H. S. Heaps [19]. Posteriormente essa relação foi denominada Lei de Heaps, sendo o expoente de Heaps λ um valor compreendido entre 0 e 1 [65]. Em 2013, Martin Gerlach e Eduardo G. Altmann [22] mostraram que, num *corpus* de mais de 5,2 milhões de livros publicados nos últimos cincos séculos e digitalizados pelo Google, a Lei de Heaps é melhor descrita por uma lei de potência com dois regimes. O espectro de tamanhos utilizados nessa dissertação ($10^2 < T < 10^6$) corresponde ao primeiro desses dois regimes. Ainda em 2013, Tomasso Pola e colaboradores [66], utilizando um único texto, observaram o mesmo comportamento descrito por Martin Gerlach e Eduardo G. Altmann [22].

Em 2001, Alexander F. Gelbukh e Grigori Sidorov [67] propuseram, a partir de textos em inglês, russo e espanhol, que diferentes línguas apresentariam diferentes expoentes de Heaps. Resultado contrário foi apresentado em 2013 por Tomasso Pola [66] para francês, italiano, inglês e alemão.

Ao analisarmos os valores dos expoentes das regressões (Tabela 3.1), obtemos o valor médio $\lambda = 0.71 \pm 0.05$ para o expoente de Heaps. Tal fato assegura que o corpus utilizado contém um conjunto representativo adequado aos estudo das propriedades estatísticas de linguagem escrita [67, 68].

Quando ordenamos de forma crescente os dez idiomas que compõe o *corpus* segundo os expoentes de Heaps apresentados na Tabela 3.1 obtemos a seguinte sequência: inglês, dinamarquês, sueco, alemão, francês, italiano, espanhol, português, húngaro e finlandês. Observamos que tal sequência ordena os idiomas também segundo os grupos linguísticos: família germânica, família latina e família urálica. Esse resultado inédito amplia a hipótese de Alexander F. Gelbukh e Grigori Sidorov [67] e aponta que o expoente de Heaps pode ser útil para uma possível classificacão de grupos linguísticos.

Outra propriedade associada à distribuição de frequências foi observada em 1949 por George Zipf [13]. O autor propôs que a frequência k_{max} do verbete mais comum em um texto seria aproximadamente dez porcento do número total de palavras. Mais recentemente, Sebastian Bernhardsson e colaboradores [68] propuseram que a frequência do verbete mais comum escala linearmente com o tamanho T do texto medido em número de palavras:

$$k_{max} = \epsilon T. \tag{3·5}$$

No entanto, os autores não discutiram qual o valor do coeficiente ϵ, limitando-se a afirmar que esse valor seria constante e maior que zero.

A partir do *corpus* de artigos da *Wikipedia* e textos literários extraímos a frequência do verbete mais comum. Ao dispormos os dados num gráfico duplo-logarítmico observamos que o coeficiente ν da regressão é menor que 1 (Tabela 3.1) caracterizando assim uma relação não linear. Na Figura 3.3 é possível notar que no regime de pequenos tamanhos de texto há maior flutuação em torno na reta cujo expoente é o valor médio $\nu = 0.92$. Quando a regressão é realizada apenas para textos literários ($T > 10^4$) o valor médio obtido para o

expoente é 1 caracterizando assim uma dependência linear entre a frequência máxima e o tamanho do texto. Tomando então apenas o *corpus* de textos literários, foram realizadas regressões lineares cujos coeficientes angulares ϵ são apresentados na Tabela 3.1. Dentre os idiomas analisados, as amostras de alemão e dinamarquês foram as que apresentaram maior erro relativo ($\Delta\epsilon$=0.16). A média global desses coeficientes angulares assumiu o valor $\epsilon = 0.048 \pm 0.015$. É importante frisar que, ao contrário do que afirmava Zipf, esse valor não é 0.1 e varia conforme a língua analisada.

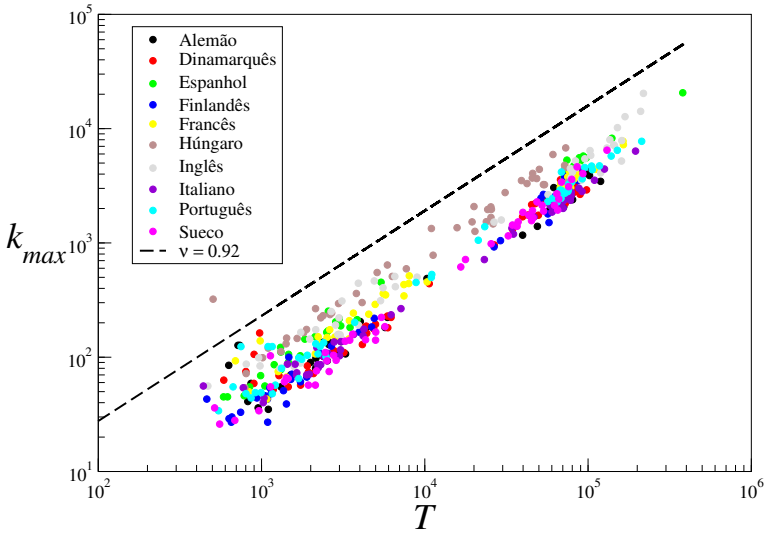


Figura 3.3 Gráfico da frequência k_{max} do verbete mais comum como função do tamanho T do texto medido em número de palavras para artigos da *Wikipedia* e textos literários (pontos coloridos). A reta tracejada corresponde a curva $k_{max} \sim T^{0.92}$.

Embora o comportamento qualitativo exposto na Figura 3.3 seja discutido na literatura [68, 69], não são conhecidos resultados quantitativos. Assim, além de caracterizar o *corpus* segundo às leis de escala previamente conhecidas (Zipf e Heaps) apresentamos resultados inéditos para a relação entre o expoente de Heaps de um idioma e seu respectivo grupo linguístico e para a relação entre a frequência máxima de um verbete e o tamanho do texto para diferentes línguas.

3.2 COMPRIMENTO DE VERBETES: ANÁLISES DE FREQUÊNCIA E ENTROPIA

Em agosto de 1851, Augustus de Morgan escreveu uma carta a um amigo na qual afirmava que o tamanho médio dos verbetes utilizados num texto poderia ser utilizado como parâmetro para identificação autoral [70]. Em 1887, Thomas Corwin Mendenhall propôs estudar a representação gráfica da relação entre o comprimento de um verbete e a sua frequência buscando identificar a "curva normal do escritor" [71]. Em 1935, George Zipf escreveu: *"Em vista dos elementos do fluxo de discurso pode-se dizer que o comprimento de uma palavra tende a manter uma relação inversa com a sua frequência*

	Idioma	β	λ	ν	ϵ
1	Alemão	1.92 ± 0.07	0.70 ± 0.01	0.89 ± 0.03	0.036 ± 0.006
2	Dinamarquês	1.88 ± 0.05	0.67 ± 0.01	0.91 ± 0.03	0.031 ± 0.005
3	Espanhol	1.96 ± 0.03	0.74 ± 0.01	0.95 ± 0.01	0.054 ± 0.002
4	Finlandês	2.09 ± 0.05	0.77 ± 0.01	0.98 ± 0.02	0.052 ± 0.007
5	Francês	1.93 ± 0.03	0.71 ± 0.01	0.91 ± 0.02	0.045 ± 0.003
6	Húngaro	2.08 ± 0.08	0.77 ± 0.01	0.88 ± 0.03	0.068 ± 0.008
7	Inglês	1.84 ± 0.06	0.60 ± 0.01	0.91 ± 0.02	0.075 ± 0.009
8	Italiano	1.94 ± 0.06	0.72 ± 0.01	0.91 ± 0.01	0.033 ± 0.002
9	Português	1.98 ± 0.06	0.75 ± 0.01	0.91 ± 0.02	0.038 ± 0.002
10	Sueco	1.92 ± 0.06	0.69 ± 0.01	0.99 ± 0.02	0.050 ± 0.003
	Média	1.95 ± 0.08	0.71 ± 0.05	0.92 ± 0.03	0.048 ± 0.015

Tabela 3.1 Expoente β médio por língua da Lei de Zipf (equação 3.1). Expoente de Heaps λ por idioma (equação 3.4). Expoente ν da regressão $k_{max} \sim T^{\nu}$. Coeficiente angular ϵ da frequência do verbete mais comum como função do tamanho T para textos literários (equação 3.5).

relativa.[...] Nota de rodapé: Não necessariamente proporcional; possivelmente alguma função matemática não-linear." [14]. Desde então o estudo das distribuições dos comprimentos dos verbetes de um texto é um dos mais vastos campos de pesquisa da linguística quantitativa [72].

Para todas as amostras de linguagem escrita do nosso *corpus* foram computadas as distribuições dos tamanhos do verbetes:

$$P(l) = \frac{N(l)}{\sum_{l=1}^{l_{max}} N(l)} = \frac{N(l)}{V} \tag{3.6}$$

onde $N(l)$ é o número de verbetes de tamanho l e V é o número total de verbetes.

Na Figura 3.4 são apresentados as distribuições para textos literários de cada uma das dez línguas. Esse conjunto selecionado cobre todo o espectro de tamanhos do *corpus*. Como podemos ver, para obras literárias a forma da distribuição $P(l)$ não depende do tamanho do texto e segue uma curva característica para cada língua.

A entropia de Shannon $H(l)$ associada a um dado tamanho l é calculada a partir da distribuição de palavras do texto. Na Figura 3.5, observamos que a forma qualitativa das curvas é semelhante para idiomas pertencentes à mesma família linguística. A família latina (espanhol, francês, italiano e português) apresenta dois picos característicos. Na família germânica (alemão, dinamarquês, inglês e sueco) o segundo pico é drasticamente reduzido. Enquanto a família urálica possui apenas um pico localizado em $l = 5$. Em Janeiro de 2014, Maria Kalimeri e colaboradores [73], apresentaram resultados para a distribuição de entropias para digramas e trigramas como função dos tamanhos dos verbetes.

Essa investigação está diretamente relacionada à busca pelas conexões entre o tamanho dos verbetes e o seu conteúdo de informação. Em 2010, Steven T. Piantadosi e colaboradores [20] propuseram que o tamanho dos verbetes seria otimizado para um aumento na eficiência da comunicação. Em sua metodologia, esses autores utilizam a probabilidade de ocorrência de um dado verbete em contextos específicos. Para construção dos contextos foi utilizado o modelo *n-gram* que pode ser compreendido como a análise dos n primeiros vizinhos do verbete de interesse.

Os resultados apresentados por Piantadosi et al. [20] apontaram uma estreita relação entre a distribuição espacial de um verbete e o seu conteúdo informacional, o que sugere uma limitação para análise frequencista da linguagem escrita.

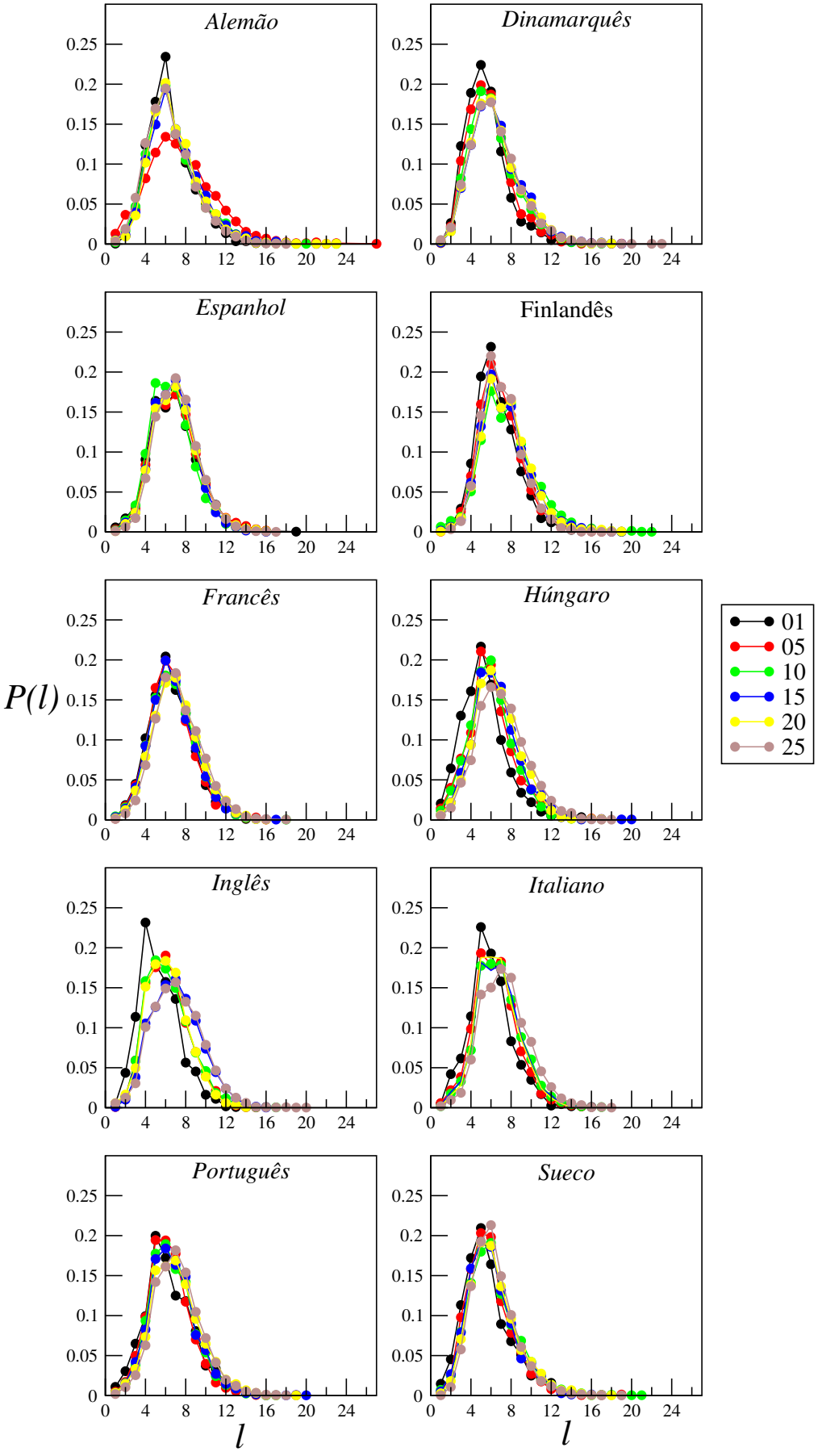


Figura 3.4 Distribuição $P(l)$ dos tamanhos dos verbetes para os textos 01, 05, 10, 15, 20 e 25 de cada língua do *corpus* de textos literários (Apêndice A).

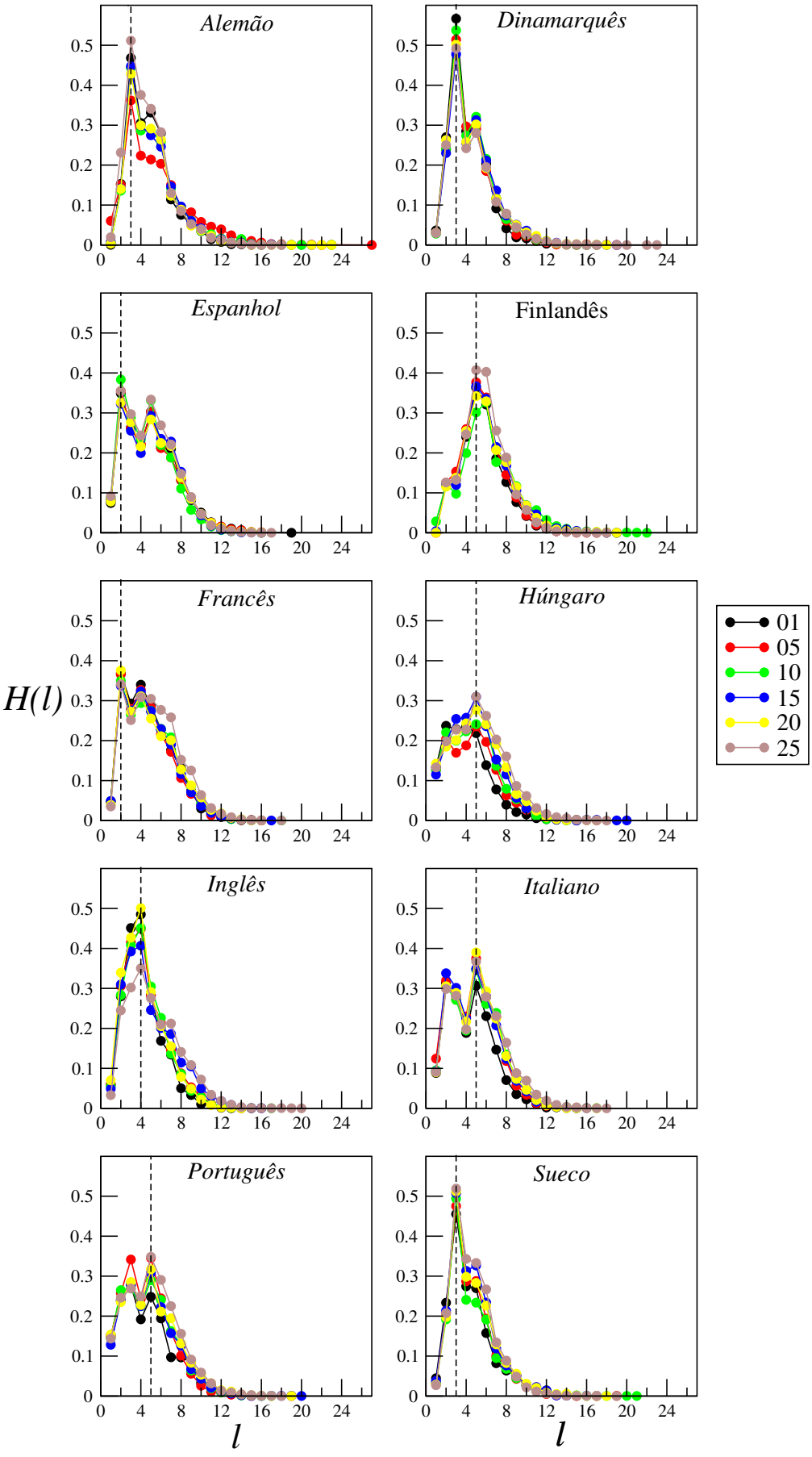


Figura 3.5 Gráfico da entropia $H(l)$ como função do tamanho l dos verbetes para os 01, 05, 10, 15, 20 e 25 de cada língua do *corpus* de textos literários (Apêndice A). As retas tracejadas indicam o valor máximo global de entropia por idioma.

CAPÍTULO 4

DISTRIBUIÇÃO ESPACIAL DE PALAVRAS EM LINGUAGEM ESCRITA

Somos verbívoros, uma espécie que vive de palavras, e o significado e o uso da linguagem estão fadados a estar entre os principais objetos de nossa ponderação, de nosso compartilhamento e de nossas disputas.

—STEVEN PINKER (Do que é Feito o Pensamento)

Neste capítulo abordamos o problema de como descrever a distribuição espacial de palavras em um texto. Investigamos, inicialmente, a relação entre o desvio padrão σ e a frequência k de ocorrência de um verbete, em seguida propomos dois modelos capazes de descrever os comportamentos limitantes para essa relação. Por fim, analisamos os efeitos das correlações espaciais sobre a entropia estrutural.

4.1 INTERMITÊNCIA COMO ESTIMADOR PARA DISTRIBUIÇÃO ESPACIAL DE PALAVRAS

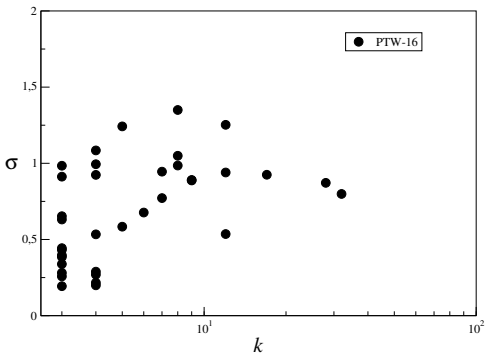
Conforme discutido no Capítulo 2, é possível observar diferentes regimes na distribuição espacial de palavras em textos. Nas últimas duas décadas diversos trabalhos têm proposto diferentes estimadores para, a partir da distribuição dos verbetes, extrair parâmetros que, dentre outras aplicações, podem ser utilizados para deteção de palavras-chave [33, 34, 35].

De modo a estudar o comportamento geral da distribuição espacial de verbetes em textos foi utilizado como estimador o desvio padrão σ da distância média entre verbetes, proposto por Ortuño e colaboradores [33]. Tal abordagem foi adotada devido o baixo custo computacional necessário para o estudo desse parâmetro. Na literatura da Linguística Quantitativa, o desvio padrão σ é também denominado como intermitência [55]. A intermitência só pode ser associada a verbetes cuja frequência seja maior que dois.

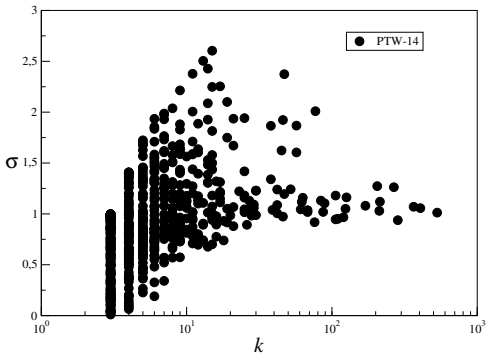
O desvio padrão cresce conforme aumenta a heterogeneidade da distribuição dos espaçamentos. Em particular uma distribuição poissoniana apresenta $\sigma = 1$. Os maiores valores do desvio padrão são tipicamente associados à palavras-chave [33, 34, 17, 35]. No Apêndice C são apresentamos o valores médios $\bar{\sigma}$ da intermiência para todos os textos utilizados em nosso estudo bem como a fração η de verbetes com desvio padrão maior que 1.

Nas Figuras 4.1 e 4.2 são apresentados os resultados do desvio padrão σ como função da frequência k para todos os verbetes de três pares de artigos da *Wikipedia* e três pares de textos literários em português (família latina), inglês (família germânica) e finlandês

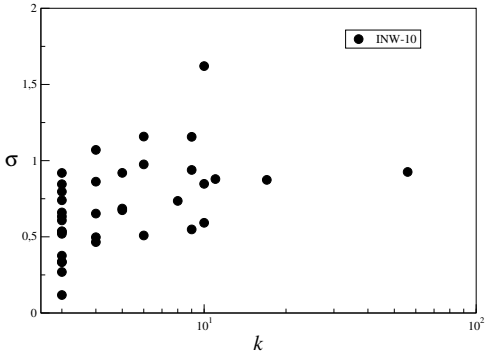
(família urálica). A forma da distribuição torna-se mais bem definida a medida que o número T de palavras do texto aumenta. Os maiores artigos da *Wikipedia* (Figura 4.1 (b), (d) e (f)) apresentam distribuições qualitativamente semelhantes àquelas associadas aos verbetes dos textos literários.



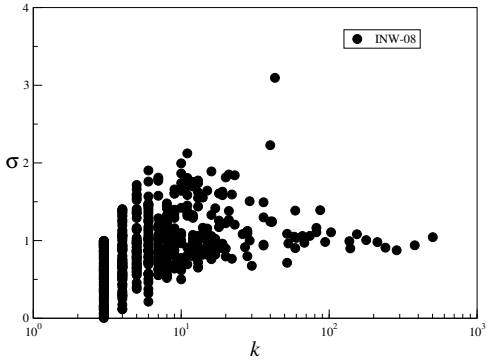
(a) Número total de palavras $T = 545$.



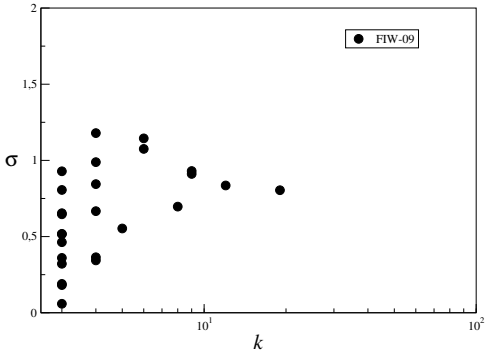
(b) Número total de palavras $T = 11039$.



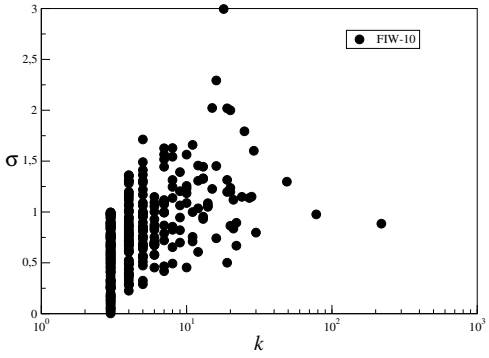
(c) Número total de palavras $T = 469$.



(d) Número total de palavras $T = 8958$.



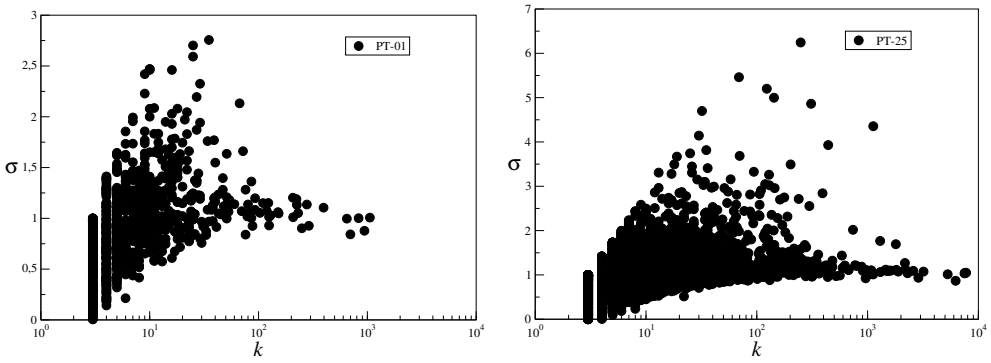
(e) Número total de palavras $T = 462$.



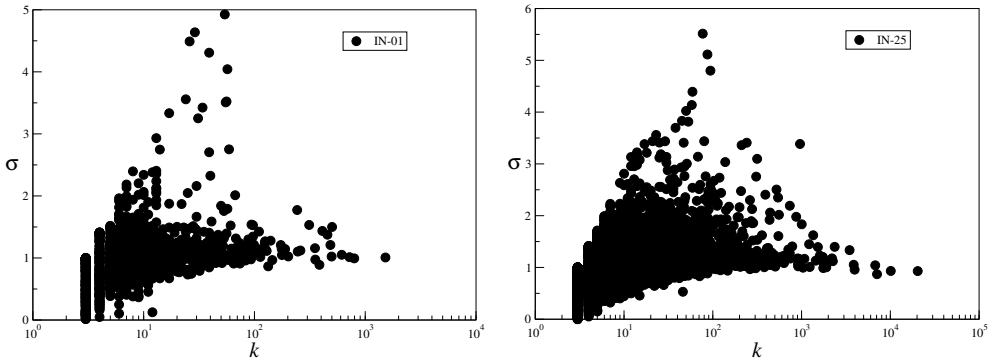
(f) Número total de palavras $T = 4904$.

Figura 4.1 Gráfico da relação entre o desvio padrão σ e número de ocorrências k para todos os verbetes de artigos da *Wikipedia* em português (PTW-16 e PTW-24), inglês (INW-10 e INW-08) e finlandês (FIW-09 e FIW-10) (Apêndice B).

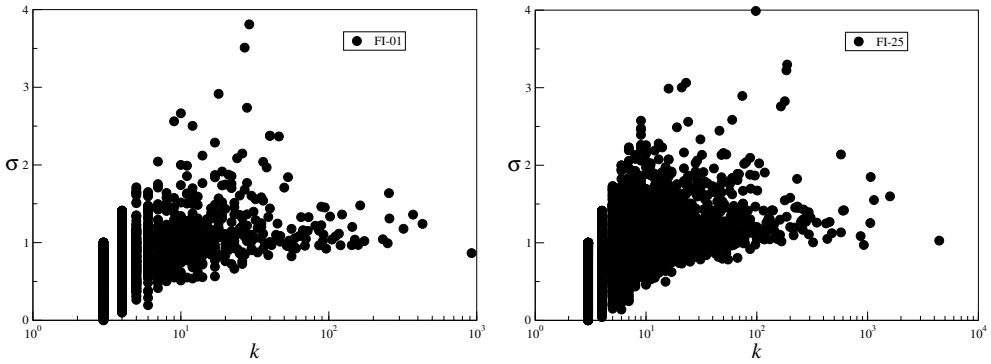
Os textos literários possuem maior número de verbetes apresentando assim distri-



(a) Número total de palavras $T = 21266$. (b) Número total de palavras $T = 212991$.



(c) Número total de palavras $T = 26540$. (d) Número total de palavras $T = 218316$.



(e) Número total de palavras $T = 26534$ (f) Número total de palavras $T = 89018$

Figura 4.2 Gráfico da relação entre o desvio padrão σ e número de ocorrências k para todos os verbetes de textos literários em português (PT-01 e PT-25), inglês (IN-01 e IN-25) e finlandês (FI-01 e FI-25) (Apêndice A).

buições mais bem definidas e portanto estes serão utilizados nos estudos relativos ao uso do desvio padrão como estimador. O menor texto apresentado tem o número total de palavras $T = 21266$ enquanto o maior texto apresentado tem $T = 218316$. Da Figura 4.2 podemos notar que o comportamento qualitativo da distribuição não depende do tamanho do texto. Outrossim a forma da distribuição apresenta-se invariante por língua ou

grupo linguístico.

Três características marcantes podem ser observadas dos gráficos expostos na Figura 4.2. Primeiro a ausência de verbetes com altos valores de desvio padrão no regime de baixas frequências. Em segundo lugar, um claro limitante inferior em todo o regime de frequências. E por fim, um comportamento poissoniano no regime de altas frequências:

$$\lim_{k\to\infty} \sigma = 1. \tag{4.1}$$

Essas características serão investigadas nas próximas seções deste capítulo.

4.2 MODELO HAMILTONIANO PARA A DISTRIBUIÇÃO ESPACIAL DE PALAVRAS

Em 2009, Pedro Carpena e colaboradores [34], utilizando uma generalização da análise estatística de níveis de energia de sistemas quânticos desordenados [53, 74], propuseram que a distribuição geométrica é um bom modelo para descrever o comportamento de palavras não relevantes. Nesse trabalho é possível extrair as seguintes expressões para a média da intermitência e seu desvio:

$$\langle\sigma\rangle = \frac{2k-1}{2k+2}\sqrt{1-\frac{k}{T}}, \tag{4.2}$$

$$sd(\sigma) = \frac{1}{\sqrt{k}(1+2.8k^{-0.865})}\sqrt{1-\frac{k}{T}}. \tag{4.3}$$

Na Figura 4.3 apresentamos o gráfico das expressões obtidas por Carpena e as quantidades equivalentes extraídas do livro *Os Maias* (PT-25 — Corpus A).

A abordagem descrita pelas equações 4.2 e 4.3 são análises estatísticas da média do desvio padrão. Embora esse resultado coincida com o comportamento médio da intermitência dos verbetes no regime de altas frequências, ele não fornece um limitante inferior para a relação entre σ e k. O perfil observado na Figura 4.2 nos induz a acreditar que a região inferior do gráfico é povoada por verbetes não relevantes cuja distribuição seria descorrelacionada.

Curiosamente, em 2004, Pedro Carpena e colaboradores [75] estudaram as distribuições de níveis de energia em cadeias com desordem correlacionada por meio do Hamiltoniano:

$$H = \sum_i \xi_i \left|i\right\rangle \left\langle i\right| + \sum_{\langle i,j\rangle} V \left|i\right\rangle \left\langle j\right|, \tag{4.4}$$

onde ξ_i representa a energia do sítio i, V é a energia de acoplamento entre os sítios i e j (primeiros vizinhos de i) e $i = \{1, 2, ..., N\}$, para um sistema de tamanho N. Neste modelo os sítios apresentam correlações de longo alcance segundo uma lei de potência. As correlações são introduzidas segundo:

$$\xi_i = \sum_{q=1}^{N/2} \left[q^{-\beta} \left(\frac{2\pi}{N}\right)^{1-\beta} \right]^{1/2} \cos\left(\frac{2\pi i q}{N} + \phi_q\right), \tag{4.5}$$

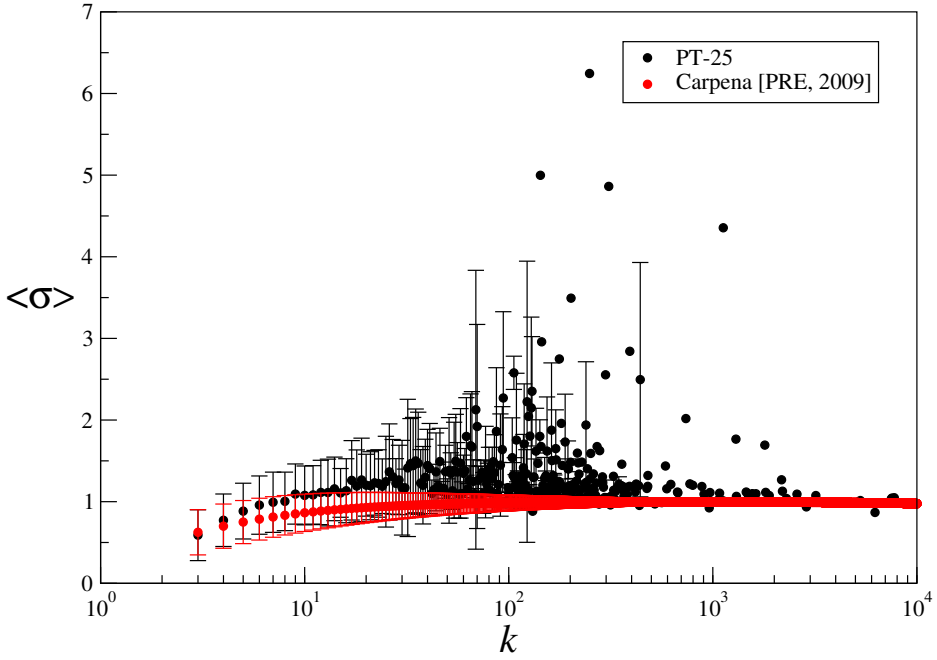


Figura 4.3 Gráfico do desvio padrão médio $\langle\sigma\rangle$ como função da frequência k para os verbetes do livro *Os Maias* (círculos pretos) e para a expressões 4.2 e 4.3 (círculos vermelhos).

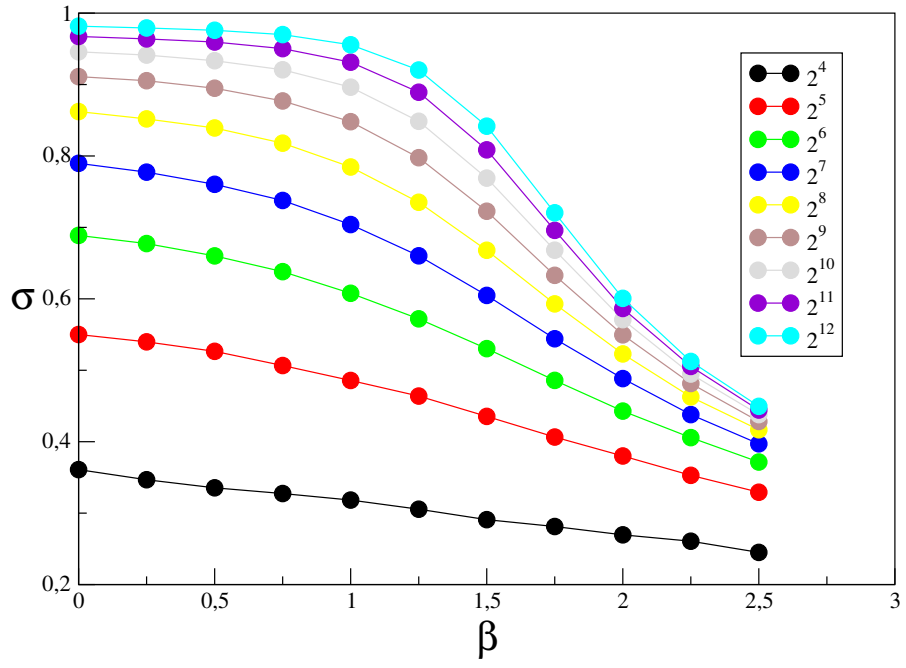


Figura 4.4 Gráfico do desvio padrão σ como função do expoente β da correlação para diferentes tamanhos de sistema. As linhas contínuas servem como guia para visualização.

onde ϕ_q são $N/2$ fases aleatórias uniformente distribuídas no intervalo $[0, 2\pi]$. Esse acoplamento gera um espectro de potência do tipo $1/q^{\beta}$. O parâmetro β controla o grau de correlação espacial, em particular $\beta = 0$ descreve um sistema totalmente descorrelacionado. Os resultados obtidos por esses autores quanto ao comportamento de σ como função do parâmetro β foram reproduzidos e são apresentados na Figura 4.4.

Seguindo a hipótese de que a região limitante inferior do gráfico é ocupada por verbetes descorrelacionados, decidimos investigar o comportamento de σ como função do tamanho N do sistema para o caso $\beta = 0$.

Na Figura 4.5 apresentamos um gráfico comparativo entre os valores de intermitência para o caso descorrelacionado com aqueles extraídos do texto *Os Maias*. Nessa abordagem podemos modelar a frequência de um verbete tomando $k = N$. É importante frisar que essa associação entre um sistema hamiltoniano descrito por 4.4 e a distribuição espacial de verbetes não havia sido reportada na literatura.

Aproveitando a característica fundamental de uma distribuição descorrelacionada, somos encorajados a propor um modelo mais simples que possa apresentar um limitante inferior para relação $\sigma(k)$ e que possua características qualitativas semelhantes às discutidas no Capítulo anterior.

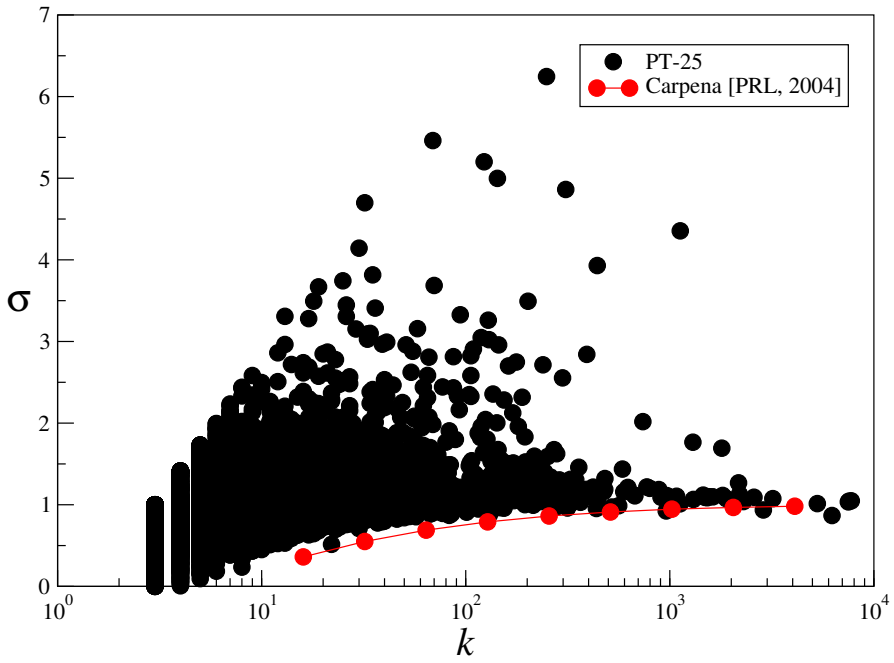


Figura 4.5 Gráfico do desvio padrão σ como função da frequência k para os verbetes do livro *Os Maias* (círculos pretos) e para os valores obtidos a partir do hamiltoniano da Equação 4.4 com $\beta = 0$ e diferentes tamanhos de sistema (círculos vermelhos). As linhas contínuas servem como guia para visualização.

4.3 MODELO DE NÚMEROS PRIMOS PARA A DISTRIBUIÇÃO DE PALAVRAS

Mapeando um texto como uma reta e sendo as posições dos verbetes definidas como números inteiros, é possível imaginar diferentes séries descorrelacionadas para representar a distribuição espacial de palavras. Muitos sistemas físicos são descritos segundo formulação de números primos [76, 77, 78], de modo que a sequência de números primos aparece como uma candidata natural para descrever a distribuição de verbetes descorrelacionados.

Para determinar a distribuição espacial de um verbete cuja frequência é k, iniciamos por colocar a primeira ocorrência na posição 2 e a partir daí em cada posição correspondente a um número primo. Em seguida são computadas as distâncias entre sucessivas ocorrências do verbete para então determinar a média e o desvio padrão σ da distribuição dos espaçamentos. Para comparar os valores da intermitência como função da frequência, o processo é realizado para frequências compreendidas entre $k = 3$ e $k = k_{max}$.

O estudo da distribuição dos espaçamentos entre números primos consecutivos tem atraído a atenção de físicos e matemáticos ao longo dos últimos dois séculos [79, 80, 81, 82]. Em 2014, Marek Wolf [83] apresentou resultados computacionais que dão apoio à hipótese de que essa distribuição é do tipo Poisson.

Na Figura 4.6, são comparados os valores de intermitência obtidos a partir da sequência de números primos com aqueles obtidos a partir do livro *Os Maias*. Observamos que para frequências superiores a $k = 10$, o modelo descreve qualitativamente a relação entre a

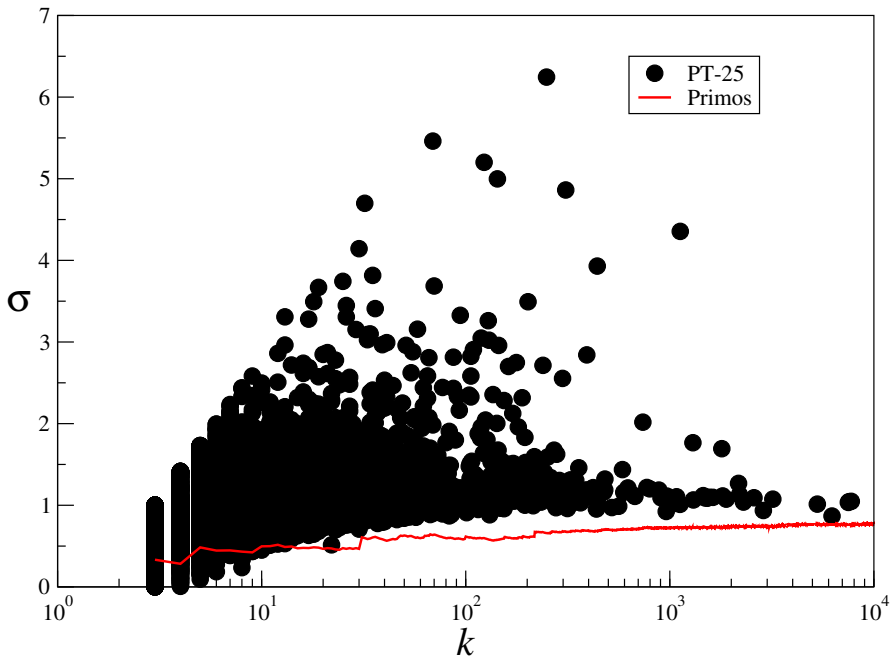


Figura 4.6 Gráfico da intermitência σ como função da frequência k para os verbetes do livro *Os Maias* (círculos pretos) e para os valores obtidos da sequência de números primos (linha vermelha).

intermitência e a frequência.

Com o objetivo de observar se a sequência de números primos obedece às mesmas leis discutidas no Capítulo anterior, computamos o tamanho mínimo do texto que comporta k_{max} números primos. Na Figura 4.7, apresentamos, em escala duplo-logarítmica, um gráfico da relação funcional entre k_{max} e T para a sequência de números primos. O comportamento qualitativo descrito pelo modelo é semelhante àquele observado em textos, conforme discutido no Capítulo anterior. Na região de interesse o número k_{max} de primos menores que T pode ser descrito segundo uma lei de potência com expoente $\nu = 0.90$, o que concorda quantitativamente com o valor médio calculado para todas as línguas. O comportamento assintótico dessa relação foi conjecturado independentente por Gauss [84] e Legendre [85] e é descrito:

$$k_{max} \sim \frac{T}{\ln(T)}. \tag{4.6}$$

O comportamento observado na Figura 4.7 aponta que o modelo de números primos para a distribuição espacial de palavras gera uma descrição adequada do limite inferior da relação $\sigma(k)$. Bem como reproduz características quantitativas próprias da linguagem escrita que não podem ser obtidas a partir do modelo hamiltoniano.

Como esse limite inferior é característico de palavras descorrelacionadas somos induzidos a concluir que o limite superior corresponde às palavras cuja distribuição é correlacionada.

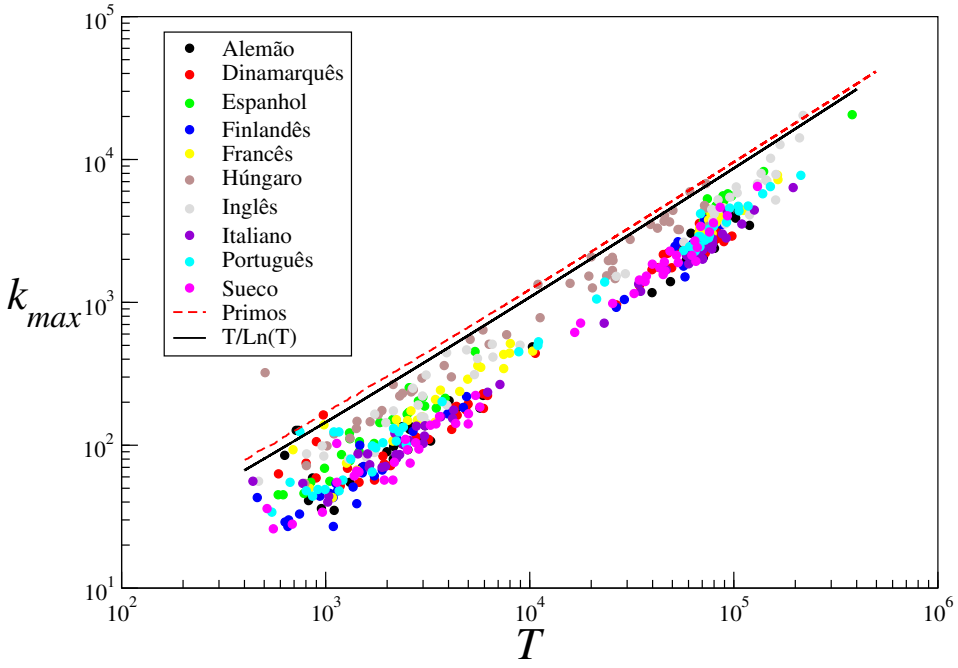


Figura 4.7 Gráfico da frequência máxima k_{max} como função do tamanho T para textos da *Wikipedia* e textos literários (pontos coloridos), para o modelo da distribuição de números primos (reta vermelha tracejada) e para a expressão assintótica dada pela equação 4.6 (linha preta contínua).

4.4 MODELO GEOMÉTRICO PARA A DISTRIBUIÇÃO DE PALAVRAS

Em oposição ao modelo totalmente descorrelacionado, podemos imaginar que a região superior da Figura 4.2 pode ser descrita por uma distribuição em que as posições do verbete são determinadas por um fator multiplicativo ℓ, em que as posições seguem uma sequência geométrica. Para um verbete cuja frequência seja k, podemos construir o conjunto $\{x\}$ das posições:

$$\{x\} = \{x_1, x_2, ..., x_k\}, \tag{4.7}$$

assumindo que a primeira ocorrência do verbete aconteça na posição $x_1 = 1$, temos:

$$x_i = \ell^{i-1}. \tag{4.8}$$

As distâncias $\{s\}$ entre sucessivas ocorrências do verbete podem ser escritas:

$$s_i = x_{i+1} - x_i = \ell^i - \ell^{i-1} \tag{4.9}$$

ou

$$s_i = \ell^i(1 - \ell^{-1}). \tag{4.10}$$

De posse de todas as $k-1$ distâncias, podemos calcular a distância média:

$$\langle s \rangle = \frac{1}{k-1}\sum_{i=1}^{k-1} s_i = \frac{1}{k-1}\frac{\ell-1}{\ell}\sum_{i=1}^{k-1}\ell^i \tag{4.11}$$

de forma que:

$$\langle s \rangle = \frac{\ell^{k-1}-1}{k-1}. \tag{4.12}$$

O segundo momento da distribuição das distâncias será:

$$\langle s^2 \rangle = \frac{1}{k-1}\sum_{i=1}^{k-1} s_i^2 = \left(\frac{\ell-1}{\ell}\right)^2 \frac{1}{k-1}\sum_{i=1}^{k-1}\ell^{2i} \tag{4.13}$$

$$\langle s^2 \rangle = \frac{(\ell-1)}{(\ell+1)}\frac{(\ell^{2(k-1)}-1)}{k-1} \tag{4.14}$$

De posse das Equações 4.12 e 4.14 é possível obter o valor da intermitência σ:

$$\sigma^2 \equiv \frac{\langle s^2 \rangle}{\langle s \rangle^2} - 1 \tag{4.15}$$

então:

$$\sigma^2 = (k-1)\frac{\ell-1}{\ell+1}\frac{(\ell^{k-1}+1)}{(\ell^{k-1}-1)} - 1 \tag{4.16}$$

Tomando $l \gg 1$ teremos:

$$\sigma^2 = (k-1) - 1 \tag{4.17}$$

Portanto o desvio padrão do modelo geométrico pode ser escrito como,

$$\sigma = \sqrt{k - 2}. \tag{4.18}$$

Na Figura 4.8, apresentamos um gráfico dos valores de intermitência previstos simultaneamente pelos modelos geométrico e de números primos em uma comparação com os maiores textos do nosso *corpus* em espanhol, inglês, italiano e finlandês. De fato os valores empiricamente observados são delimitados pelas duas curvas teóricas.

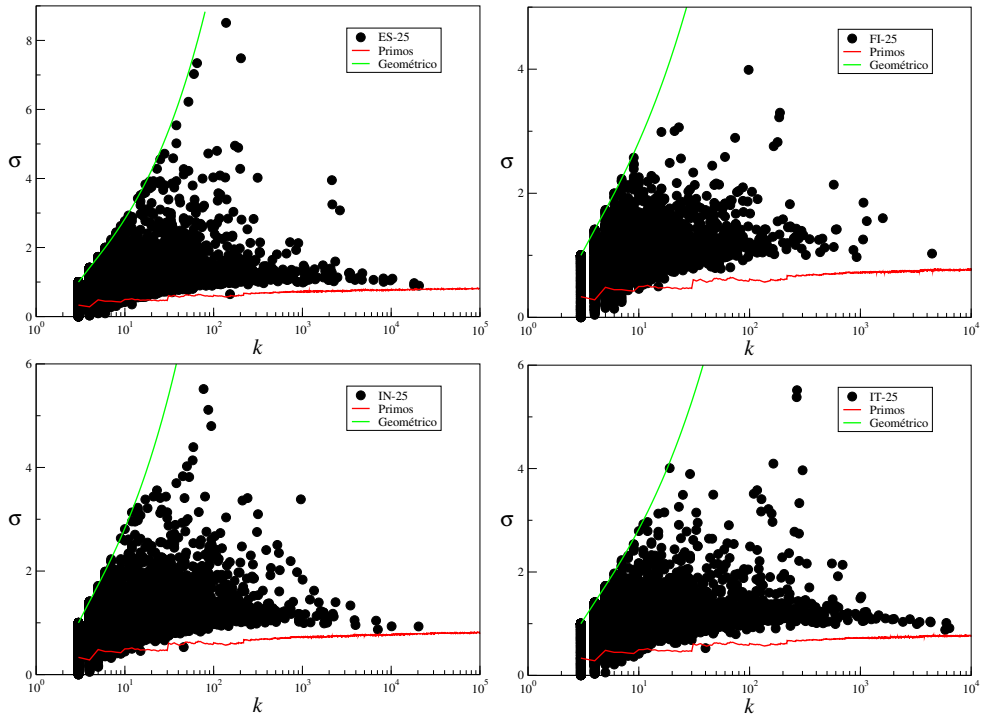


Figura 4.8 Gráfico da intermitência σ como função da frequência k para os verbetes dos livros ES-25, FI-25, IN-25 e IT-25 (círculos pretos), para os valores obtidos da sequência de números primos (linhas vermelhas) e para os valores obtidos analiticamente através da equação 4.18 (linhas verdes).

Como forma de exemplificar os verbetes que estão nos regimes descritos pelos dois modelos, destacamos na Figura 4.9 uma seleção de vinte verbetes representativos. Os valores de intermitência e frequência associados a esse grupo são apresentados na Tabela 4.1.

No Apêndice C é exposta a fração γ de verbetes que não se encontram na região delimitada pelas curvas dos modelos para todo o *corpus*.

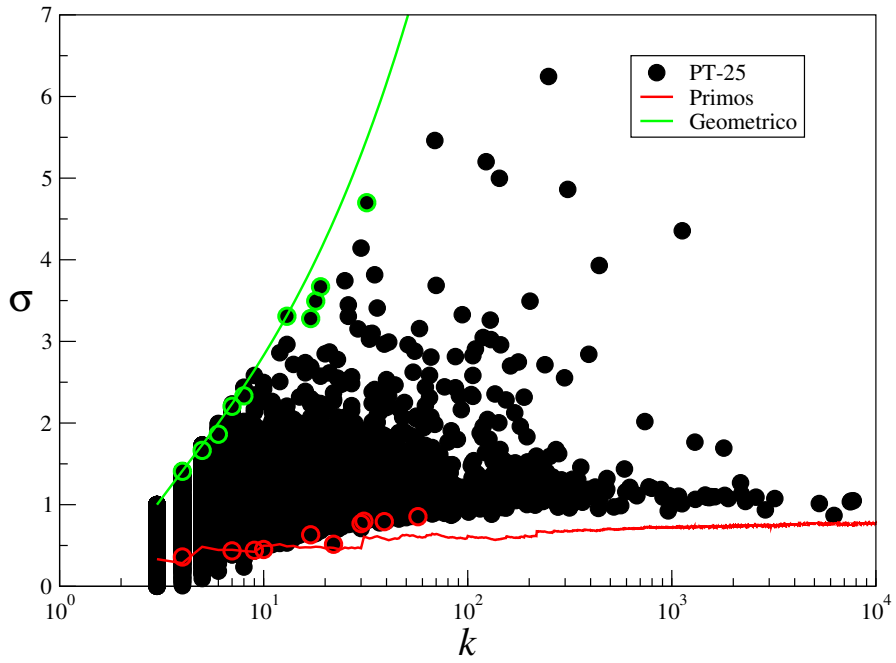


Figura 4.9 Gráfico da intermitência σ como função da frequência k para os verbetes do livro PT-25 (círculos pretos), para os valores obtidos da sequência de números primos (linha vermelha) e para os valores obtidos analiticamente através da equação 4.18 (linha verde). Os círculos verdes e vermelhos correspondem, respectivamente, aos dez verbetes representativos do modelo geométrico e de números primos apresentados na Tabela 4.1.

Verbete	Frequência	σ	Verbete	Frequência	σ
Collegios	4	1.41	Males	4	0.36
Joanninha	5	1.67	Ido	7	0.43
José	6	1.86	Mez	9	0.44
Marqueza	7	2.20	Ultimas	10	0.45
Silveiras	8	2.33	Certa	17	0.63
Pimenta	13	3.31	Pesado	22	0.52
Rughel	17	3.28	D'esta	30	0.77
Mentira	18	3.49	Qualquer	31	0.80
Lawrence	19	3.67	Escada	39	0.79
Abbade	32	4.70	N'aquela	57	0.85

Tabela 4.1 Verbetes representativos do modelo geométrico (Coluna 1) e de números primos (Coluna 4) para o livro *Os Maias*.

4.5 ENTROPIA ESPACIAL

O desvio padrão revela importantes características da distribuição espacial dos verbetes, embora ele seja construído a partir apenas de dois momentos dessa distribuição. De

forma a ampliar essa abordagem, buscamos um parâmetro que fosse capaz de quantificar com mais robustez a distribuição espacial de palavras.

Inspirados pela entropia de Shannon, apresentada no Capítulo 2, e pelo trabalho de Ali Mehri e Amir H. Darooneh [59] introduzimos a entropia:

$$H(w) = \langle \ln(p_w(d) \rangle = -\sum_{\{d\}} p_w(d) \ln p_w(d) \tag{4.19}$$

onde $p_w(d)$ representa a probabilidade de ocorrência da distância d associada ao verbete w, construída de forma similar ao cálculo da intermitência.

O comportamento típico dessa quantidade é comum a textos literários e artigos retirados da *Wikipedia* como pode ser observado nas Figuras 4.10 e 4.11. A curva dessa entropia possui concavidade para baixo de modo que seu ponto de máximo k_0 define dois regimes de $H(w)$ com a frequência.

Com os modelos apresentados anteriormente fomos capazes de definir duas situações limites: verbetes correlacionados e descorrelacionados. É possível para o modelo geométrico definir a dependência explicíta da entropia $H(w)$ com a frequência k. Para esse modelo todas as distâncias têm a mesma probabilidade $p_g(w)$ de ocorrência:

$$p_g(w) = \frac{1}{k-1}. \tag{4.20}$$

Da Equação 4.19:

$$H_g(w) = -\sum_{i=1}^{k-1} \frac{1}{k-1} \ln\left(\frac{1}{k-1}\right). \tag{4.21}$$

Assim, para o modelo geométrico:

$$H_g(w) = \ln(k-1). \tag{4.22}$$

Para uma sequência de números primos, M. Wolf [83] propôs que o comportamento assintótico da distribuição de probabilidades dos espaçamentos pode ser escrito:

$$p(d) = \frac{1}{\bar{d}(T)} e^{-d/\bar{d}(T)}, \tag{4.23}$$

Onde $\bar{d} = T/(k-1)$ representa a distância média entre ocorrências. Assim, a entropia associada à sequência dos primos será:

$$H_p(w) = 1 + \ln\left(\frac{T}{k-1}\right). \tag{4.24}$$

Na Figura 4.12, são apresentadas em escala log-linear as curvas descritas pelas Equações 4.22 e 4.24 juntamente aos valores obtidos a partir do livro *Os Maias*. Observamos que a entropia do modelo geométrico descreve de forma satisfatória a região de valores crescentes de $H(w)$ enquanto a entropia do modelo de números primos descreve adequadamente a região de inclinação negativa.

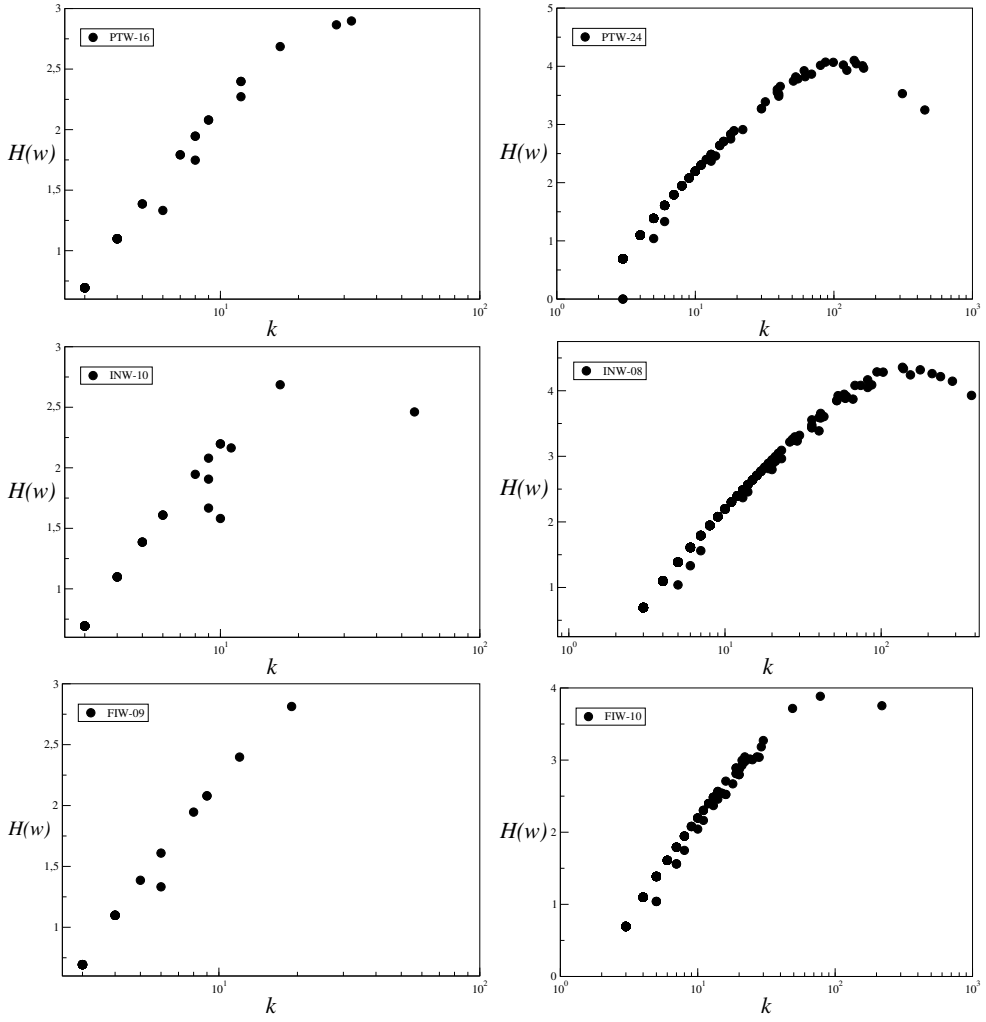


Figura 4.10 Gráfico da relação entre a entropia $H(w)$ e número de ocorrências k para todos os verbetes de artigos da *Wikipedia* em português (PTW-16 e PTW-24), inglês (INW-10 e INW-08) e finlandês (FIW-09 e FIW-10) (Apêndice B).

O valor de frequência k_0 que maximiza a entropia $H(w)$ pode ser obtido fazendo:

$$H_g(w) = H_p(w). \tag{4.25}$$

Assumindo $T \gg 1$, obtemos:

$$k_0 = \sqrt{eT}, \tag{4.26}$$

onde e é o número de Euler. Na Figura 4.12, são apresentados os valores de analíticos e experimentais de k_0 para o livro *Os Maias*.

Procedendo de forma semelhante, obtemos o valor máximo de entropia:

$$H_{max}(w) = \frac{1}{2}\ln(eT). \tag{4.27}$$

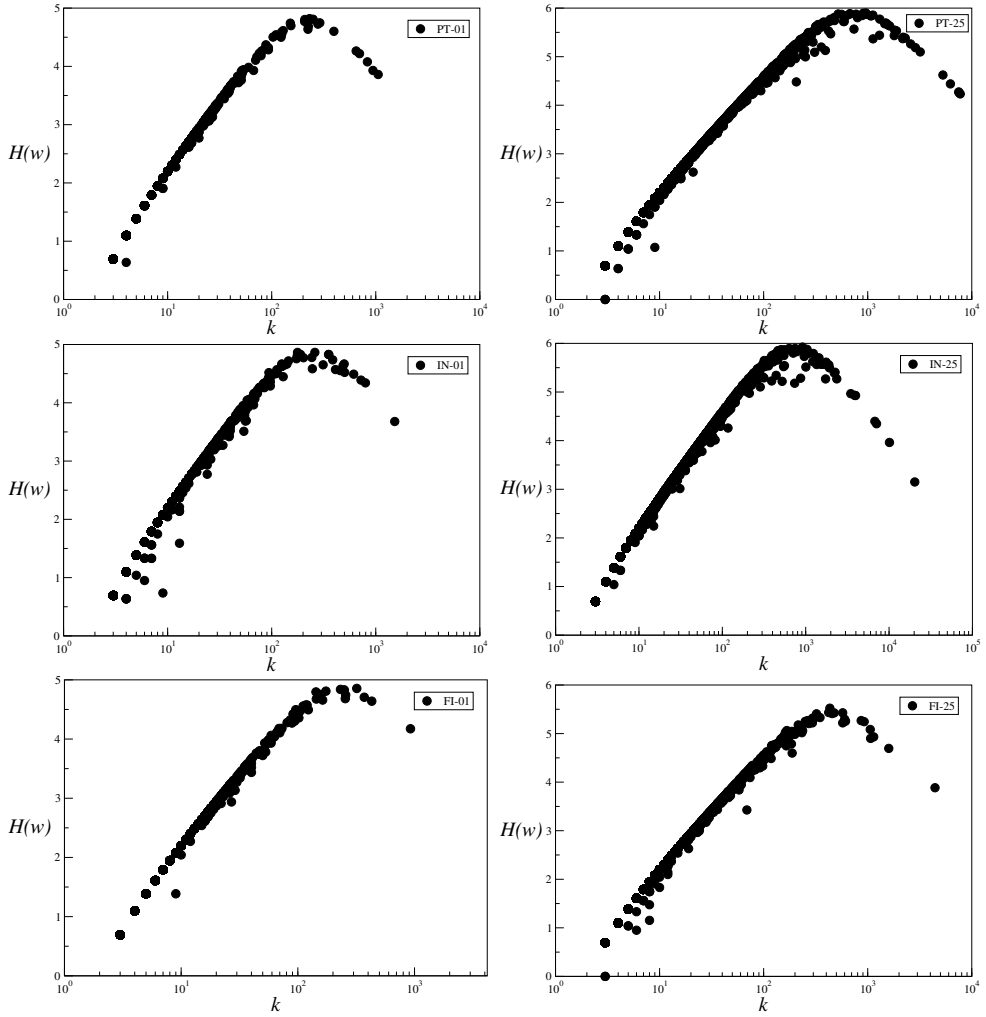


Figura 4.11 Gráfico da relação entre a entropia $H(w)$ e número de ocorrências k para todos os verbetes de textos literários em português (PT-01 e PT-25), inglês (IN-01 e IN-25) e finlandês (FI-01 e FI-25) (Apêndice A).

Utilizando todo o nosso *corpus* apresentamos na Figura 4.13 o gráfico da relação entre o tamanho T dos textos e o valor máximo de entropia associada aos verbetes. Os coeficientes α^* das regressões logarítmicas são apresentados na Tabela 4.2, a média global resultou no valor $\alpha^* = 0.46 \pm 0.03$. No gráfico podemos notar ainda que a regressão se torna mais adequada para textos literários ($T > 10^4$). Quando a regressão é realizada somente nessa região obtemos o valor médio $\alpha = 0.49 \pm 0.01$ (Tabela 4.2) coincidindo com o valor análitico dado pela Equação 4.27.

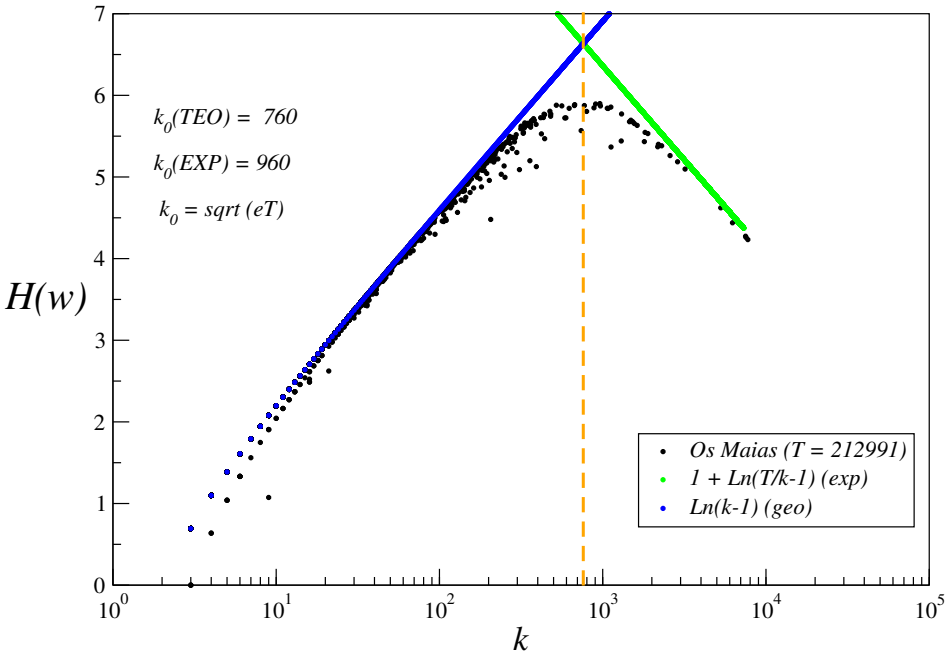


Figura 4.12 Gráfico da relação entre a entropia $H(w)$ e número de ocorrências k para todos os verbetes do livro *Os Maias* (círculos pretos), para o modelo geométrico (linha azul) e para o modelo de números primos (linha verde). A reta laranja tracejada apresenta o valor teórico $k_0 = 760$.

	Idioma	α	α^*
1	Alemão	0.49 ± 0.02	0.43 ± 0.02
2	Dinamarquês	0.49 ± 0.01	0.42 ± 0.02
3	Espanhol	0.49 ± 0.01	0.49 ± 0.01
4	Finlandês	0.49 ± 0.02	0.50 ± 0.01
5	Francês	0.49 ± 0.02	0.44 ± 0.02
6	Húngaro	0.52 ± 0.01	0.44 ± 0.02
7	Inglês	0.50 ± 0.01	0.48 ± 0.01
8	Italiano	0.51 ± 0.01	0.47 ± 0.01
9	Português	0.47 ± 0.01	0.42 ± 0.02
10	Sueco	0.49 ± 0.01	0.49 ± 0.01
	Média	0.49 ± 0.01	0.46 ± 0.03

Tabela 4.2 Valores dos coeficientes das regressões logarítmicas da entropia máxima $H_{max}(w)$ como função do tamanho T para textos literários (α) e para todo o corpus (α^*).

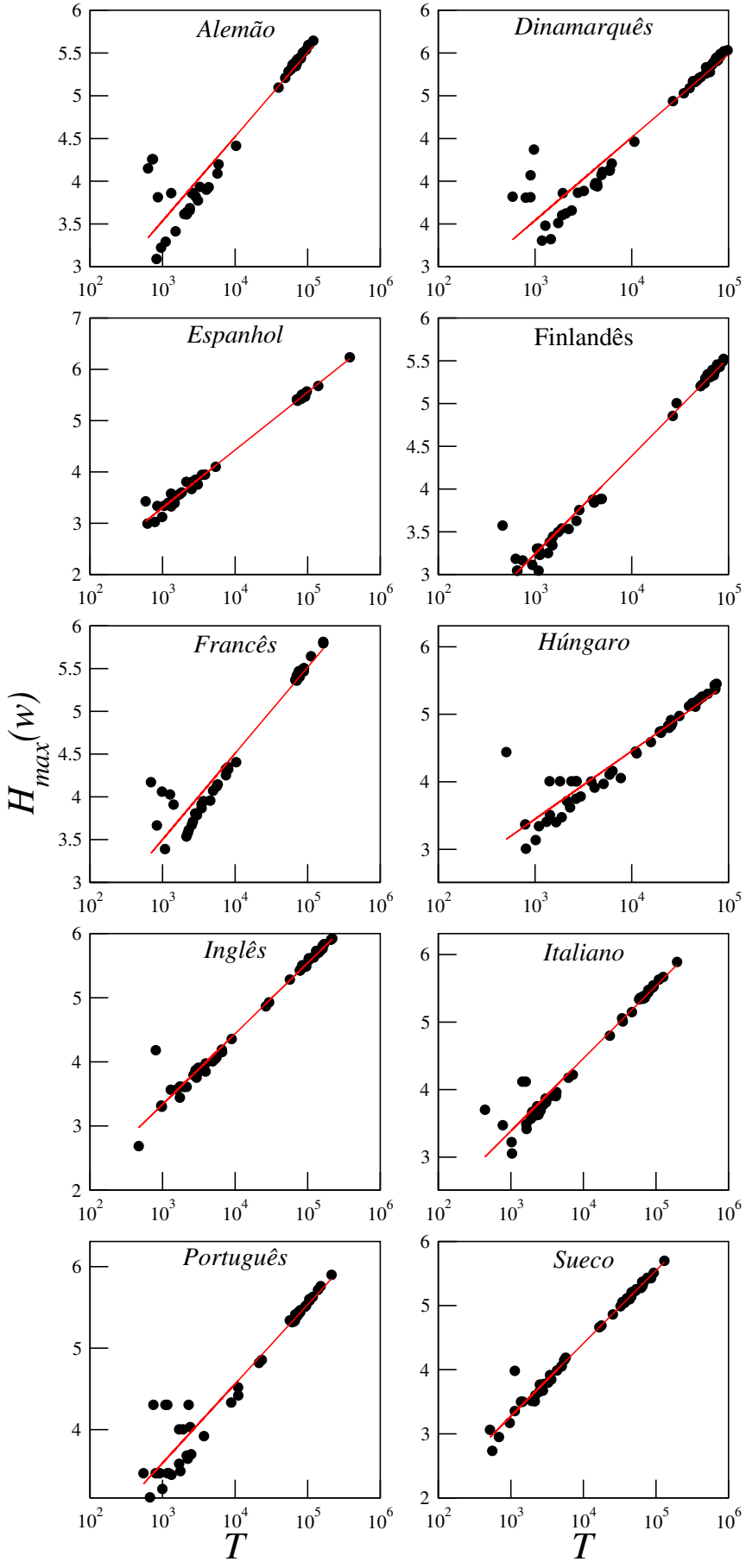


Figura 4.13 Gráfico em escala log-linear da relação entre a entropia máxima $H_{max}(w)$ e o número total de palavras T. As linhas vermelhas apresentam as regressões realizadas a partir dos dados extraídos do *corpus*.

4.6 ENTROPIA ESTRUTURAL

As expressões obtidas anteriormente para a entropia $H(w)$, associada à distribuição espacial dos verbetes, nos regimes correlacionado e descorrelacionado nos permitem definir um parâmetro que caracterize de forma global a heterogeneidade espacial do texto. Com essa idéia introduzimos a média da entropia espacial dos verbetes

$$\overline{H} = \frac{1}{V(k \geqslant 3)} \sum_{\{w\}} H(w) \tag{4.28}$$

que denominaremos *entropia estrutural*. Aqui, o termo $V(k \geqslant 3)$ representa a quantidade de verbetes que ocorrem ao menos 3 vezes.

Partindo da lei de Zipf e das definições para a entropia nos dois regimes $H_g(w)$ e $H_p(w)$ podemos obter expressões teóricas para o comportamento das principais quantidades frequencistas como a diversidade de vocabulário:

$$D = \frac{V}{T} = \frac{\sum\limits_{k=1}^{k_{max}} n(k)}{\sum\limits_{k=1}^{k_{max}} kn(k)} = \frac{\sum\limits_{k=1}^{k_{max}} k^{-\beta}}{\sum\limits_{k=1}^{k_{max}} k^{-(\beta-1)}}, \tag{4.29}$$

e a fração de palavras no regime exponencial:

$$f = \frac{\sum\limits_{k=k_o}^{k_{max}} kn(k)}{\sum\limits_{k=1}^{k_{max}} kn(k)} = 1 - \frac{\sum\limits_{k=1}^{k_o-1} kn(k)}{\sum\limits_{k=1}^{k_{max}} kn(k)} = 1 - \frac{\sum\limits_{k=1}^{k_o-1} k^{-(\beta-1)}}{\sum\limits_{k=1}^{k_{max}} k^{-(\beta-1)}}, \tag{4.30}$$

bem como a própria entropia estrutural

$$\overline{H} = \frac{1}{\sum\limits_{k=3}^{k_{max}} n(k)} \left[\sum_{k=3}^{k_o-1} H_g(w) + \sum_{k=k_0}^{k_{max}} H_p(w) \right] \tag{4.31}$$

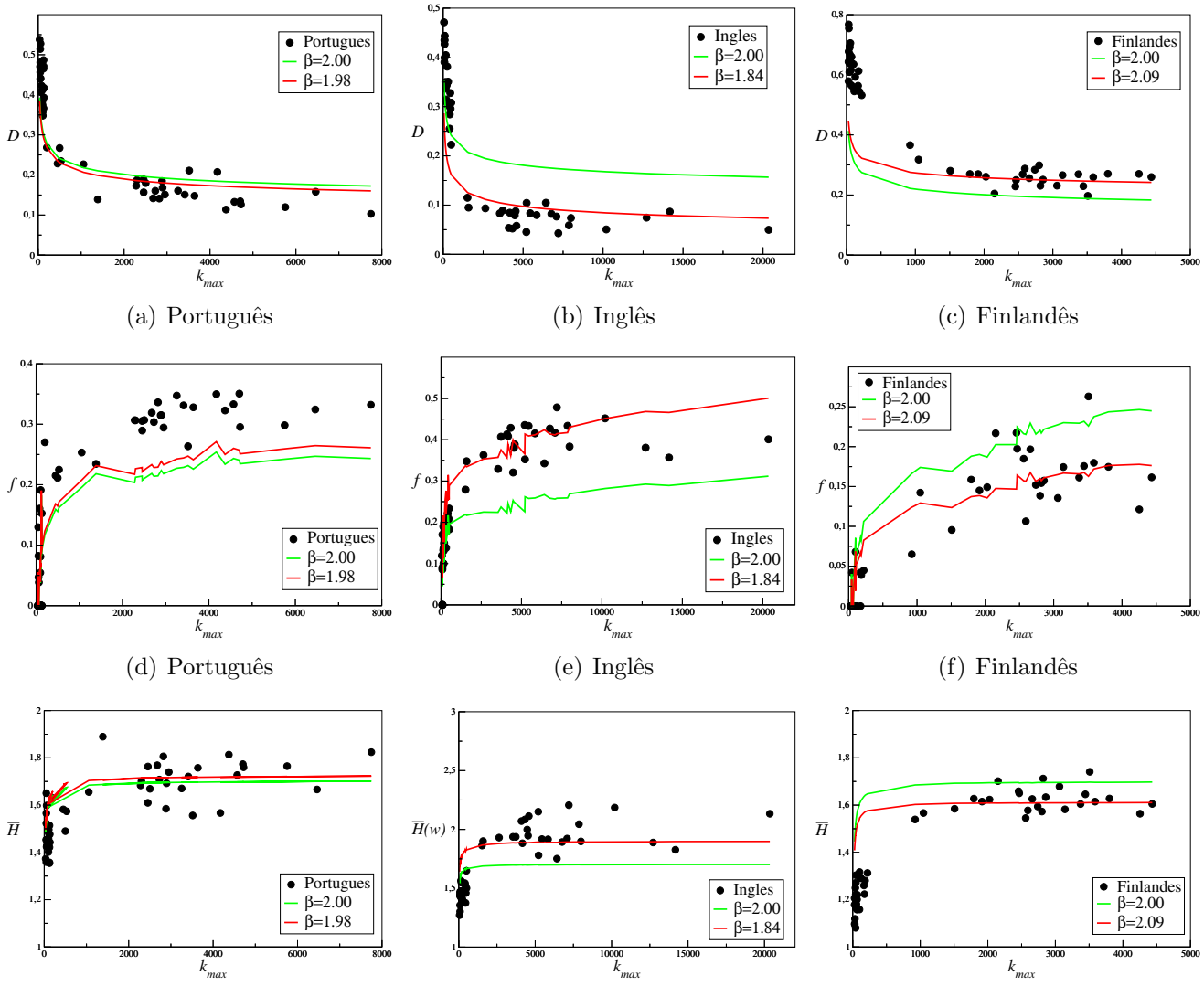
$$\overline{H} = \frac{1}{\sum\limits_{k=3}^{k_{max}} k^{-(\beta-1)}} \left[\sum_{k=3}^{k_o-1} k^{-\beta+1} Ln(k-1) + \sum_{k=k_0}^{k_{max}} k^{-\beta+1} Ln \left[\frac{eT}{(k-1)} \right] \right] \tag{4.32}$$

É instrutivo perceber que todas estas quantidades também podem ser computadas diretamente dos textos que constituem nosso *corpus* e que suas expressões envolvem parâmetros obtidos em nossa análise frequencista desenvolvida no Capítulo 2. Desta forma uma confrontação entre os valores empíricos e teóricos serve como *pedra de toque* para os esforços desenvolvidos até aqui.

Na Figura 4.14, exibimos os gráficos para os valores de diversidade D, fração f e entropia estrutural $\overline{H}$ para os 25 textos literários e 25 artigos da *Wikipedia* de três idiomas português, inglês e finlandês como função da frequência máxima k_{max} (círculos cheios).

Simultaneamente em cada subgráfico temos os respectivos valores destas grandezas calculadas por meio das Equações 4.29, 4.30 e 4.32 utilizando os expoentes para a Lei de Zipf β_{exp} obtidos experimentalmente no Capítulo anterior (linha contínua vermelha) e para $\beta = 2$ (linha contínua verde).

De modo geral as curvas com β_{exp} são mais próximas dos valores computados diretamente dos textos, em particular para a diversidade e entropia estrutural os resultados são particularmente adequados para o limite de altas frequências $k_{max} \gg 1$, enquanto que os valores previstos pelo exponte $\beta = 2$ indica comportamentos limitantes das quantidades.



(a) Português (b) Inglês (c) Finlandês

(d) Português (e) Inglês (f) Finlandês

(g) Português (h) Inglês (i) Finlandês

Figura 4.14 Gráficos da diversidade de vocabulário D, fração f de palavras no regime exponencial e entropia estrutural $\overline{H}$ como funções da frequência máxima de ocorrências k_{max} para todo o corpus em português, inglês e finlandês.

Para a situação em que $\beta = 2$, ou seja, quando temos a Lei de Zipf em sua forma original, as somas parciais que surgem na definição da fração f são denominadas número harmônico:

$$H_n \equiv \sum_{k=1}^{n} \frac{1}{k}, \tag{4.33}$$

utilizando esta definição e as leis empíricas para textos literários podemos expressar as constantes k_o e k_{max}, em termos do tamanho do texto (Equações 3.5 e 4.26). De modo que uma expresssão fechada para a fração f pode ser escrita como:

$$f_T = 1 - \frac{H_{\sqrt{eT}}}{H_{\epsilon T}}. \tag{4.34}$$

Assintoticamente o número harmônico é dado por:

$$H_n \sim Ln(n) + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} - \frac{1}{252n^6} + ..., \tag{4.35}$$

onde γ é denominada constante de Euler-Mascheroni. Um importante limitante para H_n é dado por $H_n < H_{max}$ [86, 87]:

$$H_{max} = Ln(n + 1/2) + \gamma + \frac{1}{24n^2}, \tag{4.36}$$

de onde resulta:

$$f_T = 1 - \frac{Ln(\sqrt{eT} + 1/2) + \gamma + \frac{1}{24eT}}{Ln(\epsilon T + 1/2) + \gamma + \frac{1}{24\epsilon^2 T^2}}. \tag{4.37}$$

Na Figura 4.15, apresentamos gráficos para o comportamento de $f(T)$ para os 25 textos literários de todas as línguas de nosso *corpus*, como uma função do tamanho T dos textos e uma comparação com a expressão analítica obtida na equação acima. Nesses gráficos utilizamos os valores de ϵ encontrados empiricamente no Capítulo 2 (Tabela 3.1).

Ao observarmos os gráficos percebemos que as línguas urálicas, finlandês e húngaro, possuem um comportamento qualitativo distinto das famílias germânica e latina. Enquanto que para estas duas últimas a curva $f(T)$ estabelece um limitante inferior para os dados empíricos, para a primeira o caso se revela oposto. A previsão analítica formulada a partir da lei de Zipf e dos diferentes tipos de correlação oferece um comportamento limitante inferior para esta quantidade. De fato, este resultado indica aque além da peculiar diferença dos tamanhos dos verbetes assinalada no Capítulo 2 a disposição espacial de palavras das línguas urálicas possuem características singulares, distintas dos outros grupos.

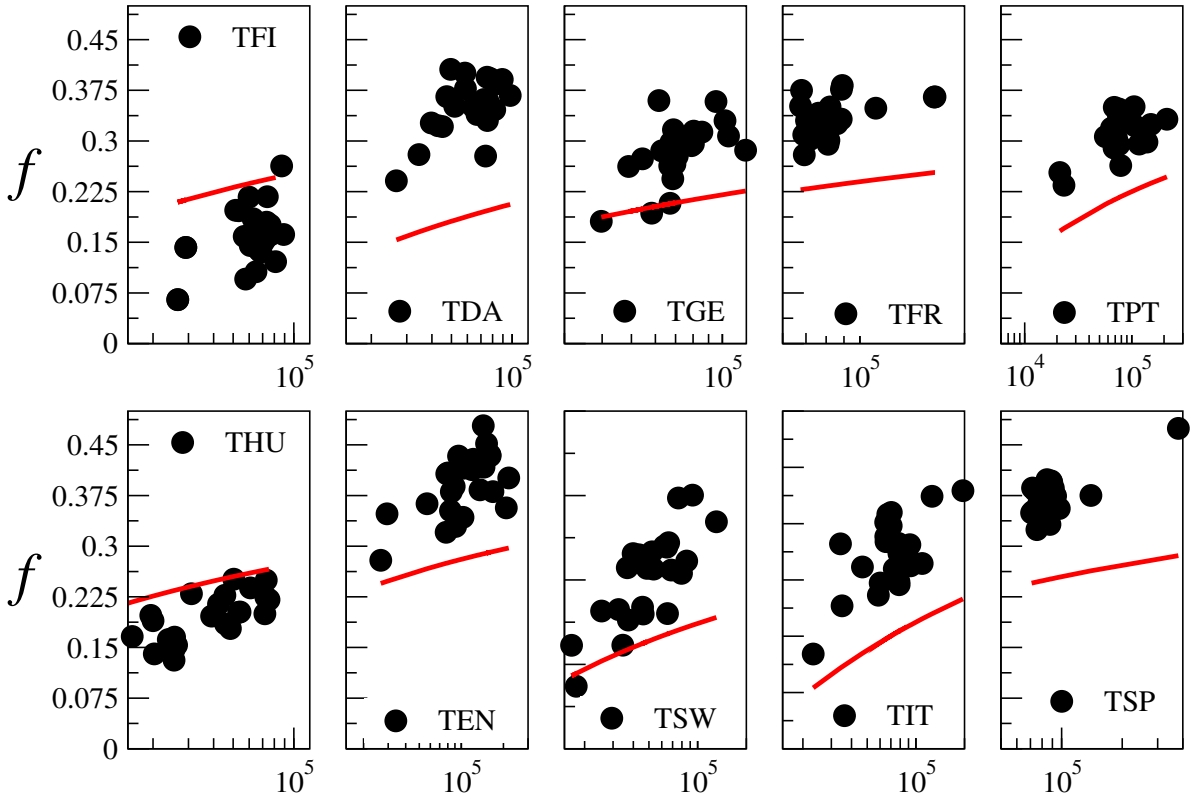


Figura 4.15 Gráfico da fração de palavras f com frequência acima do limiar k_o de máxima entropia estrutural, como função do tamanho do texto T, para todos os 25 textos literários do *corpus* (círculos), a linha vermelha indica a expressão analítica (Equação 4.34).

4.7 FRAÇÃO CRÍTICA DE EMBARALHAMENTO

De forma a apreciar o papel desempenhando pela fração f no limiar das correlações espaciais dos verbetes, implementamos um processo de embaralhamento dos textos. O método consiste em tomar uma fração p de palavras, sorteadas aleatoriamente, e trocar suas posições, a seguir, para diferentes graus de embaralhamento, são calculadas diversas grandezas como a intermitência média:

$$\overline{\sigma} = \frac{1}{V(k \geqslant 3)} \sum_{\{w\}} \sigma_w \tag{4.38}$$

e a entropia estrutural $\overline{H}$.

Nas Figuras 4.16 e 4.17, são apresentados os gráficos, em escala log-linear, da intermitência média dos verbetes $\overline{\sigma}$ e da entropia estrutual média $\overline{H}$ como função da fração de embaralhamento p, para três textos literários de diferentes famílias linguísticas: inglês (TEN-01), finlandês (TFI-02) e português (TPT-01). Em todos os casos as quantidades estão normalizadas pelos seus valores para o caso original ou seja, sem embaralhamento.

Curiosamente o comportamento das grandezas $\overline{\sigma}$ e $\overline{H}$ revela-se similar àquele observado em f sistemas físicos que apresentam transições de fase [88], com a diminuição

monotônica da intermitência e aumento da entropia, independente do grupo linguístico. Interessa-nos aqui estimar os valores p^0 que determinam os patamares finais destas quantidades.

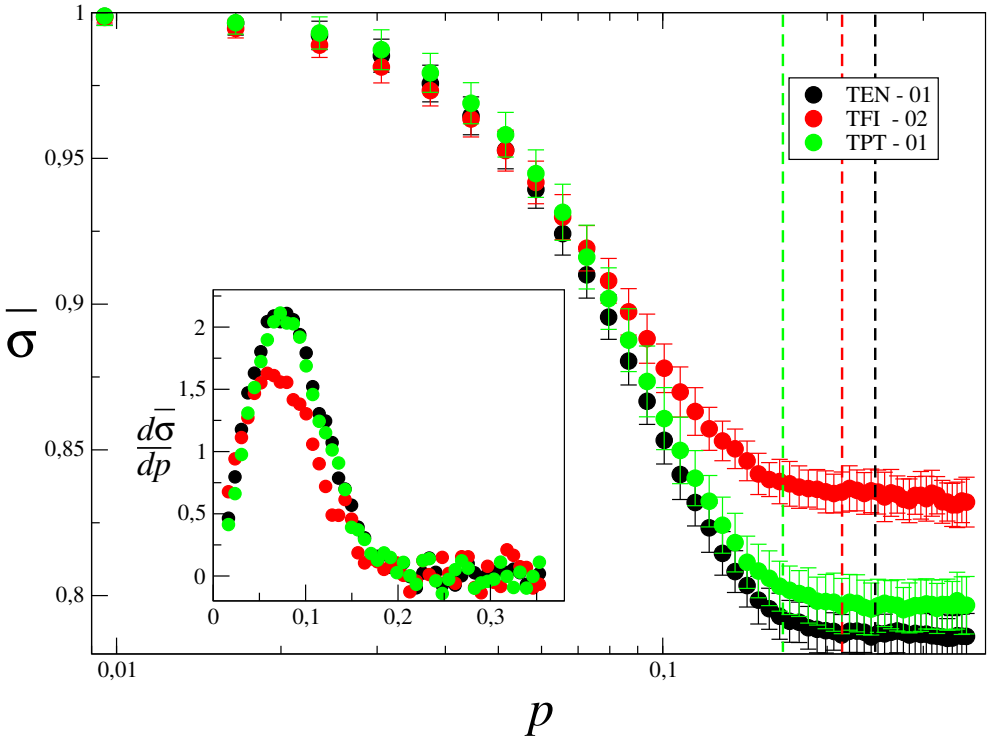


Figura 4.16 Gráfico em escala log-linear do comportamento da intermitência média dos verbetes $\overline{\sigma}$ com a fração de embaralhamento p para três textos literários de diferentes famílias linguísticas. As barras de erro correspondem ao desvio quadrático para 100 amostras. No sub-gráfico exibimos o comportamento da derivada em relação a fração p.

Consideramos uma estimativa conservadora na qual a fração p^0 pode ser estimada quando as flutuações das derivadas de $\overline{\sigma}$ e $\overline{H}$ com relação a p são da ordem 0.1 para a intermitência e 10^{-3} para a entropia. Nos subgráficos apresentados nas Figuras 4.16 e 4.17 são exibidos os comportamentos das derivadas das quantidades de interesse, de onde podem ser extraídos dos valores $p^0_{\overline{\sigma}}$ e $p^0_{\overline{H}}$, apresentados na Tabela 1. De forma geral os valores teóricos f_T são compatíveis com aqueles observados dos dados empíricos f e os obtidos a partir da análise da entropia estrutural, enquanto que os valores $p^0_{\overline{\sigma}}$ fornecem um limitante superior para o valor de embaralhamento a partir do qual os patamares tanto da intermitência média quanto da entropia permanecem inalterados. Este resultado indica que a fração f de palavras está associada ao conjunto de verbetes que possuem correlações de longo alcance ao longo do texto.

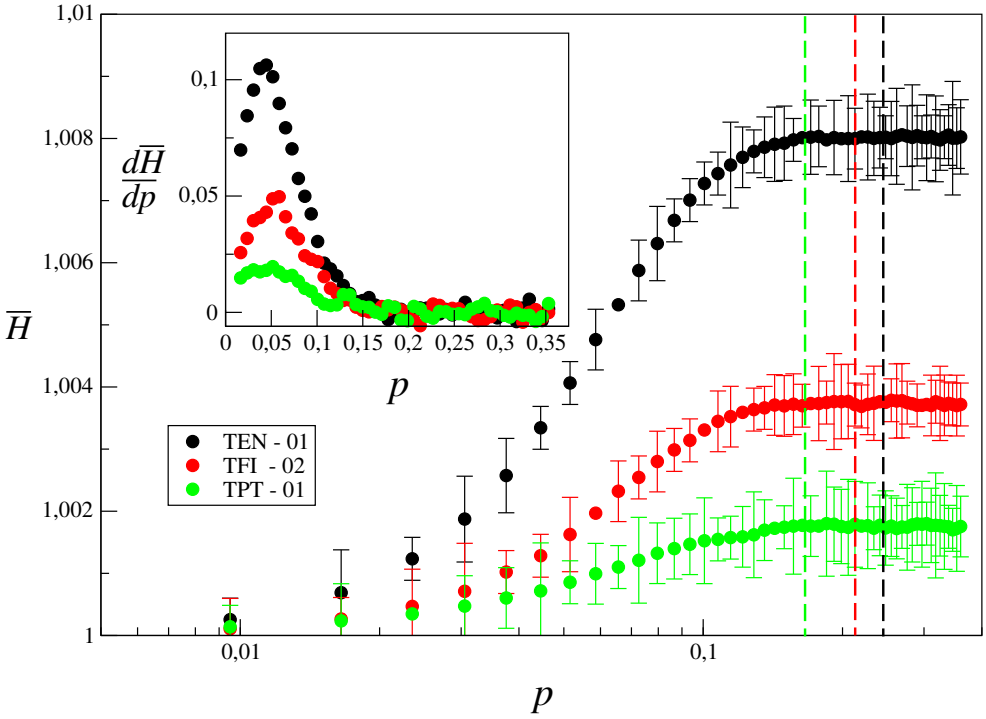


Figura 4.17 Gráfico em escala log-linear do comportamento da entropia média dos verbetes $\overline{H}$ com a fração de embaralhamento p para três textos literários de diferentes famílias linguísticas. As barras de erro correspondem ao desvio quadrático para 100 amostras. No sub-gráfico exibimos o comportamento da derivada em relação a fração p.

	Texto	f	f_T	$p^0_{\overline{\sigma}}$	$p^0_{\overline{H}}$
1	TEN-01	0.280	0.245	0.28 ± 0.07	0.21 ± 0.07
2	TFI-02	0.142	0.213	0.30 ± 0.07	0.22 ± 0.07
3	TPT-01	0.253	0.166	0.33 ± 0.07	0.23 ± 0.07

Tabela 4.3 Estimativas do valor da fração f de verbetes com frequência acima de k_o extraídas do texto, previsão teórica f_T. Nas duas últimas colunas estimada pelas derivadas da intermitência média $f_{\overline{\sigma}}$ e da entropia média $f_{\overline{H}}$.

CAPÍTULO 5

CONCLUSÕES E PERSPECTIVAS

Para eles, porém, este foi apenas o começo da verdadeira história. Toda a vida deles neste mundo e todas as suas aventuras em Nárnia haviam sido apenas a capa e a primeira página do livro. Agora, finalmente, estavam começando o Capítulo Um da Grande História que ninguém na terra jamais leu: a história que continua eternamente e na qual cada capítulo é muito melhor do que o anterior.

—C. S. LEWIS (A Última Batalha)

Nesta dissertação estudamos os aspectos frequencistas e espaciais da distribuição de verbetes em textos e o papel dessas quantidades sobre a informação contida em linguagem escrita. Ao longo de nossa abordagem, todos os nossos resultados teóricos foram comparados com aqueles obtidos de um *corpus* composto por 500 textos que incluem obras literárias e artigos da *Wikipedia* de diversas épocas em 10 idiomas distribuídos em 3 famílias linguísticas: germânica (alemão, dinamarquês, inglês e sueco), latina (espanhol, italiano, francês e português) e urálica (finlandês e húngaro).

Ao explorarmos a relação de escala entre o vocabulário V e o tamanho dos textos T, obtivemos o valor médio $\lambda = 0.71 \pm 0.05$ para o expoente de Heaps (Equação 3.4) . Foi constatado que o ordenamento de línguas a partir desse expoente gera um agrupamento segundo grupos linguísticos. A partir do *corpus* de artigos da *Wikipedia* e textos literários, apresentamos resultados empíricos que apontaram que há uma relação funcional não linear entre a frequência máxima k_{max} e o número total de palavras do texto T. Discutimos que para textos literários ($T > 10^4$) o valor médio obtido para o expoente ν é 1 e obtivemos os valores do coeficiente angular da relação $k_{max} = \epsilon T$ para todos os idiomas. Embora esse comportamento qualitativo fosse discutido na literatura [68, 69], não eram conhecidos resultados quantitativos.

Assim, além de caracterizar e validar o *corpus* segundo às leis de escala previamente conhecidas (Zipf e Heaps) apresentamos resultados inéditos para a relação entre o expoente de Heaps de um idioma e seu respectivo grupo linguístico e para a relação entre a frequência máxima de um verbete e o tamanho do texto para diferentes línguas.

Quando analisamos as características morfológicas dos símbolos a partir da distribuição de tamanho $P(l)$ dos verbetes obtivemos sua respectiva entropia. Observamos que a forma qualitativa das curvas para esse parâmetro é semelhante para idiomas pertencentes a mesma família linguística. A família latina apresenta dois picos característicos e família germânica tem um segundo pico drasticamente reduzido. Essas duas famílias apresentam um cauda na distribuição para longos tamanhos de verbetes. Já a família urálica possui apenas um pico localizado em $l = 5$.

Inspirados pela estatística de níveis de energia de sistemas quânticos desordenados, utilizamos o desvio padrão (ou intermitência) σ como métrica para estudarmos a distribuição espacial de verbetes. Como forma de compreender o comportamento característico da intermitência σ com a frequência k, elaboramos uma analogia entre um sistema descorrelacionado descrito por hamiltoniano do tipo *tight-binding* e a distribuição espacial de verbetes.

Aproveitando a característica fundamental de uma distribuição descorrelacionada, propusemos o modelo de números primos como um limitante inferior mais simples para relação $\sigma(k)$. O comportamento qualitativo descrito pelo modelo de números primos demonstrou ser semelhante àquele observado em textos. Na região de interesse o número k_{max} de primos menores que T pode ser descrito segundo uma lei de potência com expoente $\nu = 0.90$ enquanto, obtivemos empiricamente, para textos literários: $\nu = 0.92 \pm 0.03$.

Para descrever o limite superior que corresponderia às palavras cuja distribuição é correlacionada, propusemos uma distribuição em que as posições do verbete seguem uma sequência geométrica. Também foi obtido o valor analítico para o desvio padrão desse modelo geométrico.

A partir de gráficos dos valores de intermitência previstos simultaneamente pelos modelos geométrico e de números primos em uma comparação com os maiores textos do nosso *corpus*, observamos que os valores empiricamente observados são adequadamente delimitados pelas duas curvas teóricas.

Embora o desvio padrão tenha revelado importantes características da distribuição espacial dos verbetes, apresentamos a entropia espacial $H(w)$ como um parâmetro capaz de quantificar com mais robustez a informação estrutural contida nos textos do nosso *corpus*. O comportamento típico de $H(w)$ se demonstrou comum a textos literários e artigos retirados da *Wikipedia*.

Seguindo o modelo geométrico, obtivemos uma expressão analítica para a dependência explicíta da entropia $H(w)$ com a frequência k para os verbetes correlacionados. Partindo do comportamento assintótico da distribuição de probabilidade dos espaçamentos de números primos, apresentamos a entropia espacial associada aos verbetes descorrelacionados.

Tais expressões analíticas foram apresentadas juntamente aos valores obtidos a partir do livro *Os Maias*. Sendo observado que a entropia do modelo geométrico descreve de forma satisfatória a região de valores crescentes de $H(w)$ enquanto a entropia do modelo de números primos descreve adequadamente a região de inclinação negativa. Os coeficientes das regressões logaritmicas do valor máximo de entropia $H(w)$ como função do tamanho T do texto coicidiram com a previsão analítica dada pelos modelos propostos.

Partindo da lei de Zipf e das formulações para a entropia nos dois regimes, obtivemos expressões teóricas para o comportamento da diversidade de vocabulário, da fração de palavras no regime exponencial e entropia estrutural. De modo geral as curvas desses três parâmetros com β_{exp} são mais próximas dos valores computados diretamente dos textos.

Ao observarmos os gráficos, com $\beta = 2$, da fração de palavras no regime exponencial percebemos que as línguas urálicas (finlandês e húngaro) possuem um comportamento qualitativo distinto das famílias germânica e latina.

Implementamos um processo de embaralhamento dos textos para estudarmos o papel

desempenhando pelas correlações espaciais de longo alcance dos verbetes. Para distintos níveis de embaralhamento foram computadas o desvio padrão médio $\overline{\sigma}$ e a entropia estrutural $\overline{H}$. É notável que comportamento das grandezas $\overline{\sigma}$ e $\overline{H}$ tenha se demonstrado similar aquele observado em diversos sistemas físicos que apresentam transições de fase. A partir de um dado limiar de embaralhamento p^0, essas quantidades mantiveram-se inalteradas.

No intuito de aprimorar a compreensão sobre o papel desempenhado pela fração f na estrutura de informação é importante modificar a forma de embaralhamento estabelecendo análises quanto a esse raio de ação.

Outra importante perspectiva de desdobramento desse trabalho é investigar modelos que possuam correlações parciais entre seus elementos. Uma primeira idéia é utilizar sequências de números co-primos.

Possíveis conexões entre a entropia espacial e a relevância de um verbete constituem um profícuo caminho a ser investigado.

APÊNDICE A

CORPUS - LIVROS

Como conjunto inicial de amostras de linguagem escrita utilizamos obras literárias disponíveis no *Project Gutenberg* [89]. Foram selecionados vinte e cinco livros, essencialmente romances, em dez idiomas: alemão, dinamarquês, espanhol, finlandês, francês, húngaro, inglês, italiano, português e sueco. De forma a obter obras representativas da literatura brasileira utilizamos também livros disponíveis no *Portal Domínio Público* [90]. Nas páginas seguintes apresentamos de forma detalhada todas as obras estudadas. Para cada livro é informado sua referência que é o código que representa o livro nos Capítulos 01, 02, 03 e 04 bem como no Apêndice C. Ao lado do título do livro são explicitados o(s) autor(es) e o release da sua referida fonte sendo PT: Project Gutenberg e DP: Domínio Público.

No Apêndice B descrevemos os artigos retirados da *Wikipedia*. Para esse conjunto são apresentados sua referência, o termo na língua original, a data de acesso e seu respectivo endereço eletrônico.

Alemão			
Referência	Livro	Autor(es)	Release
DE-01	Isabella von Ägypten	Achim von Arnim	PG May, 2000 Etext #2190
DE-02	Heidis Lehr- und Wanderjahre	Johanna Spyri	PG February, 2005 Ebook #7511
DE-03	Aus einer kleinen Garnison	Fritz Oswald Bilse	PG January 20, 2014 EBook #44719
DE-04	Der Wehrwolf	Hermann Löns	PG October 2, 2007 EBook #22824
DE-05	Die Luftschiffahrt der Gegenwart	Hermann Hoernes	PG April 10, 2013 EBook #42489
DE-06	Die Achatnen Kugeln	Kasimir Edschmid	PG March 27, 2012 EBook #39277
DE-07	Der Trotzkopf	Emmy von Rhoden	PG February 17, 2010 Ebook #31309
DE-08	Indienfahrt	Waldemar Bonsels	PG January 20, 2008 eBook #24377
DE-09	Die Räuberbande	Leonhard Frank	PG October 19, 2009 EBook #30281
DE-10	Die Frau von dreißig Jahren	Honoré de Balzac	PG August 11, 2008 EBook #26261
DE-11	Aus meinem Leben, Erster Teil	August Bebel	PG May 5, 2004 EBook #12267
DE-12	Celsissimus	Arthur Achleitner	PG November 4, 2004 EBook #13953
DE-13	Rittmeister Brand; Bertram Vogelweid	Marie Ebner von Eschenbach	PG February 8, 2010 eBook #31233
DE-14	In Purpurner Finsterniß	Michael Georg Conrad	PG April 29, 2012 EBook #39565
DE-15	Komödiantinnen	Walter Bloem	PG January 12, 2014 EBook #44647
DE-16	Der Pilger Kamanita	Karl Adolph Gjellerup	PG February 7, 2005 eBook #14962
DE-17	Luthers Glaube	Ricarda Octavia Huch	PG April 12, 2012 eBook #39430
DE-18	Die Wahlverwandtschaften	Johann Wolfgang von Goethe	PG November, 2000 Etext #2403
DE-19	Charles Fourier	August Bebel	PG October 21, 2006 EBook #19596
DE-20	Der Mann im Mond	Wilhelm Hauff	PG September 13, 2004 eBook #13451
DE-21	Die Liebesbriefe der Marquise	Lily Braun	PG April 29, 2013 eBook #42617
DE-22	Effi Briest	Theodor Fontane	PG January 18, 2010 EBook #5323
DE-23	Briefe an eine Freundin	Wilhelm von Humboldt	PG June 11, 2007 EBook #21801
DE-24	Reise in die Aequinoctial-Gegenden des neuen Continents. Band 1.	Wilhelm von Humboldt	PG September 3, 2007 Ebook #22492
DE-25	Lichtenstein	Wilhelm Hauff	PG October, 2004 EBook #6726

Dinamarquês			
Referência	Livro	Autor(es)	Release
DK-01	Guds Fred	Peter Nansen	PG July 24, 2013 EBook #43295
DK-02	Kaptajnen paa 15 Aar	Jules Verne	PG August 6, 2010 EBook #33360
DK-03	Ved vejen	Herman Bang	PG August 13, 2004 EBook #13175
DK-04	Etienne Gerards Bedrifter	Arthur Conan Doyle	PG July 12, 2009 EBook #29392
DK-05	Tine	Herman Bang	PG January 11, 2004 EBook #10686
DK-06	Julies Dagbog	Peter Nansen	PG January 7, 2012 EBook #38515
DK-07	To verdener	Knud Hjortø	PG October 13, 2012 Ebook #41045
DK-08	Faedra	Herman Bang	PG March 1, 2004 Ebook #11396
DK-09	Hans Råskov	Knud Hjortø	PG January 31, 2013 Ebook #41956
DK-10	Judith Fürste	Adda Ravnkilde	PG April 22, 2012 Ebook #39510
DK-11	Hvad Skovsøen gemte	Palle Rosenkrantz	July 21, 2013 Ebook #43275
DK-12	Af mit Levned	Johan Louis Ussing	PG June 15, 2011 Ebook #36430
DK-13	Slægten	Gustav Wied	PG October 1, 2011 EBook #37594
DK-14	Kongens Fald	Johannes Vilhelm Jensen	PG August 2, 2011 Ebook #36942
DK-15	En Nihilist	Stepniak	PG July 12, 2009 Ebook #29392
DK-16	Bjørneæt	Carit Etlar	PG September 21, 2013 Ebook #43781
DK-17	Fru Marie Grubbe	Jens Peter Jacobsen	PG November 24, 2013 Ebook #44275
DK-18	Ludvigsbakke	Herman Bang	January 25, 2004 EBook #10829
DK-19	Stuk	Herman Bang	June 24, 2004 EBook #12698
DK-20	Absalons Brønd	Sophus Bauditz	PG July 21, 2012 Ebook #40291
DK-21	Doktor Nikola	Guy Boothby	PG January 28, 2008 Ebook #24447
DK-22	Ved Nytaarstid i Nøddebo Præstegaard	Henrik Scharling	PG December 4, 2011 Ebook #38220
DK-23	Haabløse Slægter	Herman Bang	PG February 18, 2004 Ebook #11139
DK-24	Minna	Karl Gjellerup	PG August 28, 2010 Ebook #33562
DK-25	Germanernes Lærling	Karl Gjellerup	PG November 3, 2012 Ebook #41277

Espanhol			
Referência	Livro	Autor(es)	Release
ES-01	El tesoro misterioso	William Tufnell Le Queux	PG August 28, 2009 EBook #29830
ES-02	Juanita La Larga	Juan Valera	PGFebruary 22, 2011 EBook #16484
ES-03	Fiebre de amor (Dominique)	Eugène Fromentin	PG September 2, 2008 EBook #26508
ES-04	Oriente	Vicente Blasco Ibáñez	PG July 9, 2012 EBook #40182
ES-05	El origen del pensamiento	Armando Palacio Valdés	PG October 7, 2011 EBook #30535
ES-06	Cádiz	Benito Pérez Galdós	PG June 23, 2007 EBook #21906
ES-07	Silas Marner	George Eliot	PG March 13, 2008 EBook #24823
ES-08	El Mar	Jules Michelet	PG August 12, 2008 EBook #26284
ES-09	Entre naranjos	Vicente Blasco Ibáñez	PG September 28, 2009 EBook #30122
ES-10	Quilito	Carlos Maria Ocanto	PG October 14, 2007 EBook #23035
ES-11	Los pazos de Ulloa	Emilia Pardo Bazán	PG March 16, 2006 EBook #18005
ES-12	Las inquietudes de Shanti Andia	Pío Baroja	PG July 8, 2004 EBook #12848
ES-13	La letra escarlata	Nathaniel Hawthorne	PG August 6, 2011 EBook #36990
ES-14	Años de juventud del doctor Angélico	Armando Palacio Valdés	PG June 13, 2012 EBook #39990
ES-15	La gloria de don Ramiro	Enrique Larreta	PG September 6, 2009 EBook #29920
ES-16	Amaury	Alexandre Dumas	PG April 4, 2008 EBook #24988
ES-17	Angelina	Rafael Delgado	PG June 17, 2005 EBook #16082
ES-18	Su único hijo	Leopoldo Alas	PG December 17, 2005 EBook #17341
ES-19	En el Fondo del Abismo	Jorge Ohnet	PG December 2, 2004 EBook #14236
ES-20	Arroz y tartana	Vicente Blasco Ibáñez	PG August 2, 2005 EBook #16413
ES-21	La gaviota	Fernán Caballero	PG November 23, 2007 EBook #23600
ES-22	La maja desnuda	Vicente Blasco Ibáñez	PG June 24, 2013 EBook #43030
ES-23	La guardia blanca	Arthur Conan Doyle	PG June 17, 2011 EBook #36453
ES-24	Pequeñeces	Luis Coloma	PG December 3, 2006 EBook #20011
ES-25	Don Quijote	Miguel de Cervantes	PG April 27, 2010 EBook #2000

Finlandês			
Referência	Livro	Autor(es)	Release
FI-01	Vaihdokas	Juho Reijonen	PG January 30, 2005 EBook #14840
FI-02	Mennyt	Santeri Alkio	PG June 21, 2013 eBook #43000
FI-03	Elsa	Teuvo Pakkala	PG October 13, 2004 EBook #13733
FI-04	Laulu tulipunaisesta kukasta	Johannes Linnankoski	PG June 29, 2004 EBook #12780
FI-05	Puukkojunkkarit	Santeri Alkio	PG November 9, 2004 EBook #13991
FI-06	Vuonna 2000 Katsaus vuoteen 1887	Edward Bellamy	PG September 14, 2005 EBook #16694
FI-07	Ylhäiset ja alhaiset	K. J. Gummerus	PG November 30, 2004 EBook #14214
FI-08	Pikku mies	A. Daudet	PG April 28, 2013 EBook #42609
FI-09	Alroy	Benjamin D'Israeli	PG February 26, 2008 EBook #24687
FI-10	Palestiinassa	Kaarle August Hildén	PG December 23, 2005 EBook #17380
FI-11	Erämaan nuijamiehet	Santeri Ivalo	PG October 5, 2013 EBook #43890
FI-12	Heikki Helmikangas	Eero Sissala	PG March 26, 2007 EBook #20905
FI-13	Jerin veli Erään koiran elämä ja seikkailut	Jack London	PG July 20, 2013 EBook #43258
FI-14	Härkmanin pojat	Betty Elfving	PG April 17, 2005 EBook #15637
FI-15	Koston henki	August Blanche	PG June 28, 2008 EBook #25924
FI-16	Matka-kuvaelmia Englannista	Otto Funcke	PG February 7, 2011 EBook #35202
FI-17	Valkoisia kanervakukkia	Mathilda Roos	PG May 22, 2012 EBook #39756
FI-18	Häpeäpilkku	Ludwig Anzengruber	PG December 16, 2011 EBook #38322
FI-19	Sisaret	Georg Ebers	PG August 7, 2011 EBook #37001
FI-20	Marianne-rouva	Victoria Benedictsson	PG September 14, 2012 EBook #40761
FI-21	Yrjänä Kailanen ja hänen poikansa	Gustaf Schröder	PG September 5, 2005 EBook #16652
FI-22	Veneh'ojalaiset	Arvid Järnefelt	PG April 27, 2004 EBook #12182
FI-23	Seitsemän veljestä	Aleksis Kivi	PG April 7, 2004 EBook #11940
FI-24	Vilun-ihana	Berthold Auerbach	PG November 12, 2007 EBook #23461
FI-25	Panu	Juhani Aho	PG October 25, 2004 EBook #13850

Francês			
Referência	Livro	Autor(es)	Release
FR-01	Germaine	Edmond About	PG April 1, 2006 EBook #18092
FR-02	André Cornélis	Paul Bourget	PG November 25, 2007 EBook #23616
FR-03	La fille des indiens rouges	Émile Chevalier	PG April 26, 2006 EBook #18263
FR-04	Le Médecin des Dames de Néans	René Boylesve	PG February 20, 2009 EBook #28124
FR-05	Biribi	Georges Darien	PG August 8, 2005 EBook #16492
FR-06	La tombe de fer	Hendrik Conscience	PG December 6, 2005 EBook #17242
FR-07	Argent et Noblesse	Henri Conscience	PG December 13, 2005 EBook #17298
FR-08	L'américaine	Jules Claretie	PG March 28, 2006 EBook #18064
FR-09	Fierabras	Jehan Bagnyon	PG November 27, 2013 EBook #44301
FR-10	Le Crépuscule des Dieux	Élémir Bourges	PG February 6, 2013 EBook #42036
FR-11	Le chemin qui descend	Henri Ardel	PG January 20, 2010 EBook #31032
FR-12	Le vieux muet	Jean-Baptiste Caouette	PG November 25, 2004 EBook #14151
FR-13	Les grands froids	Émile Bouant	PG September 17, 2013 EBook #43760
FR-14	Le Guaranis	Gustave Aimard	PG January 20, 2014 EBook #44715
FR-15	Miss Rovel	Victor Cherbuliez	PG April 6, 2009 EBook #28523
FR-16	Pile et face	Lucien Biart	PG March 19, 2006 EBook #18014
FR-17	Mademoiselle Clocque	René Boylesve	PG July 23, 2006 EBook #18899
FR-18	Le Blé qui lève	René Bazin	PG February 1, 2010 EBook #31154
FR-19	Les parisiennes de Paris	Théodore de Banville	PG March 4, 2006 EBook #17915
FR-20	La Maison	Henry Bordeaux	PG June 19, 2004 EBook #12646
FR-21	Ce que disait la flamme	Hector Bernier	PG December 20, 2004 EBook #14399
FR-22	Un Coeur de femme	Paul Bourget	PG November 11, 2013 EBook #44161
FR-23	Madame Bovary	Gustave Flaubert	PG November 28, 2011 EBook #14155
FR-24	Belle-Rose	Amédée Achard	PG February 20, 2006 EBook #17808
FR-25	Germinal	Emile Zola	PG May, 2004 EBook #5711

Húngaro			
Referência	Livro	Autor(es)	Release
HU-01	Testamentum és Hat levél	Elek Benedek	PG December 9, 2012 eBook #41587
HU-02	Béla, a buta	Dezső Kosztolányi	PG September 13, 2012 EBook #40748
HU-03	Éjszaka	Sándor Bródy	PG March 15, 2014 eBook #45147
HU-04	Bukfenc	Gyula Krúdy	PG December 2, 2012 EBook #41539
HU-05	Az arany szalamandra	Ferenc Donászy	PG May 10, 2006 EBook #18365
HU-06	Carinus; A nagyenyedi két füzfa	Mór Jókai	PG December 20, 2012 eBook #41670
HU-07	A Mester	Miklós Surányi	PG January 27, 2007 EBook #19744
HU-08	Emberek	Sándor Bródy	PG September 23, 2013 EBook #43801
HU-09	Nyomor	Sándor Bródy	PG September 11, 2013 EBook #43694
HU-10	Esik a hó	Frigyes Karinthy	PG September 5, 2012 eBook #40669
HU-11	Az akarat szabadságáról	Arthur Schopenhauer	PG March 2, 2013 EBook #42242
HU-12	A három galamb	Lehel Kádár	PG September 8, 2012 EBook #40715
HU-13	Különféle magyarok meg egyéb népek	István Tömörkény	PG December 16, 2008 EBook #27546
HU-14	Magyar élet	István Bársony	PG December 19, 2008 EBook #27565
HU-15	Magyar népmesék	János Erdélyi	PG January 18, 2012 EBook #38605
HU-16	Magyarhon szépségei; A legvitézebb huszár	Mór Jókai	PG December 10, 2012 EBook #41601
HU-17	A przemysli repülő Regény a nagy háborúból	Kurt Matull	PG August 22, 2012 eBook #40561
HU-18	Eredeti népmesék	Lsszló Arany	PG February 13, 2012 eBook #38852
HU-19	Az Atlasz-család	Gergely Csiki	PG January 2, 2009 EBook #27685
HU-20	Grimm testvérek összegyüjtött meséi	Jacob Grimm Wilhelm Grimm	PG June 26, 2012 EBook #40088
HU-21	A vörös regina	Árpád Abonyi	PG December 27, 2010 EBook #34759
HU-22	Elbeszélések	Gergely Csiky	PG August 11, 2013 EBook #43443
HU-23	Álomvilág	Zoltán Ambrus	PG March 9, 2013 eBook #42286
HU-24	Szirmay Ilona	József Gaál	PG June 14, 2010 EBook #32816
HU-25	Végzetes tévedés	Lenke Beniczkyné Bajza	PG June 29, 2010 EBook #33026

Inglês			
Referência	Livro	Autor(es)	Release
IN-01	Alice's Adventures in Wonderland	Lewis Carroll	PG June 25, 2008 EBook #11
IN-02	Through the Looking-Glass	Charles Dodgson, AKA Lewis Carroll	PG December 29, 2008 EBook #12
IN-03	The Mysterious Affair at Styles	Agatha Christie	PG January 26, 2013 EBook #863
IN-04	Jerusalem	Selma Lagerloef	PG May 16, 2005 eBook #15837
IN-05	The Picture of Dorian Gray	Oscar Wilde	PG July 2, 2011 EBook #174
IN-06	The Interesting Narrative of the Life of Olaudah Equiano	Olaudah Equiano	PG March 17, 2005 EBook #15399
IN-07	The Scarlet Letter	Nathaniel Hawthorne	PG December 18, 2011 EBook #33
IN-08	The Million-Dollar Suitcase	Alice M. P. Newberry	PG August 31, 2009 EBook #29877
IN-09	At the Back of the North Wind	George MacDonald	PG July 8, 2008 EBook #225
IN-10	The Lee Shore	Rose Macaulay	PG August 28, 2005 eBook #16612
IN-11	Pascal's Pensées	Blaise Pascal	PG April 27, 2006 EBook #18269
IN-12	American Notes for General Circulation	Charles Dickens	PG February 18, 2013 eBook #675
IN-13	Gulliver's Travels	Jonathan Swift	PG June 15, 2009 eBook #829
IN-14	Sense and Sensibility	Jane Austen	PG May 25, 2008 EBook #161
IN-15	Pride and Prejudice	Jane Austen	PG October 12, 2012 EBook #1342
IN-16	A Pair of Blue Eyes	Thomas Hardy	PG July 8, 2008 EBook #224
IN-17	A Tale of Two Cities	Charles Dickens	PG November 28, 2009 EBook #98
IN-18	An Essay Concerning Humane Understanding	John Locke	PG January 6, 2004 EBook #10615
IN-19	Jude the Obscure	Thomas Hardy	PG September 13, 2005 eBook #153
IN-20	On the Origin of Species	Charles Darwin	PG January 22, 2013 EBook #1228
IN-21	Emma	Jane Austen	PG January 21, 2010 Etext #158
IN-22	Dracula	Bram Stoker	PG August 16, 2013 EBook #345
IN-23	Nostromo: A Tale of the Seaboard	Joseph Conrad	PG January 9, 2006 EBook #2021
IN-24	Moby Dick	Herman Melville	PG January 3, 2009 EBook #2701
IN-25	Narrative of the Voyages and Services of the Nemesis from 1840 to 1843	William Hutcheon Hall William Dallas Bernard	PG September 8, 2013 EBook #43669

Italiano			
Referência	Livro	Autor(es)	Release
IT-01	Ninnoli	Gerolamo Rovetta	PG March 1, 2009 EBook #28231
IT-02	Divina Commedia di Dante: Purgatorio	Dante Alighieri	PG August, 1997 Etext #1010
IT-03	I sogni dell'Anarchico	Ugo Mioni	PG April 26, 2008 EBook #25175
IT-04	Come l'onda...	Luigi Capuana	PG April 28, 2013 eBook #42610
IT-05	Vae victis!	Annie Vivanti	PG December 09, 2011 EBook #38259
IT-06	Il mistero del poeta	Antonio Fogazzaro	PG September 4, 2007 EBook #22504
IT-07	Tra cielo e terra	Anton Giulio Barrili	PG March 20, 2009 EBook #28374
IT-08	Il ritratto del diavolo	Anton Giulio Barrili	PG February 25, 2006 EBook #17858
IT-09	Gli 'ismi' contemporanei	Luigi Capuana	PG March 14, 2009 EBook #28325
IT-10	L'amore di Loredana	Luciano Zùccoli	PG November 16, 2010 EBook #34346
IT-11	Dal primo piano alla soffitta	Enrico Castelnuovo	PG December 13, 2009 EBook #30663
IT-12	Il peccato di Loreta	Alberto Boccardi	PG November 4, 2008 eBook #27158
IT-13	I coniugi Varedo	Enrico Castelnuovo	PG September 19, 2009 eBook #30030
IT-14	L'undecimo comandamento	Anton Giulio Barrili	PG March 13, 2009 EBook #28321
IT-15	Nana a Milano	Cletto Arrighi	PG January 17, 2013 EBook #9302
IT-16	Nella lotta	Enrico Castelnuovo	PG September 19, 2009 eBook #30032
IT-17	Il bacio della contessa Savina	Antonio Caccianiga	PG January 25, 2011 EBook #35065
IT-18	La disfatta	Alfredo Oriani	PG December 9, 2006 EBook #20061
IT-19	Castel Gavone	Anton Giulio Barrili	PG April 26, 2008 EBook #25181
IT-20	Ettore Fieramosca	Massimo D'Azeglio	PG January 30, 2014 EBook #44797
IT-21	La carità del prossimo	Vittorio Bersezio	PG April 26, 2008 EBook #25179
IT-22	Della storia d'Italia, v. 1-2	Cesare Balbo	PG November 14, 2006 EBook #19808
IT-23	La favorita del Mahdi	Emilio Salgari	PG April 26, 2008 EBook #25180
IT-24	Mater dolorosa	Gerolamo Rovetta	PG May 21, 2009 EBook #28910
IT-25	Manfredo Palavicino	Giuseppe Rovani	PG November 22, 2003 eBook #10215

Português			
Referência	Livro	Autor(es)	Release
PT-01	Descobrimento das Filippinas	Caetano Alberto	PG June 26, 2009 EBook #29243
PT-02	Saudades: história de menina e moça	Bernardim Ribeiro	PG January 6, 2009 EBook #27725
PT-03	Dona Guidinha do Poço	Manuel de Oliveira Paiva	DP 1986
PT-04	A Morte Vence	João José Grave	PG December 3, 2007 EBook #23687
PT-05	Dom Casmurro	Machado de Assis	DP 2081
PT-06	O Mysterio da Estrada de Cintra	Eça de Queiroz Ramalho Ortigão	PG February 12, 2007 EBook #20574
PT-07	O triste fim de Policarpo Quaresma	Afonso H. de Lima Barreto	DP 2028
PT-08	Viagens na Minha Terra	João Almeida Garrett	PG January 22, 2008 EBook #24401
PT-09	Maria Dusá	Lindolfo Rocha	DP 16838
PT-10	A Cidade e as Serras	Eça de Queirós	PG February 28, 2008 EBook #18220
PT-11	O Ateneu	Raul Pompéia	DP 2020
PT-12	A falência	Júlia Lopes de Almeida	DP 7552
PT-13	Quincas Borba	Machado de Assis	DP 2128
PT-14	Senhora	José de Alencar	DP 2026
PT-15	O Matuto	Franklin Távora	DP 1812
PT-16	Amor Crioulo	Abel Botelho	PG March 26, 2008 EBook #24919
PT-17	O Cortiço	Aluísio Azevedo	DP 1723
PT-18	Motta Coqueiro ou A pena de morte	José do Patrocínio	DP 7550
PT-19	As Vítimas-Algozes	Joaquim Manuel de Macedo	DP 2134
PT-20	Os retirantes	José do Patrocínio	DP 7551
PT-21	O Guarani	José de Alencar	DP 1843
PT-22	O Primo Bazilio	Eça de Queirós	PG June 13, 2013 EBook #42942
PT-23	O crime do padre Amaro	Eça de Queirós	PG April 13, 2010 EBook #31971
PT-24	Os Sertões	Euclides da Cunha	DP 2163
PT-25	Os Maias	Eça de Queirós	PG October 16, 2012 EBook #40409

Sueco			
Referência	Livro	Autor(es)	Release
SE-01	Jordens Inre	Otto Witt	PG August 16, 2009 EBook #29707
SE-02	Lifsbilder från finska hem 1 Bland fattigt folk	Minna Canth	PG February 6, 2007 EBook #20518
SE-03	Blindskår	Minna Canth	PG September 6, 2008 EBook #26547
SE-04	Noveller	Minna Canth	PG September 6, 2008 EBook #26546
SE-05	De vandrande djåknarne	Viktor Rydberg	PG November 15, 2011 EBook #9827
SE-06	Det går an	Carl Jonas L. Almqvist	PG January 11, 2005 EBook #14670
SE-07	Singoalla	Viktor Rydberg	PG April 26, 2009 EBook #28610
SE-08	Fårvillelser	Hjalmar Sôderberg	PG June 19, 2007 EBook #21862
SE-09	Mor i Sutre	Hjalmar Bergman	PG November 6, 2005 EBook #17015
SE-10	Inferno	August Strindberg	PG September 8, 2009 EBook #29935
SE-11	I Vårbrytningen	August Strindberg	PG November 7, 2010 EBook #34236
SE-12	Om Lars Johansson (Lucidor den olycklige)	Josef Linck	PG April 10, 2009 EBook #28539
SE-13	Boken om lille-bror Ett äktenskaps roman	Gustaf af Geijerstam	PG April 2, 2009 EBook #28473
SE-14	David Ramms arv	Dan Andersson	PG May 5, 2006 EBook #18317
SE-15	En roman om förste konsuln	Mathilda Malling	PG December 18, 2007 EBook #23891
SE-16	Modern	Ernst Ahlgren Axel Lundegård	PG April 25, 2005 EBook #15703
SE-17	Barnen ifran Frostmofjaellet	Laura Fitinghoff	PG November 12, 2011 EBook #9828
SE-18	Himlauret eller det profetiska ordet	F. Franson	PG May 7, 2005 EBook #15786
SE-19	Elsa Finne I-II	Axel Lundegård	PG May 12, 2005 EBook #15821
SE-20	Eros' begravning	Hjalmar Bergman	PG June 14, 2004 EBook #12613
SE-21	Bannlyst	Selma Lagerlôf	PG March 14, 2012 EBook #39147
SE-22	Vi Bookar, Krokar och Rothar	Hjalmar Bergman	PG April 28, 2005 EBook #15724
SE-23	Hemsôborna	August Strindberg	PG September 25, 2009 EBook #30078
SE-24	Dagdrömmar En man utan humor I	Gustaf Hellström	PG May 31, 2005 EBook #15959
SE-25	Folkungatrådet	Verner von Heidenstam	PG September 4, 2004 EBook #13371

APÊNDICE B

CORPUS - WIKIPEDIA

Alemão			
Referência	Verbete	Data de Acesso	Endereço
DEW-01	Art (Biologie)	19/07/2014	http://de.wikipedia.org/wiki/Art_(Biologie)
DEW-02	Automobil	19/07/2014	http://de.wikipedia.org/wiki/Automobil
DEW-03	Bakterien	19/07/2014	http://de.wikipedia.org/wiki/Bakterien
DEW-04	Berg	19/07/2014	http://de.wikipedia.org/wiki/Berg
DEW-05	Bevölkerung	19/07/2014	http://de.wikipedia.org/wiki/Bevölkerung
DEW-06	Biometrie	19/07/2014	http://de.wikipedia.org/wiki/Biometrie
DEW-07	Boden (Bodenkunde)	19/07/2014	http://de.[...]/Boden_(Bodenkunde)
DEW-08	Galileo Galilei	19/07/2014	http://de.wikipedia.org/wiki/Galileo_Galilei
DEW-09	Glauben	19/07/2014	http://de.wikipedia.org/wiki/Glauben
DEW-10	Leben	19/07/2014	http://de.wikipedia.org/wiki/Leben
DEW-11	Lunge	19/07/2014	http://de.wikipedia.org/wiki/Lunge
DEW-12	Mann	19/07/2014	http://de.wikipedia.org/wiki/Mann
DEW-13	Mathematik	19/07/2014	http://de.wikipedia.org/wiki/Mathematik
DEW-14	Mensch	19/07/2014	http://de.wikipedia.org/wiki/Mensch
DEW-15	Moral	19/07/2014	http://de.wikipedia.org/wiki/Moral
DEW-16	Natur	19/07/2014	http://de.wikipedia.org/wiki/Natur
DEW-17	Pädagogik	19/07/2014	http://de.wikipedia.org/wiki/Pädagogik
DEW-18	Philosophie	19/07/2014	http://de.wikipedia.org/wiki/Philosophie
DEW-19	Sumer	19/07/2014	http://de.wikipedia.org/wiki/Sumer
DEW-20	Text	19/07/2014	http://de.wikipedia.org/wiki/Text
DEW-21	Tod	19/07/2014	http://de.wikipedia.org/wiki/Tod
DEW-22	Verbrechen	19/07/2014	http://de.wikipedia.org/wiki/Verbrechen
DEW-23	Vernunft	19/07/2014	http://de.wikipedia.org/wiki/Vernunft
DEW-24	Vulkan	19/07/2014	http://de.wikipedia.org/wiki/Vulkan
DEW-25	Zeit	19/07/2014	http://de.wikipedia.org/wiki/Zeit
Dinamarquês			
DKW-01	Armenien	20/07/2014	http://da.wikipedia.org/wiki/Armenien
DKW-02	Bibelen	20/07/2014	http://da.wikipedia.org/wiki/Bibelen
DKW-03	Big Bang	20/07/2014	http://da.wikipedia.org/wiki/Big_Bang
DKW-04	Bryst	20/07/2014	http://da.wikipedia.org/wiki/Bryst
DKW-05	Etnologi	20/07/2014	http://da.wikipedia.org/wiki/Etnologi
DKW-06	Fodbold	20/07/2014	http://da.wikipedia.org/wiki/Fodbold
DKW-07	Fotografi	20/07/2014	http://da.wikipedia.org/wiki/Fotografi
DKW-08	Frihed	20/07/2014	http://da.wikipedia.org/wiki/Frihed
DKW-09	Georges Simenon	20/07/2014	http://da.wikipedia.org/wiki/Georges_Simenon
DKW-10	Golfstrømmen	20/07/2014	http://da.wikipedia.org/wiki/Golfstrømmen

DKW-11	Klimaændring	20/07/2014	http://da.wikipedia.org/wiki/Klimaændring
DKW-12	Kød	20/07/2014	http://da.wikipedia.org/wiki/Kød
DKW-13	Magt	20/07/2014	http://da.wikipedia.org/wiki/Magt
DKW-14	Maya	20/07/2014	http://da.wikipedia.org/wiki/Maya
DKW-15	Merkur (planet)	20/07/2014	http://da.wikipedia.org/wiki/Merkur_(planet)
DKW-16	Natur	20/07/2014	http://da.wikipedia.org/wiki/Natur
DKW-17	Odin	20/07/2014	http://da.wikipedia.org/wiki/Odin
DKW-18	Penge	20/07/2014	http://da.wikipedia.org/wiki/Penge
DKW-19	Psykose	20/07/2014	http://da.wikipedia.org/wiki/Psykose
DKW-20	Rødlos	20/07/2014	http://da.wikipedia.org/wiki/Rødlos
DKW-21	Stat	20/07/2014	http://da.wikipedia.org/wiki/Stat
DKW-22	Træ(organisme)	20/07/2014	http://da.wikipedia.org/wiki/Træ_(organisme)
DKW-23	Universitet	20/07/2014	http://da.wikipedia.org/wiki/Universitet
DKW-24	Vegetarisme	20/07/2014	http://da.wikipedia.org/wiki/Vegetarisme
DKW-25	WFP	20/07/2014	http://da.[...]/World_Food_Programme
		Espanhol	
ESW-01	Agua dulce	19/07/2014	http://es.wikipedia.org/wiki/Agua_dulce
ESW-02	Bazo	19/07/2014	http://es.wikipedia.org/wiki/Bazo
ESW-03	Biosfera	19/07/2014	http://es.wikipedia.org/wiki/Biosfera
ESW-04	Codificación neural	19/07/2014	http://es.wikipedia.org/wiki/Codificación_neural
ESW-05	Deidad	19/07/2014	http://es.wikipedia.org/wiki/Deidad
ESW-06	Dinero	19/07/2014	http://es.wikipedia.org/wiki/Dinero
ESW-07	Educación	19/07/2014	http://es.wikipedia.org/wiki/Educación
ESW-08	Escritura	19/07/2014	http://es.wikipedia.org/wiki/Escritura
ESW-09	Ingravidez	19/07/2014	http://es.wikipedia.org/wiki/Ingravidez
ESW-10	José Santos de la Hera	19/07/2014	http://es.[...]/José_Santos_de_la_Hera
ESW-11	Juegos Nemeos	19/07/2014	http://es.wikipedia.org/wiki/Juegos_Nemeos
ESW-12	Leucemia	19/07/2014	http://es.wikipedia.org/wiki/Leucemia
ESW-13	Litoral (geografía)	19/07/2014	http://es.wikipedia.org/wiki/Litoral_(geografía)
ESW-14	Materia oscura	19/07/2014	http://es.wikipedia.org/wiki/Materia_oscura
ESW-15	Mente	19/07/2014	http://es.wikipedia.org/wiki/Mente
ESW-16	Miedo	19/07/2014	http://es.wikipedia.org/wiki/Miedo
ESW-17	Mitología	19/07/2014	http://es.wikipedia.org/wiki/Mitología
ESW-18	Moral	19/07/2014	http://es.wikipedia.org/wiki/Moral
ESW-19	Playa	19/07/2014	http://es.wikipedia.org/wiki/Playa
ESW-20	Positivismo	19/07/2014	http://es.wikipedia.org/wiki/Positivismo
ESW-21	Razón	19/07/2014	http://es.wikipedia.org/wiki/Razon
ESW-22	Roca	19/07/2014	http://es.wikipedia.org/wiki/Roca
ESW-23	Romanticismo	19/07/2014	http://es.wikipedia.org/wiki/Romanticismo
ESW-24	Sachapuyos	19/07/2014	http://es.wikipedia.org/wiki/Sachapuyos
ESW-25	Sueño	19/07/2014	http://es.wikipedia.org/wiki/Sueño
		Finlandês	
FIW-01	Antropologia	19/07/2014	http://fi.wikipedia.org/wiki/Antropologia
FIW-02	Aurinkokunta	19/07/2014	http://fi.wikipedia.org/wiki/Aurinkokunta
FIW-03	Baletti	19/07/2014	http://fi.wikipedia.org/wiki/Baletti
FIW-04	Diabetes	19/07/2014	http://fi.wikipedia.org/wiki/Diabetes

FIW-05	Ensimmäinen maailmansota	19/07/2014	http://fi.[...]/Ensimmäinen_maailmansota
FIW-06	Eurooppa	19/07/2014	http://fi.wikipedia.org/wiki/Eurooppa
FIW-07	Fysiikka	19/07/2014	http://fi.wikipedia.org/wiki/Fysiikka
FIW-08	Ihminen	19/07/2014	http://fi.wikipedia.org/wiki/Ihminen
FIW-09	Kirjasintyyppi	19/07/2014	http://fi.wikipedia.org/wiki/Kirjasintyyppi
FIW-10	Klassismin musiikki	19/07/2014	http://fi.[...]/Klassismin_musiikki
FIW-11	Kuolema	19/07/2014	http://fi.wikipedia.org/wiki/Kuolema
FIW-12	Liberalismi	19/07/2014	http://fi.wikipedia.org/wiki/Liberalismo
FIW-13	Liikunta	19/07/2014	http://fi.wikipedia.org/wiki/Liikunta
FIW-14	Meemi	19/07/2014	http://fi.wikipedia.org/wiki/Meemi
FIW-15	Metafysiikka	19/07/2014	http://fi.wikipedia.org/wiki/Metafysiikka
FIW-16	Modernismi	19/07/2014	http://fi.wikipedia.org/wiki/Modernismi
FIW-17	Ohjelmistotuotanto	19/07/2014	http://fi.[...]/Ohjelmistotuotanto
FIW-18	Ooppera	19/07/2014	http://fi.wikipedia.org/wiki/Ooppera
FIW-19	Päättely	19/07/2014	http://fi.wikipedia.org/wiki/Päättely
FIW-20	Rock	19/07/2014	http://fi.wikipedia.org/wiki/Rock
FIW-21	Romaaninen tyyli	19/07/2014	http://fi.[...]/Romaaninen_tyyli
FIW-22	Telenovela	19/07/2014	http://fi.wikipedia.org/wiki/Telenovela
FIW-23	Tietoteoria	19/07/2014	http://fi.wikipedia.org/wiki/Tietoteoria
FIW-24	Tulivuori	19/07/2014	http://fi.wikipedia.org/wiki/Tulivuori
FIW-25	Universaali	19/07/2014	http://fi.wikipedia.org/wiki/Universaali
	Francês		
FRW-01	Communication	19/07/2014	http://fr.wikipedia.org/wiki/Communication
FRW-02	Crime	19/07/2014	http://fr.wikipedia.org/wiki/Crime
FRW-03	Culture	19/07/2014	http://fr.wikipedia.org/wiki/Culture
FRW-04	Église (institution)	19/07/2014	http://fr.wikipedia.org/wiki/Église_(institution)
FRW-05	Espérance (vertu)	19/07/2014	http://fr.wikipedia.org/wiki/Espérance_(vertu)
FRW-06	Fable	19/07/2014	http://fr.wikipedia.org/wiki/Fable
FRW-07	Homme	19/07/2014	http://fr.wikipedia.org/wiki/Homme
FRW-08	Information	19/07/2014	http://fr.wikipedia.org/wiki/Information
FRW-09	Emmanuel Kant	19/07/2014	http://fr.wikipedia.org/wiki/Kant
FRW-10	Morale	19/07/2014	http://fr.wikipedia.org/wiki/Morale
FRW-11	Mort	19/07/2014	http://fr.wikipedia.org/wiki/Mort
FRW-12	Nutrition	19/07/2014	http://fr.wikipedia.org/wiki/Nutrition
FRW-13	Orage	19/07/2014	http://fr.wikipedia.org/wiki/Orage
FRW-14	Publicité	19/07/2014	http://fr.wikipedia.org/wiki/Publicit'e
FRW-15	Raison	19/07/2014	http://fr.wikipedia.org/wiki/Raison
FRW-16	Règne (biologie)	19/07/2014	http://fr.wikipedia.org/wiki/Règne_(biologie)
FRW-17	Réseautage social	19/07/2014	http://fr.[...]/Réseautage_social
FRW-18	Rivière	19/07/2014	http://fr.wikipedia.org/wiki/Rivière
FRW-19	Serveur informatique	19/07/2014	http://fr.[...]/Serveur_informatique
FRW-20	Temps	19/07/2014	http://fr.wikipedia.org/wiki/Temps
FRW-21	Théorie	19/07/2014	http://fr.wikipedia.org/wiki/Théorie
FRW-22	Vérité en Philosophie	19/07/2014	http://fr.[...]/Vérité_en_Philosophie
FRW-23	Vie	19/07/2014	http://fr.wikipedia.org/wiki/Vie
FRW-24	Vieillesse	19/07/2014	http://fr.wikipedia.org/wiki/Vieillesse

FRW-25	Volcan	19/07/2014	http://fr.wikipedia.org/wiki/Volcan
		Húngaro	
HUW-01	Atlanti-óceán	19/07/2014	http://hu.wikipedia.org/wiki/Atlanti-óceán
HUW-02	Autóbusz	19/07/2014	http://hu.wikipedia.org/wiki/Autóbusz
HUW-03	Benzin	19/07/2014	http://hu.wikipedia.org/wiki/Benzin
HUW-04	Civilizáció	19/07/2014	http://hu.wikipedia.org/wiki/Civilizáció
HUW-05	Élet	19/07/2014	http://hu.wikipedia.org/wiki/Élet
HUW-06	Filozófia	19/07/2014	http://hu.wikipedia.org/wiki/Filozófia
HUW-07	Fizika	19/07/2014	http://hu.wikipedia.org/wiki/Fizika
HUW-08	Idő	19/07/2014	http://hu.wikipedia.org/wiki/Idő
HUW-09	José Saramago	19/07/2014	http://hu.wikipedia.org/wiki/José_Saramago
HUW-10	Labdarúgás	19/07/2014	http://hu.wikipedia.org/wiki/Labdarúgás
HUW-11	Mágnesség	19/07/2014	http://hu.wikipedia.org/wiki/Mágnesség
HUW-12	Magyarok	19/07/2014	http://hu.wikipedia.org/wiki/Magyarok
HUW-13	Matematika	19/07/2014	http://hu.wikipedia.org/wiki/Matematika
HUW-14	Mobiltelefon	19/07/2014	http://hu.wikipedia.org/wiki/Mobiltelefon
HUW-15	Művészet	19/07/2014	http://hu.wikipedia.org/wiki/Művészet
HUW-16	Nagyagy	19/07/2014	http://hu.wikipedia.org/wiki/Nagyagy
HUW-17	ONU	19/07/2014	http://hu.wikipedia.org/wiki/ONU
HUW-18	Opera (színmű)	19/07/2014	http://hu.wikipedia.org/wiki/Opera_(színmű)
HUW-19	Ősrobbanás	19/07/2014	http://hu.wikipedia.org/wiki/Ősrobbanás
HUW-20	Szabadság (filozófia)	19/07/2014	http://hu.wikipedia.org/wiki/Szabadság_(filozófia)
HUW-21	Szív	19/07/2014	http://hu.wikipedia.org/wiki/Szív
HUW-22	Szó	19/07/2014	http://hu.wikipedia.org/wiki/Szó
HUW-23	Szociológia	19/07/2014	http://hu.wikipedia.org/wiki/Szociológia
HUW-24	Úszóhólyag	19/07/2014	http://hu.wikipedia.org/wiki/Úszóhólyag
HUW-25	Zongora	19/07/2014	http://hu.wikipedia.org/wiki/Zongora
		Inglês	
INW-01	Aeneid	19/07/2014	http://en.wikipedia.org/wiki/Aeneid
INW-02	Atacama Desert	19/07/2014	http://en.wikipedia.org/wiki/Atacama_Desert
INW-03	Belief	19/07/2014	http://en.wikipedia.org/wiki/Belief
INW-04	Book	19/07/2014	http://en.wikipedia.org/wiki/Book
INW-05	Bourgeoisie	19/07/2014	http://en.wikipedia.org/wiki/Bourgeoisie
INW-06	Charter of Liberties	19/07/2014	http://en.wikipedia.org/wiki/Charter_of_Liberties
INW-07	Civil liberties	19/07/2014	http://en.wikipedia.org/wiki/Civil_liberties
INW-08	Common sense	19/07/2014	http://en.wikipedia.org/wiki/Common_sense
INW-09	Cosmogony	19/07/2014	http://en.wikipedia.org/wiki/Cosmogony
INW-10	Elastic-rebound theory	19/07/2014	http://en.wikipedia.org/wiki/Elastic-rebound_theory
INW-11	Eye	19/07/2014	http://en.wikipedia.org/wiki/Eye
INW-12	Fungicide	19/07/2014	http://en.wikipedia.org/wiki/Fungicide
INW-13	Future	19/07/2014	http://en.wikipedia.org/wiki/Future
INW-14	Globe	19/07/2014	http://en.wikipedia.org/wiki/Globe
INW-15	Idea	19/07/2014	http://en.wikipedia.org/wiki/Idea
INW-16	Learning	19/07/2014	http://en.wikipedia.org/wiki/Learning
INW-17	Levant	19/07/2014	http://en.wikipedia.org/wiki/Levant
INW-18	Lider	19/07/2014	http://en.wikipedia.org/wiki/Lider

INW-19	Paper	19/07/2014	http://en.wikipedia.org/wiki/Paper
INW-20	Plateau	19/07/2014	http://en.wikipedia.org/wiki/Plateau
INW-21	Politics	19/07/2014	http://en.wikipedia.org/wiki/Politics
INW-22	Relief	19/07/2014	http://en.wikipedia.org/wiki/Relief
INW-23	Sand	19/07/2014	http://en.wikipedia.org/wiki/Sand
INW-24	Semiotics	19/07/2014	http://en.wikipedia.org/wiki/Semiotics
INW-25	Tool	19/07/2014	http://en.wikipedia.org/wiki/Tool
		Italiano	
ITW-01	Adolescenza	21/07/2014	http://it.wikipedia.org/wiki/Adolescenza
ITW-02	Balletto	21/07/2014	http://it.wikipedia.org/wiki/Balletto
ITW-03	Biosfera	21/07/2014	http://it.wikipedia.org/wiki/Biosfera
ITW-04	Bullismo	21/07/2014	http://it.wikipedia.org/wiki/Bullismo
ITW-05	Colpa (diritto)	21/07/2014	http://it.wikipedia.org/wiki/Colpa_(diritto)
ITW-06	Conoscenza	21/07/2014	http://it.wikipedia.org/wiki/Conoscenza
ITW-07	Cultura	21/07/2014	http://it.wikipedia.org/wiki/Cultura
ITW-08	Empatia	21/07/2014	http://it.wikipedia.org/wiki/Empatia
ITW-09	Eroe	21/07/2014	http://it.wikipedia.org/wiki/Eroe
ITW-10	Fiume	21/07/2014	http://it.wikipedia.org/wiki/Fiume
ITW-11	Isaac Asimov	21/07/2014	http://it.wikipedia.org/wiki/Isaac_Asimov
ITW-12	Lettura	21/07/2014	http://it.wikipedia.org/wiki/Lettura
ITW-13	Lingua (linguistica)	21/07/2014	http://it.wikipedia.org/wiki/Lingua_(linguistica)
ITW-14	Milizia	21/07/2014	http://it.wikipedia.org/wiki/Milizia
ITW-15	Montagna	21/07/2014	http://it.wikipedia.org/wiki/Montagna
ITW-16	Morte	21/07/2014	http://it.wikipedia.org/wiki/Morte
ITW-17	Prosa	21/07/2014	http://it.wikipedia.org/wiki/Prosa
ITW-18	Rete sociale	21/07/2014	http://it.wikipedia.org/wiki/Rete_sociale
ITW-19	Scienza	21/07/2014	http://it.wikipedia.org/wiki/Scienza
ITW-20	Scuola	21/07/2014	http://it.wikipedia.org/wiki/Scuola
ITW-21	Senilità	21/07/2014	http://it.wikipedia.org/wiki/Senilità
ITW-22	Sociologia	21/07/2014	http://it.wikipedia.org/wiki/Sociologia
ITW-23	Teatro	21/07/2014	http://it.wikipedia.org/wiki/Teatro
ITW-24	Uomo	21/07/2014	http://it.wikipedia.org/wiki/Uomo
ITW-25	Vita	21/07/2014	http://it.wikipedia.org/wiki/Vita
		Português	
PTW-01	Agronomia	18/07/2014	http://pt.wikipedia.org/wiki/Agronomia
PTW-02	Azul	18/07/2014	http://pt.wikipedia.org/wiki/Azul
PTW-03	Banco Mundial	18/07/2014	http://pt.wikipedia.org/wiki/Banco_Mundial
PTW-04	Biosfera	18/07/2014	http://pt.wikipedia.org/wiki/Biosfera
PTW-05	Biotecnologia	18/07/2014	http://pt.wikipedia.org/wiki/Biotecnologia
PTW-06	Cartografia	18/07/2014	http://pt.wikipedia.org/wiki/Cartografia
PTW-07	Cidade	18/07/2014	http://pt.wikipedia.org/wiki/Cidade
PTW-08	Civilização	18/07/2014	http://pt.wikipedia.org/wiki/Civilização
PTW-09	Confiança	18/07/2014	http://pt.wikipedia.org/wiki/Confiança
PTW-10	Conhecimento	18/07/2014	http://pt.wikipedia.org/wiki/Conhecimento
PTW-11	Estatística	18/07/2014	http://pt.wikipedia.org/wiki/Estatística
PTW-12	Fé	18/07/2014	http://pt.wikipedia.org/wiki/Fé
PTW-13	Filosofia	18/07/2014	http://pt.wikipedia.org/wiki/Filosofia

PTW-14	Informação	18/07/2014	http://pt.wikipedia.org/wiki/Informação
PTW-15	Logistica	18/07/2014	http://pt.wikipedia.org/wiki/Logística
PTW-16	Nutrição	18/07/2014	http://pt.wikipedia.org/wiki/Nutrição
PTW-17	Mitologia	18/07/2014	http://pt.wikipedia.org/wiki/Mitologia
PTW-18	Materia	18/07/2014	http://pt.wikipedia.org/wiki/Materia
PTW-19	Oceanografia	18/07/2014	http://pt.wikipedia.org/wiki/Oceanografia
PTW-20	Ostra	18/07/2014	http://pt.wikipedia.org/wiki/Ostra
PTW-21	Política	18/07/2014	http://pt.wikipedia.org/wiki/Política
PTW-22	Rede Social	18/07/2014	http://pt.wikipedia.org/wiki/Rede_social
PTW-23	Safira	18/07/2014	http://pt.wikipedia.org/wiki/Safira
PTW-24	Tempo	18/07/2014	http://pt.wikipedia.org/wiki/Tempo
PTW-25	Vida	18/07/2014	http://pt.wikipedia.org/wiki/Vida
		Sueco	
SEW-01	Afrikas litteratur	19/07/2014	http://sv.wikipedia.org/wiki/Afrikas_litteratur
SEW-02	Demokrati	19/07/2014	http://sv.wikipedia.org/wiki/Demokrati
SEW-03	Exoplanet	19/07/2014	http://sv.wikipedia.org/wiki/Exoplanet
SEW-04	Filosofi	19/07/2014	http://sv.wikipedia.org/wiki/Filosofi
SEW-05	Fortplantning	19/07/2014	http://sv.wikipedia.org/wiki/Fortplantning
SEW-06	Fysik	19/07/2014	http://sv.wikipedia.org/wiki/Fysik
SEW-07	Inbördeskrig	19/07/2014	http://sv.wikipedia.org/wiki/Inbördeskrig
SEW-08	Intelligens	19/07/2014	http://sv.wikipedia.org/wiki/Intelligens
SEW-09	Internet	19/07/2014	http://sv.wikipedia.org/wiki/Internet
SEW-10	Jordbävning	19/07/2014	http://sv.wikipedia.org/wiki/Jordbävning
SEW-11	Jorden	19/07/2014	http://sv.wikipedia.org/wiki/Jorden
SEW-12	Klimat	19/07/2014	http://sv.wikipedia.org/wiki/Klimat
SEW-13	Känsla	19/07/2014	http://sv.wikipedia.org/wiki/Känsla
SEW-14	Konst	19/07/2014	http://sv.wikipedia.org/wiki/Konst
SEW-15	Kunskap	19/07/2014	http://sv.wikipedia.org/wiki/Kunskap
SEW-16	Matfotografi	19/07/2014	http://sv.wikipedia.org/wiki/Matfotografi
SEW-17	Metafysik	19/07/2014	http://sv.wikipedia.org/wiki/Metafysik
SEW-18	Naturvetenskap	19/07/2014	http://sv.wikipedia.org/wiki/Naturvetenskap
SEW-19	Nobelpriset	19/07/2014	http://sv.wikipedia.org/wiki/Nobelpriset
SEW-20	Rationalism	19/07/2014	http://sv.wikipedia.org/wiki/Rationalism
SEW-21	Sociologi	19/07/2014	http://sv.wikipedia.org/wiki/Sociologi
SEW-22	Svensk humor	19/07/2014	http://sv.wikipedia.org/wiki/Svensk_humor
SEW-23	Terrorism	19/07/2014	http://sv.wikipedia.org/wiki/Terrorism
SEW-24	Vetenskap	19/07/2014	http://sv.wikipedia.org/wiki/Vetenskap
SEW-25	Wikipedia	19/07/2014	http://sv.wikipedia.org/wiki/Wikipedia

APÊNDICE C

RESULTADOS DO CORPUS

Para todos os textos utilizados, apresentamos os resultados dos seguintes parâmetros:

- T: Número de palavras;
- V: Número de verbetes;
- D: Diversidade (T/V);
- η: Fração de verbetes com desvio padrão $\sigma > 1$;
- $\bar{\sigma}$: Desvio padrão médio;
- γ: Fração de verbetes limitados pelas curvas dos modelos geométrico e de números primos;
- k_{max}: Frequência máxima;
- $\overline{H}$: Entropia estrutural;
- θ: Fração de verbetes com frequência superior a k_0;
- $\bar{l}_v$: Comprimento médio dos verbetes;
- $\bar{l}_{vr}$: Comprimento médio dos verbetes com desvio padrão $\sigma > 1$.

Referência	T	V	D	η	σ	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
DEW-01	4317	1532	0.35	0.27	0.83	0.14	173	1.50	0.07	8.94	6.05
DEW-02	2857	1303	0.46	0.30	0.83	0.10	108	1.47	0.07	8.76	5.25
DEW-03	3270	1353	0.41	0.34	0.90	0.09	107	1.52	0.06	8.67	5.66
DEW-04	1101	589	0.53	0.15	0.74	0.11	35	1.41	0.03	7.58	4.00
DEW-05	5824	397	0.48	0.24	0.79	0.07	41	1.31	0.10	8.01	4.71
DEW-06	2153	987	0.46	0.24	0.78	0.16	98	1.47	0.08	9.55	5.75
DEW-07	2361	1073	0.45	0.26	0.79	0.14	103	1.54	0.13	9.05	4.48
DEW-08	5708	2265	0.40	0.25	0.80	0.12	182	1.49	0.06	8.25	5.64
DEW-09	862	408	0.47	0.18	0.73	0.15	59	1.48	0.00	7.54	4.42
DEW-10	2143	1000	0.47	0.26	0.80	0.13	80	1.47	0.00	8.04	4.50
DEW-11	2372	975	0.41	0.30	0.83	0.14	135	1.44	0.10	8.64	5.50
DEW-12	953	498	0.52	0.14	0.76	0.09	36	1.53	0.00	8.45	3.88
DEW-13	3069	1337	0.44	0.27	0.80	0.16	137	1.44	0.08	8.45	5.17
DEW-14	5906	2293	0.39	0.31	0.85	0.13	223	1.52	0.11	8.93	5.31
DEW-15	630	352	0.56	0.21	0.71	0.15	85	1.38	0.00	8.61	4.12
DEW-16	1510	688	0.46	0.30	0.84	0.11	70	1.54	0.09	8.42	4.19
DEW-17	1995	899	0.45	0.27	0.81	0.12	90	1.50	0.00	9.36	6.24
DEW-18	10307	3269	0.32	0.29	0.84	0.11	488	1.49	0.16	9.12	6.96
DEW-19	736	424	0.58	0.07	0.68	0.09	127	1.26	0.00	7.43	5.50
DEW-20	717	393	0.55	0.12	0.65	0.22	127	1.27	0.00	8.40	6.00
DEW-21	2583	1135	0.44	0.33	0.84	0.19	129	1.53	0.00	8.79	4.71
DEW-22	1315	594	0.45	0.19	0.75	0.15	55	1.51	0.00	8.31	5.53
DEW-23	2601	1076	0.41	0.27	0.81	0.11	135	1.52	0.00	8.10	4.54
DEW-24	2528	1141	0.45	0.28	0.82	0.12	93	1.53	0.00	8.40	4.86
DEW-25	4024	1515	0.38	0.24	0.81	0.12	205	1.40	0.15	8.45	5.68
DE-01	39730	7290	0.18	0.29	0.84	0.10	1170	1.70	0.18	8.19	5.56
DE-02	54376	5657	0.10	0.43	0.98	0.08	2442	2.00	0.27	7.64	5.36
DE-03	48941	8670	0.18	0.34	0.88	0.11	1396	1.71	0.26	8.48	6.05
DE-04	61637	6868	0.11	0.34	0.87	0.10	3048	1.88	0.36	7.70	5.56
DE-05	58324	10666	0.18	0.38	0.95	0.09	1958	1.71	0.19	9.15	6.80
DE-06	66957	10999	0.16	0.33	0.87	0.11	2917	1.75	0.29	8.15	5.66
DE-07	69442	8604	0.12	0.34	0.90	0.09	2880	1.83	0.30	8.34	5.92
DE-08	67543	11193	0.17	0.30	0.84	0.11	2418	1.71	0.30	8.82	6.19
DE-09	68509	10170	0.15	0.32	0.88	0.10	2549	1.76	0.24	8.57	6.32
DE-10	66937	10327	0.15	0.36	0.92	0.10	2409	1.76	0.26	8.59	6.08
DE-11	62951	10796	0.17	0.36	0.91	0.09	2297	1.74	0.29	9.35	7.36
DE-12	67087	12407	0.18	0.33	0.90	0.10	2337	1.65	0.21	8.63	6.31
DE-13	69642	11771	0.17	0.28	0.83	0.12	2412	1.71	0.27	8.66	6.02
DE-14	68852	13056	0.19	0.33	0.88	0.11	2239	1.69	0.32	8.99	6.15
DE-15	70902	12299	0.17	0.34	0.89	0.10	2066	1.71	0.28	8.90	6.17
DE-16	71618	11245	0.16	0.35	0.91	0.10	2454	1.71	0.30	8.81	6.46
DE-17	72963	10166	0.14	0.37	0.92	0.09	2440	1.78	0.29	8.82	6.79
DE-18	78991	10889	0.14	0.30	0.84	0.11	2591	1.79	0.29	8.80	6.52
DE-19	80160	12409	0.15	0.39	0.96	0.08	4247	1.72	0.30	9.29	7.18
DE-20	80367	11580	0.14	0.34	0.89	0.10	2389	1.78	0.31	8.51	6.07
DE-21	85565	13177	0.15	0.26	0.80	0.12	2851	1.75	0.31	8.78	6.54
DE-22	95274	11157	0.12	0.34	0.87	0.10	4088	1.88	0.36	8.78	6.20
DE-23	102056	10107	0.10	0.32	0.85	0.10	3889	1.89	0.33	9.04	6.80
DE-24	104800	15106	0.14	0.34	0.90	0.09	4371	1.77	0.31	8.67	6.84
DE-25	119360	11654	0.10	0.39	0.94	0.08	3458	1.93	0.29	7.72	5.82

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
DKW-01	4213	1593	0.38	0.34	0.89	0.14	182	1.43	0.08	7.49	5.80
DKW-02	5905	1929	0.33	0.35	0.89	0.10	229	1.50	0.15	7.53	5.64
DKW-03	4861	1522	0.31	0.31	0.83	0.13	177	1.47	0.20	7.96	5.75
DKW-04	4146	1489	0.36	0.34	0.85	0.12	129	1.52	0.00	7.36	5.19
DKW-05	1279	607	0.47	0.23	0.75	0.14	69	1.42	0.05	7.72	5.68
DKW-06	4379	1356	0.31	0.38	0.93	0.10	187	1.48	0.12	7.08	5.15
DKW-07	898	409	0.46	0.12	0.72	0.14	59	1.43	0.00	6.69	3.00
DKW-08	2082	827	0.40	0.26	0.79	0.13	72	1.40	0.11	7.20	4.16
DKW-09	10649	3304	0.31	0.28	0.84	0.12	440	1.53	0.13	7.41	5.18
DKW-10	974	461	0.47	0.15	0.73	0.09	163	1.35	0.00	6.96	4.18
DKW-11	2778	1046	0.38	0.25	0.80	0.14	114	1.53	0.10	7.63	5.49
DKW-12	802	425	0.53	0.22	0.78	0.08	75	1.44	0.00	7.26	2.91
DKW-13	1458	668	0.46	0.18	0.77	0.15	55	1.57	0.00	7.79	3.85
DKW-14	1179	592	0.50	0.12	0.81	0.07	52	1.41	0.04	7.22	4.89
DKW-15	5932	1840	0.31	0.31	0.88	0.12	181	1.51	0.12	7.39	5.05
DKW-16	1936	746	0.39	0.25	0.83	0.09	74	1.51	0.00	7.36	4.62
DKW-17	6212	1963	0.32	0.24	0.80	0.11	223	1.49	0.17	7.00	5.47
DKW-18	900	416	0.46	0.24	0.74	0.14	106	1.45	0.00	7.36	5.00
DKW-19	1735	663	0.38	0.27	0.84	0.14	57	1.49	0.00	7.43	6.72
DKW-20	4408	1469	0.33	0.30	0.87	0.11	162	1.49	0.12	7.01	4.95
DKW-21	2393	808	0.34	0.34	0.89	0.07	102	1.56	0.04	7.36	5.04
DKW-22	4991	1705	0.34	0.36	0.90	0.12	194	1.53	0.14	7.55	5.24
DKW-23	587	340	0.58	0.11	0.67	0.18	63	1.44	0.00	7.16	5.25
DKW-24	3198	1270	0.40	0.27	0.81	0.12	110	1.44	0.06	7.33	5.11
DKW-25	1907	825	0.43	0.14	0.76	0.11	84	1.43	0.04	7.69	3.75
DK-01	26642	5650	0.21	0.28	0.85	0.10	960	1.70	0.24	7.26	5.01
DK-02	34493	6091	0.18	0.33	0.87	0.10	1316	1.69	0.28	7.40	5.26
DK-03	39690	5239	0.13	0.41	0.97	0.08	1692	1.87	0.33	7.13	5.36
DK-04	42879	5921	0.14	0.31	0.86	0.09	1753	1.76	0.32	7.22	5.48
DK-05	45115	5523	0.12	0.42	0.96	0.07	2163	1.85	0.32	7.11	5.43
DK-06	47401	6703	0.14	0.31	0.86	0.10	2206	1.80	0.37	7.39	5.05
DK-07	49595	5982	0.12	0.31	0.87	0.10	1747	1.82	0.41	7.30	5.13
DK-08	51909	6807	0.13	0.41	0.94	0.09	2430	1.83	0.35	7.25	5.30
DK-09	58292	7209	0.12	0.33	0.89	0.10	2201	1.80	0.40	7.20	5.21
DK-10	58666	8057	0.14	0.32	0.88	0.10	2186	1.77	0.38	7.61	5.39
DK-11	64240	8160	0.13	0.41	0.97	0.08	2391	1.80	0.35	7.43	5.75
DK-12	66992	10926	0.16	0.35	0.91	0.09	2274	1.70	0.34	8.19	6.17
DK-13	67781	9075	0.13	0.37	0.92	0.09	3569	1.79	0.35	7.44	5.57
DK-14	69991	9840	0.14	0.32	0.88	0.11	2975	1.76	0.35	7.24	5.31
DK-15	72506	7948	0.11	0.35	0.89	0.09	2214	1.83	0.36	7.72	5.89
DK-16	73918	11503	0.16	0.35	0.90	0.09	2921	1.75	0.28	7.77	5.80
DK-17	74278	11784	0.16	0.35	0.90	0.10	3966	1.72	0.35	7.50	5.20
DK-18	75178	7348	0.10	0.41	0.97	0.08	3199	1.95	0.39	7.39	5.71
DK-19	75439	9869	0.13	0.39	0.95	0.09	3303	1.82	0.33	8.03	5.81
DK-20	75859	9731	0.13	0.34	0.90	0.09	3063	1.80	0.36	7.99	5.87
DK-21	77898	6809	0.09	0.39	0.93	0.08	3047	1.91	0.39	7.34	5.80
DK-22	82014	8056	0.10	0.41	0.96	0.09	2777	1.88	0.35	7.47	5.78
DK-23	82681	8807	0.11	0.40	0.96	0.08	3128	1.88	0.38	7.57	5.73
DK-24	89397	11417	0.13	0.30	0.84	0.10	2662	1.79	0.39	8.00	5.64
DK-25	97863	14517	0.15	0.33	0.89	0.10	2910	1.72	0.37	8.17	5.77

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
ESW-01	1464	522	0.36	0.24	0.87	0.05	106	1.54	0.27	6.81	3.65
ESW-02	1048	458	0.44	0.18	0.71	0.19	56	1.47	0.00	7.21	5.27
ESW-03	1196	517	0.43	0.27	0.80	0.08	86	1.60	0.13	7.43	4.38
ESW-04	2566	760	0.30	0.26	0.83	0.10	253	1.67	0.15	7.34	5.53
ESW-05	585	297	0.51	0.26	0.80	0.11	45	1.61	0.00	7.01	2.00
ESW-06	2993	955	0.32	0.31	0.86	0.11	217	1.57	0.15	7.27	5.00
ESW-07	3055	1062	0.35	0.28	0.81	0.10	188	1.42	0.14	7.56	4.72
ESW-08	3483	1198	0.34	0.33	0.90	0.11	182	1.46	0.16	7.70	5.91
ESW-09	1308	492	0.38	0.06	0.61	0.22	122	1.44	0.15	6.85	5.80
ESW-10	1316	485	0.37	0.27	0.82	0.09	78	1.65	0.12	6.99	4.35
ESW-11	783	394	0.50	0.16	0.72	0.18	46	1.50	0.06	6.41	2.75
ESW-12	1842	718	0.39	0.22	0.81	0.07	144	1.44	0.13	7.42	5.20
ESW-13	1314	528	0.40	0.16	0.75	0.11	111	1.43	0.13	7.66	6.31
ESW-14	5394	1431	0.27	0.25	0.79	0.19	453	1.44	0.23	7.55	5.84
ESW-15	2507	934	0.37	0.34	0.87	0.11	165	1.55	0.13	8.08	5.86
ESW-16	3853	1396	0.36	0.26	0.81	0.13	213	1.49	0.18	7.43	5.40
ESW-17	2127	811	0.38	0.23	0.79	0.12	151	1.42	0.18	7.26	4.46
ESW-18	2554	907	0.36	0.25	0.82	0.12	170	1.54	0.16	7.36	5.34
ESW-19	846	350	0.41	0.14	0.77	0.10	55	1.56	0.00	6.65	2.86
ESW-20	987	411	0.42	0.17	0.72	0.16	69	1.46	0.00	7.70	4.82
ESW-21	1729	633	0.37	0.31	0.89	0.07	103	1.63	0.12	7.50	5.40
ESW-22	1459	544	0.37	0.24	0.78	0.15	90	1.45	0.10	7.60	6.05
ESW-23	1700	736	0.43	0.22	0.78	0.15	99	1.58	0.16	7.35	3.44
ESW-24	619	326	0.53	0.05	0.65	0.15	45	1.41	0.00	6.61	2.50
ESW-25	2846	927	0.33	0.26	0.81	0.10	204	1.43	0.25	7.28	4.71
ES-01	70955	8961	0.13	0.33	0.87	0.10	3691	1.78	0.35	7.81	6.26
ES-02	71720	11599	0.16	0.30	0.86	0.11	3676	1.61	0.39	7.77	5.81
ES-03	72229	10587	0.15	0.26	0.79	0.13	4448	1.67	0.35	8.03	6.01
ES-04	74175	12022	0.16	0.35	0.89	0.10	5304	1.64	0.38	7.85	6.59
ES-05	75454	11300	0.15	0.31	0.87	0.10	3653	1.68	0.32	7.92	6.36
ES-06	76618	11806	0.15	0.34	0.89	0.10	3747	1.69	0.35	7.82	6.06
ES-07	77022	9528	0.12	0.36	0.91	0.10	4040	1.78	0.36	7.83	6.21
ES-08	80329	13098	0.16	0.32	0.87	0.10	4352	1.64	0.33	8.02	6.42
ES-09	81712	11969	0.15	0.31	0.87	0.11	4727	1.64	0.38	7.93	6.34
ES-10	83078	12024	0.14	0.31	0.86	0.11	4179	1.69	0.37	7.70	5.82
ES-11	83751	15035	0.18	0.25	0.81	0.13	4633	1.58	0.38	7.87	5.80
ES-12	84683	11477	0.14	0.38	0.96	0.08	4641	1.74	0.40	7.72	6.22
ES-13	86100	10473	0.12	0.34	0.89	0.09	5244	1.80	0.36	8.01	6.53
ES-14	87591	12659	0.14	0.31	0.87	0.11	4482	1.68	0.33	7.99	6.42
ES-15	88470	14424	0.16	0.27	0.81	0.12	5567	1.61	0.36	7.78	6.16
ES-16	88774	11292	0.13	0.32	0.87	0.09	4182	1.76	0.40	7.86	6.10
ES-17	89282	12403	0.14	0.33	0.86	0.10	4498	1.71	0.37	7.61	5.99
ES-18	89348	11696	0.13	0.33	0.89	0.09	5145	1.71	0.39	7.84	6.17
ES-19	89540	11067	0.12	0.33	0.87	0.10	4271	1.78	0.39	7.87	6.28
ES-20	90990	13636	0.15	0.30	0.85	0.11	5358	1.62	0.39	7.99	6.38
ES-21	92092	12883	0.14	0.36	0.92	0.09	4464	1.72	0.38	7.60	5.96
ES-22	93429	12698	0.14	0.30	0.85	0.11	5755	1.68	0.37	8.04	6.42
ES-23	97432	12687	0.13	0.32	0.89	0.10	5536	1.75	0.36	7.71	6.20
ES-24	139928	18455	0.13	0.35	0.91	0.10	8243	1.70	0.38	7.94	6.51
ES-25	380247	22909	0.06	0.41	0.97	0.08	20586	1.98	0.47	7.83	6.65

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
FIW-01	937	647	0.69	0.18	0.75	0.18	44	1,22	0.00	9.14	6.62
FIW-02	1087	693	0.64	0.15	0.72	0.16	43	1,08	0.00	8.49	5.25
FIW-03	4088	2504	0.61	0.17	0.73	0.16	174	1,22	0.00	8.95	6.47
FIW-04	1420	865	0.61	0.22	0.79	0.17	39	1,25	0.00	8.90	6.58
FIW-05	2679	1589	0.59	0.26	0.78	0.13	125	1,29	0.00	9.16	7.26
FIW-06	1534	1012	0.66	0.18	0.69	0.22	71	1,16	0.00	8.80	6.87
FIW-07	4726	2572	0.54	0.31	0.90	0.11	184	1,28	0.04	9.62	8.08
FIW-08	3964	2234	0.56	0.25	0.82	0.13	166	1,26	0.04	9.11	5.78
FIW-09	462	321	0.69	0.12	0.65	0.15	43	1,18	0.00	9.68	6.00
FIW-10	4904	2609	0.53	0.25	0.80	0.13	219	1,31	0.04	9.23	6.73
FIW-11	1363	874	0.64	0.17	0.71	0.17	51	1,16	0.00	8.54	6.38
FIW-12	1512	969	0.64	0.18	0.76	0.07	64	1,20	0.00	9.29	4.60
FIW-13	1469	933	0.64	0.21	0.79	0.09	100	1,16	0.07	9.82	6.00
FIW-14	1144	755	0.66	0.24	0.72	0.19	48	1,27	0.04	8.64	5.71
FIW-15	2231	1216	0.55	0.27	0.83	0.09	111	1,29	0.00	8.80	6.69
FIW-16	1123	792	0.71	0.17	0.67	0.19	54	1,18	0.00	8.82	4.56
FIW-17	2867	1594	0.56	0.22	0.77	0.15	101	1,32	0.00	9.82	7.55
FIW-18	745	562	0.75	0.07	0.64	0.18	33	1,12	0.00	8.95	3.50
FIW-19	1090	631	0.58	0.12	0.68	0.17	27	1,18	0.00	9.46	6.60
FIW-20	1048	700	0.67	0.20	0.76	0.10	46	1,30	0.00	8.28	6.00
FIW-21	658	505	0.77	0.04	0.60	0.17	30	1,10	0.00	8.80	2.00
FIW-22	632	428	0.68	0.09	0.55	0.32	29	1,10	0.00	8.30	5.00
FIW-23	1893	1070	0.57	0.23	0.74	0.26	67	1,22	0.00	8.71	5.84
FIW-24	1732	1065	0.61	0.25	0.76	0.15	61	1,16	0.00	8.51	6.04
FIW-25	653	420	0.64	0.13	0.72	0.13	27	1,21	0.00	8.52	6.83
FI-01	26534	9709	0.37	0.28	0.83	0.12	923	1,54	0.06	8.48	6.11
FI-02	29078	9239	0.32	0.28	0.84	0.11	1047	1,57	0.14	8.79	6.27
FI-03	51457	12857	0.25	0.32	0.88	0.11	2465	1,65	0.20	8.61	6.40
FI-04	53280	13644	0.26	0.35	0.91	0.09	2658	1,63	0.20	8.58	6.35
FI-05	56677	15290	0.27	0.33	0.88	0.10	1793	1,63	0.16	8.62	6.30
FI-06	57655	16174	0.28	0.30	0.86	0.10	1508	1,58	0.10	9.75	7.39
FI-07	59942	12276	0.20	0.39	0.94	0.08	2151	1,7	0.22	8.38	6.49
FI-08	60757	16380	0.27	0.31	0.86	0.11	1914	1,61	0.15	9.08	6.70
FI-09	62792	16885	0.27	0.31	0.85	0.12	2561	1,55	0.18	9.03	6.86
FI-10	64905	18672	0.29	0.31	0.86	0.11	2593	1,58	0.11	9.29	6.93
FI-11	67138	20070	0.30	0.29	0.84	0.11	2801	1,57	0.14	9.32	6.89
FI-12	67835	17033	0.25	0.34	0.88	0.10	2857	1,63	0.16	8.77	6.41
FI-13	68036	18102	0.27	0.32	0.89	0.11	3142	1,58	0.17	9.21	6.75
FI-14	68644	15873	0.23	0.32	0.87	0.10	3061	1,68	0.14	8.63	6.53
FI-15	70289	18333	0.26	0.32	0.88	0.11	2026	1,62	0.15	9.12	6.72
FI-16	70543	20049	0.28	0.32	0.86	0.12	2739	1,59	0.15	9.61	7.01
FI-17	71886	19444	0.27	0.31	0.87	0.10	3801	1,63	0.17	9.19	6.66
FI-18	72073	16632	0.23	0.29	0.84	0.11	2819	1,71	0.15	8.93	6.70
FI-19	73275	19006	0.26	0.30	0.85	0.10	3588	1,61	0.18	9.20	6.83
FI-20	73836	16887	0.23	0.28	0.82	0.12	2455	1,66	0.22	9.23	6.56
FI-21	76319	17510	0.23	0.33	0.88	0.10	3444	1,65	0.18	8.96	6.91
FI-22	76915	20697	0.27	0.32	0.90	0.11	3373	1,6	0.16	9.29	6.87
FI-23	81218	21947	0.27	0.33	0.89	0.10	4253	1,56	0.12	8.77	6.55
FI-24	87043	17149	0.20	0.30	0.85	0.10	3509	1,74	0.26	8.99	6.55
FI-25	89018	23128	0.26	0.32	0.87	0.10	4434	1,61	0.16	8.60	6.41

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
FRW-01	5563	1777	0.32	0.35	0.91	0.11	360	1.50	0.14	7.80	5.90
FRW-02	2135	736	0.34	0.30	0.85	0.16	107	1.48	0.14	7.30	5.35
FRW-03	7969	2197	0.28	0.32	0.88	0.11	441	1.56	0.26	7.97	6.29
FRW-04	981	472	0.48	0.21	0.74	0.16	139	1.41	0.00	7.09	4.46
FRW-05	1419	544	0.38	0.30	0.84	0.12	67	1.53	0.00	6.35	3.97
FRW-06	7457	2336	0.31	0.31	0.85	0.09	344	1.55	0.18	7.19	5.07
FRW-07	692	321	0.46	0.27	0.74	0.17	93	1.33	0.00	6.38	4.06
FRW-08	2520	1028	0.41	0.26	0.84	0.11	149	1.52	0.10	7.67	6.26
FRW-09	3650	1186	0.32	0.33	0.87	0.11	243	1.52	0.13	7.41	4.79
FRW-10	2812	1010	0.36	0.32	0.85	0.11	150	1.47	0.16	7.48	5.09
FRW-11	4538	1447	0.32	0.34	0.89	0.08	238	1.55	0.18	7.27	4.52
FRW-12	3440	1189	0.35	0.31	0.85	0.12	145	1.51	0.14	7.40	5.65
FRW-13	4968	1345	0.27	0.27	0.84	0.09	290	1.54	0.09	7.39	5.40
FRW-14	7444	2431	0.33	0.28	0.85	0.12	432	1.48	0.24	7.73	6.01
FRW-15	2294	824	0.36	0.31	0.85	0.12	131	1.56	0.13	7.37	5.04
FRW-16	2617	908	0.35	0.40	0.95	0.10	174	1.64	0.07	7.45	6.47
FRW-17	2222	860	0.39	0.29	0.81	0.10	151	1.52	0.07	7.54	5.03
FRW-18	1079	464	0.43	0.24	0.82	0.09	43	1.64	0.00	6.19	2.65
FRW-19	5754	1513	0.26	0.34	0.91	0.09	351	1.56	0.20	7.52	6.35
FRW-20	10335	2385	0.23	0.33	0.89	0.08	456	1.58	0.23	7.69	5.63
FRW-21	1265	494	0.39	0.22	0.78	0.10	75	1.54	0.00	6.02	3.50
FRW-22	2974	1048	0.35	0.30	0.83	0.14	158	1.48	0.09	7.52	5.03
FRW-23	834	424	0.51	0.12	0.73	0.14	50	1.37	0.00	6.87	4.57
FRW-24	3451	1296	0.38	0.29	0.86	0.12	209	1.45	0.17	7.92	4.98
FRW-25	8030	2194	0.27	0.35	0.89	0.11	515	1.55	0.21	7.54	5.90
FR-01	67525	9728	0.14	0.29	0.83	0.12	3050	1.72	0.35	7.69	5.78
FR-02	67957	9310	0.14	0.28	0.82	0.12	3312	1.73	0.37	7.87	6.10
FR-03	68984	11769	0.17	0.27	0.82	0.12	3050	1.63	0.31	7.87	5.96
FR-04	69225	11046	0.16	0.29	0.83	0.12	3223	1.68	0.28	7.96	5.87
FR-05	70235	10136	0.14	0.34	0.89	0.10	2924	1.75	0.33	7.73	6.00
FR-06	70722	8427	0.12	0.35	0.89	0.10	3306	1.80	0.31	7.92	6.25
FR-07	70880	8234	0.12	0.34	0.89	0.09	2948	1.83	0.33	7.84	6.26
FR-08	71883	8706	0.12	0.36	0.90	0.10	2959	1.83	0.30	7.72	6.04
FR-09	74120	7141	0.10	0.45	1.01	0.07	4540	1.95	0.34	7.37	6.16
FR-10	75151	12011	0.16	0.25	0.80	0.13	3786	1.66	0.33	7.83	6.02
FR-11	76086	9537	0.13	0.28	0.84	0.11	3161	1.76	0.34	7.97	6.18
FR-12	76315	10013	0.13	0.37	0.95	0.08	3448	1.76	0.32	7.79	6.32
FR-13	77365	9906	0.13	0.42	0.99	0.08	4172	1.75	0.32	7.82	6.58
FR-14	78339	10377	0.13	0.33	0.89	0.10	3041	1.75	0.34	7.94	6.51
FR-15	78438	11070	0.14	0.27	0.82	0.12	3339	1.73	0.32	7.83	6.09
FR-16	81026	11326	0.14	0.30	0.85	0.10	3642	1.73	0.30	7.84	6.17
FR-17	81479	11407	0.14	0.28	0.84	0.11	3756	1.73	0.30	7.99	6.25
FR-18	82088	10282	0.13	0.32	0.87	0.10	3624	1.79	0.35	7.67	6.09
FR-19	86015	12830	0.15	0.31	0.86	0.12	3869	1.68	0.33	7.87	6.13
FR-20	88374	13126	0.15	0.25	0.78	0.14	3798	1.63	0.33	8.15	6.25
FR-21	88513	12940	0.15	0.26	0.81	0.13	4169	1.68	0.38	8.14	6.35
FR-22	89002	10821	0.12	0.27	0.81	0.11	4586	1.78	0.38	7.98	6.17
FR-23	111177	14559	0.13	0.25	0.79	0.12	4406	1.73	0.35	8.10	6.36
FR-24	163648	13695	0.08	0.36	0.93	0.09	7769	1.88	0.37	7.97	6.51
FR-25	164529	15350	0.09	0.31	0.86	0.11	7232	1.86	0.37	8.12	6.68

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
HUW-01	1015	630	0.62	0.22	0.71	0.18	99	1.22	0.10	7.49	3.80
HUW-02	1652	845	0.51	0.23	0.77	0.13	146	1.37	0.09	7.78	6.26
HUW-03	805	516	0.64	0.18	0.75	0.18	72	1.25	0.00	7.49	3.83
HUW-04	2963	1447	0.49	0.26	0.80	0.16	295	1.30	0.10	8.08	5.57
HUW-05	4131	1610	0.39	0.21	0.83	0.07	301	1.44	0.16	8.78	6.26
HUW-06	5105	2264	0.44	0.28	0.85	0.11	551	1.32	0.15	8.43	6.51
HUW-07	1424	808	0.57	0.22	0.80	0.12	147	1.28	0.10	8.26	6.33
HUW-08	7703	3143	0.41	0.28	0.84	0.13	594	1.40	0.17	8.35	5.12
HUW-09	791	547	0.69	0.13	0.61	0.26	54	1.36	0.00	7.37	5.75
HUW-10	10978	3786	0.34	0.39	0.97	0.09	1337	1.47	0.15	8.22	6.07
HUW-11	505	304	0.60	0.15	0.60	0.23	322	1.10	0.00	6.82	3.00
HUW-12	5899	2655	0.45	0.29	0.81	0.15	641	1.30	0.14	7.73	5.01
HUW-13	2626	1350	0.51	0.22	0.73	0.17	236	1.25	0.09	8.81	6.20
HUW-14	1414	883	0.62	0.18	0.78	0.12	131	1.24	0.00	7.68	3.17
HUW-15	1812	898	0.50	0.19	0.74	0.14	164	1.25	0.00	7.71	4.67
HUW-16	2381	1080	0.45	0.26	0.82	0.11	231	1.29	0.10	8.12	5.26
HUW-17	2690	1315	0.49	0.24	0.78	0.14	244	1.29	0.09	7.84	5.90
HUW-18	6319	2804	0.44	0.23	0.77	0.15	509	1.37	0.14	7.43	5.19
HUW-19	1886	955	0.51	0.21	0.74	0.16	175	1.28	0.14	7.95	5.36
HUW-20	1315	646	0.49	0.18	0.75	0.18	112	1.28	0.09	7.63	5.50
HUW-21	2141	1017	0.48	0.21	0.73	0.20	266	1.19	0.12	7.61	4.30
HUW-22	2305	1123	0.49	0.27	0.80	0.11	221	1.39	0.12	7.98	4.92
HUW-23	3808	1951	0.51	0.24	0.79	0.12	361	1.29	0.12	8.51	5.45
HUW-24	1099	604	0.55	0.14	0.67	0.25	120	1.27	0.11	7.67	5.62
HUW-25	2649	1347	0.51	0.19	0.80	0.08	248	1.37	0.12	7.71	5.34
HU-01	11236	4024	0.36	0.30	0.86	0.12	781	1.51	0.14	7.50	4.70
HU-02	15766	5881	0.37	0.21	0.78	0.13	1358	1.50	0.17	7.81	5.15
HU-03	19485	6774	0.35	0.25	0.81	0.11	1528	1.50	0.20	7.79	5.05
HU-04	20006	7417	0.37	0.24	0.80	0.13	2081	1.50	0.19	8.03	5.40
HU-05	20289	7370	0.36	0.28	0.83	0.12	1262	1.50	0.14	7.95	5.52
HU-06	23812	8567	0.36	0.28	0.83	0.12	1533	1.54	0.16	7.72	5.18
HU-07	24275	9753	0.40	0.23	0.80	0.12	1962	1.44	0.16	8.39	5.22
HU-08	25392	8087	0.32	0.32	0.86	0.10	2067	1.56	0.15	7.76	5.35
HU-09	25414	8543	0.34	0.29	0.86	0.10	1673	1.54	0.13	7.68	5.05
HU-10	25579	8337	0.33	0.34	0.91	0.10	1967	1.51	0.17	7.75	5.37
HU-11	26172	7368	0.28	0.34	0.89	0.10	1467	1.61	0.15	8.31	6.26
HU-12	31039	9212	0.30	0.29	0.85	0.11	2759	1.59	0.23	8.17	5.68
HU-13	39067	11441	0.29	0.33	0.91	0.09	3515	1.59	0.20	7.74	5.24
HU-14	42261	11663	0.28	0.29	0.85	0.11	3308	1.61	0.21	7.97	5.37
HU-15	45491	8673	0.19	0.44	1.03	0.09	3699	1.66	0.23	7.80	5.90
HU-16	45792	14309	0.31	0.30	0.86	0.11	4058	1.53	0.18	7.95	5.34
HU-17	48445	13145	0.27	0.29	0.85	0.11	3660	1.60	0.18	8.39	5.92
HU-18	50315	8372	0.17	0.40	0.99	0.11	4405	1.70	0.25	7.66	5.93
HU-19	53261	12113	0.23	0.33	0.89	0.10	3237	1.66	0.20	8.24	6.06
HU-20	54098	15260	0.28	0.29	0.84	0.11	3234	1.58	0.20	8.15	5.96
HU-21	60924	10277	0.17	0.41	0.99	0.10	5939	1.71	0.24	7.77	5.96
HU-22	71823	16234	0.23	0.31	0.86	0.10	4766	1.64	0.20	8.43	6.16
HU-23	72647	17169	0.24	0.37	0.94	0.09	4572	1.64	0.22	8.30	6.17
HU-24	72946	18058	0.25	0.33	0.90	0.10	6785	1.62	0.25	8.39	5.85
HU-25	75271	13651	0.18	0.27	0.82	0.11	3411	1.72	0.22	8.37	6.35

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
INW-01	6550	2017	0.31	0.35	0.90	0.11	509	1.50	0.18	6.43	4.66
INW-02	2671	1018	0.38	0.24	0.79	0.13	251	1.40	0.19	6.22	4.69
INW-03	3115	969	0.31	0.37	0.93	0.11	143	1.56	0.19	6.87	5.00
INW-04	6603	1876	0.28	0.34	0.92	0.11	414	1.53	0.22	6.74	5.59
INW-05	3914	1282	0.33	0.26	0.84	0.12	444	1.38	0.20	7.14	5.38
INW-06	2147	736	0.34	0.23	0.80	0.11	174	1.55	0.17	6.20	4.19
INW-07	1971	690	0.35	0.28	0.84	0.10	144	1.44	0.13	6.90	5.61
INW-08	8958	1994	0.22	0.33	0.89	0.09	502	1.65	0.23	7.07	6.13
INW-09	806	344	0.43	0.12	0.75	0.11	87	1.30	0.00	6.67	4.62
INW-10	469	221	0.47	0.11	0.72	0.08	56	1.27	0.12	5.73	5.75
INW-11	5548	1416	0.26	0.34	0.90	0.09	405	1.54	0.19	6.88	5.60
INW-12	1297	506	0.39	0.22	0.75	0.22	80	1.43	0.00	6.58	4.70
INW-13	2980	1023	0.34	0.35	0.90	0.14	221	1.41	0.19	7.11	5.35
INW-14	1740	704	0.40	0.34	0.84	0.14	165	1.48	0.13	6.24	4.31
INW-15	3029	1019	0.34	0.25	0.81	0.08	157	1.48	0.10	6.89	4.31
INW-16	3928	1222	0.31	0.29	0.87	0.12	191	1.46	0.20	7.06	5.27
INW-17	960	426	0.44	0.26	0.79	0.06	99	1.46	0.00	6.36	5.64
INW-18	3106	1089	0.35	0.32	0.89	0.10	310	1.38	0.18	6.83	5.42
INW-19	2939	1024	0.35	0.33	0.91	0.12	158	1.42	0.14	6.43	4.92
INW-20	979	426	0.44	0.21	0.77	0.11	84	1.36	0.09	5.86	3.75
INW-21	4901	1449	0.30	0.30	0.85	0.13	465	1.46	0.21	7.08	5.51
INW-22	2840	893	0.31	0.28	0.82	0.11	191	1.55	0.15	6.63	5.28
INW-23	1583	688	0.43	0.26	0.80	0.18	83	1.45	0.09	6.51	4.12
INW-24	5357	1621	0.30	0.36	0.89	0.12	312	1.54	0.14	7.18	5.90
INW-25	1733	692	0.40	0.29	0.80	0.11	89	1.47	0.17	6.59	4.65
IN-01	26540	3050	0.11	0.38	0.95	0.08	1522	1.86	0.28	6.11	4.83
IN-02	29393	2792	0.09	0.41	0.99	0.08	1588	1.90	0.35	6.33	5.11
IN-03	56870	5312	0.09	0.40	0.95	0.07	2639	1.93	0.36	7.10	5.79
IN-04	77882	6176	0.08	0.39	0.97	0.08	4479	2.00	0.32	6.88	5.41
IN-05	78705	7009	0.09	0.39	0.93	0.09	3732	1.94	0.41	7.13	5.67
IN-06	80147	6774	0.08	0.44	0.99	0.08	4167	1.88	0.41	7.05	5.96
IN-07	83683	8741	0.10	0.31	0.87	0.10	5221	1.78	0.35	7.60	5.97
IN-08	84857	7433	0.09	0.38	0.93	0.09	4535	1.95	0.38	6.99	5.52
IN-09	88887	5149	0.06	0.49	1.06	0.06	4585	2.11	0.39	6.62	5.24
IN-10	90913	7560	0.08	0.45	0.99	0.07	3546	1.94	0.33	7.28	5.67
IN-11	95587	7973	0.08	0.51	1.08	0.06	5442	1.92	0.43	6.70	6.15
IN-12	102837	10771	0.10	0.33	0.88	0.10	6424	1.75	0.34	7.47	5.93
IN-13	103826	8274	0.08	0.39	0.95	0.08	5841	1.92	0.42	7.26	6.11
IN-14	119460	6404	0.05	0.41	0.93	0.08	4103	2.07	0.41	7.59	6.38
IN-15	121747	6378	0.05	0.37	0.91	0.08	4331	2.08	0.43	7.64	6.29
IN-16	130594	10738	0.08	0.35	0.90	0.10	6765	1.89	0.43	7.58	5.92
IN-17	136046	10042	0.07	0.41	0.97	0.09	7991	1.90	0.38	7.56	6.01
IN-18	143620	6157	0.04	0.58	1.23	0.06	7203	2.21	0.48	7.33	6.69
IN-19	145179	11134	0.08	0.34	0.90	0.09	7086	1.92	0.42	7.65	5.94
IN-20	160769	9467	0.06	0.47	1.01	0.08	7864	2.04	0.43	7.03	5.67
IN-21	151406	7652	0.05	0.58	1.20	0.04	10199	2.19	0.45	7.38	6.97
IN-22	160303	7271	0.05	0.41	0.94	0.07	5204	2.15	0.44	7.64	6.28
IN-23	168362	12542	0.07	0.36	0.91	0.09	12716	1.89	0.38	7.66	6.29
IN-24	209887	18185	0.09	0.40	0.95	0.08	14173	1.83	0.36	7.76	6.10
IN-25	218316	10863	0.05	0.48	1.04	0.06	20351	2.13	0.40	7.60	6.77

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
ITW-01	1036	502	0.48	0.20	0.77	0.12	44	1.37	0.00	7.32	4.82
ITW-02	3047	1266	0.42	0.17	0.72	0.16	116	1.46	0.10	7.26	4.16
ITW-03	1024	504	0.49	0.28	0.76	0.15	40	1.33	0.00	7.11	4.59
ITW-04	4238	1585	0.37	0.32	0.87	0.09	176	1.54	0.07	7.49	5.34
ITW-05	1635	653	0.40	0.21	0.78	0.12	70	1.45	0.04	7.52	5.48
ITW-06	3006	1132	0.38	0.28	0.83	0.07	103	1.57	0.00	7.70	4.90
ITW-07	3063	1123	0.37	0.22	0.77	0.16	137	1.45	0.04	7.82	5.39
ITW-08	3663	1364	0.37	0.30	0.86	0.11	158	1.51	0.09	7.59	5.68
ITW-09	1890	915	0.48	0.20	0.75	0.15	70	1.48	0.04	6.80	3.57
ITW-10	1645	630	0.38	0.25	0.82	0.13	73	1.52	0.04	7.06	4.87
ITW-11	6216	2146	0.35	0.34	0.87	0.11	235	1.55	0.13	7.37	5.44
ITW-12	2184	873	0.40	0.31	0.83	0.16	121	1.53	0.11	7.55	4.71
ITW-13	2208	857	0.39	0.30	0.85	0.06	86	1.58	0.04	7.46	5.00
ITW-14	2786	1176	0.42	0.28	0.82	0.08	136	1.44	0.05	7.32	5.20
ITW-15	1963	832	0.42	0.26	0.81	0.14	71	1.50	0.00	7.22	4.67
ITW-16	2303	1025	0.45	0.26	0.77	0.15	86	1.46	0.04	7.40	4.06
ITW-17	440	267	0.61	0.13	0.77	0.13	56	1.56	0.00	7.12	2.33
ITW-18	773	374	0.48	0.21	0.83	0.11	54	1.62	0.00	7.03	4.40
ITW-19	2563	939	0.37	0.33	0.89	0.09	108	1.61	0.04	8.04	5.46
ITW-20	7146	2530	0.35	0.31	0.83	0.13	266	1.49	0.13	7.68	5.26
ITW-21	1440	664	0.46	0.27	0.80	0.11	87	1.60	0.00	6.77	3.83
ITW-22	1605	740	0.46	0.20	0.79	0.10	87	1.50	0.00	7.59	4.25
ITW-23	4188	1584	0.38	0.26	0.79	0.15	154	1.56	0.04	7.79	5.13
ITW-24	2227	950	0.43	0.32	0.85	0.09	81	1.51	0.00	7.50	4.61
ITW-25	2401	958	0.40	0.22	0.80	0.10	101	1.49	0.00	7.69	5.36
IT-01	23161	5486	0.24	0.34	0.89	0.09	715	1.70	0.18	7.61	5.27
IT-02	34159	6418	0.19	0.24	0.77	0.14	1346	1.74	0.28	6.46	4.58
IT-02	34940	6986	0.20	0.41	0.97	0.08	1195	1.67	0.23	7.60	5.93
IT-04	46695	9223	0.20	0.34	0.88	0.10	1688	1.68	0.26	8.10	5.97
IT-05	58568	9402	0.16	0.39	0.96	0.09	2130	1.74	0.24	7.64	5.85
IT-06	60137	9181	0.15	0.34	0.88	0.09	1860	1.78	0.25	7.74	5.91
IT-07	65121	9442	0.14	0.35	0.90	0.10	2331	1.81	0.29	7.88	6.07
IT-08	65075	10797	0.17	0.35	0.90	0.10	2202	1.72	0.30	8.01	6.10
IT-09	65335	12962	0.20	0.35	0.91	0.10	2557	1.65	0.28	8.09	6.30
IT-10	67002	10233	0.15	0.33	0.89	0.10	2405	1.73	0.31	7.99	6.05
IT-11	67857	12102	0.18	0.29	0.82	0.13	2106	1.68	0.28	8.06	6.02
IT-12	69182	11244	0.16	0.30	0.85	0.11	2284	1.71	0.29	8.22	6.09
IT-13	70503	10553	0.15	0.35	0.89	0.11	2323	1.77	0.31	7.98	6.27
IT-14	70309	12072	0.17	0.29	0.84	0.11	2005	1.71	0.30	8.13	6.01
IT-15	73489	10930	0.15	0.35	0.91	0.10	2586	1.78	0.26	7.93	6.14
IT-16	78761	11818	0.15	0.34	0.89	0.10	2781	1.77	0.25	7.87	6.15
IT-17	78005	12150	0.16	0.31	0.86	0.11	2362	1.74	0.28	8.15	6.13
IT-18	78534	13252	0.17	0.31	0.86	0.10	2540	1.68	0.27	8.16	6.42
IT-19	78570	12267	0.16	0.30	0.84	0.12	2635	1.72	0.25	8.34	6.28
IT-20	87888	13012	0.15	0.32	0.87	0.11	3013	1.77	0.28	7.85	6.04
IT-21	91804	12993	0.14	0.34	0.90	0.10	2822	1.78	0.28	8.17	6.24
IT-22	92696	13744	0.15	0.40	0.96	0.09	4348	1.77	0.26	8.09	6.77
IT-23	109436	13357	0.12	0.37	0.93	0.09	3538	1.82	0.26	8.13	6.48
IT-24	125958	15315	0.12	0.32	0.88	0.10	4431	1.77	0.32	8.32	6.57
IT-25	195258	20435	0.10	0.36	0.91	0.10	6361	1.83	0.33	8.39	6.75

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
PTW-01	669	353	0.53	0.20	0.69	0.08	55	1.42	0.08	7.18	5.25
PTW-02	994	478	0.48	0.17	0.72	0.13	49	1.57	0.13	6.82	2.78
PTW-03	1762	676	0.38	0.30	0.91	0.04	96	1.51	0.05	6.85	4.53
PTW-04	1689	687	0.41	0.26	0.79	0.17	65	1.43	0.04	7.27	4.48
PTW-05	1327	561	0.42	0.10	0.68	0.17	80	1.46	0.16	6.69	3.62
PTW-06	2149	845	0.39	0.25	0.82	0.14	131	1.48	0.15	6.47	4.42
PTW-07	8835	2019	0.23	0.42	0.94	0.09	451	1.58	0.22	7.15	5.62
PTW-08	1096	517	0.47	0.16	0.67	0.17	124	1.35	0.00	7.38	5.25
PTW-09	748	364	0.49	0.19	0.69	0.23	124	1.36	0.00	6.59	4.27
PTW-10	2285	837	0.37	0.30	0.83	0.12	124	1.51	0.00	7.40	5.40
PTW-11	1162	541	0.47	0.18	0.72	0.21	124	1.42	0.00	7.40	4.54
PTW-12	2481	930	0.37	0.26	0.80	0.13	101	1.46	0.08	7.36	4.52
PTW-13	10964	2929	0.27	0.35	0.90	0.09	499	1.49	0.21	8.00	6.52
PTW-14	1908	681	0.36	0.25	0.80	0.14	104	1.48	0.00	7.24	5.35
PTW-15	1692	697	0.41	0.22	0.76	0.14	99	1.40	0.00	7.45	5.96
PTW-16	545	293	0.54	0.14	0.68	0.19	34	1.37	0.00	6.95	3.60
PTW-17	1164	512	0.44	0.20	0.78	0.10	48	1.52	0.00	6.59	3.62
PTW-18	3720	999	0.27	0.38	0.92	0.08	202	1.63	0.27	6.31	4.77
PTW-19	863	406	0.47	0.20	0.76	0.12	44	1.45	0.00	7.08	3.42
PTW-20	916	418	0.46	0.36	0.86	0.10	49	1.65	0.00	6.92	3.94
PTW-21	2219	772	0.35	0.24	0.84	0.08	107	1.50	0.19	7.26	5.06
PTW-22	1215	536	0.44	0.22	0.79	0.12	57	1.60	0.05	7.18	3.88
PTW-23	803	413	0.51	0.11	0.72	0.09	48	1.36	0.00	6.47	2.83
PTW-24	11039	2593	0.23	0.35	0.92	0.11	530	1.57	0.22	7.71	6.41
PTW-25	2403	1002	0.42	0.15	0.78	0.12	134	1.44	0.00	7.51	4.57
PT-01	21266	4820	0.23	0.34	0.90	0.09	1055	1.66	0.25	7.64	5.92
PT-02	23284	3240	0.14	0.36	0.91	0.08	1387	1.89	0.23	7.10	5.23
PT-03	56710	10626	0.19	0.34	0.88	0.10	2302	1.70	0.31	7.40	5.40
PT-04	60693	11465	0.19	0.20	0.75	0.13	2454	1.61	0.31	8.20	5.86
PT-05	64624	9164	0.14	0.35	0.89	0.10	2678	1.77	0.32	7.58	5.82
PT-06	64625	11186	0.17	0.36	0.92	0.09	2282	1.68	0.31	7.89	6.03
PT-07	66807	12015	0.18	0.30	0.85	0.12	2504	1.67	0.31	7.49	6.00
PT-08	67361	11333	0.17	0.36	0.92	0.10	2898	1.69	0.32	7.67	5.78
PT-09	68269	10702	0.16	0.34	0.92	0.09	2456	1.76	0.29	7.76	5.87
PT-10	68616	12686	0.18	0.32	0.87	0.10	2884	1.58	0.31	7.70	5.96
PT-11	68681	14254	0.21	0.26	0.81	0.13	4172	1.57	0.35	7.81	5.77
PT-12	72164	11570	0.16	0.29	0.83	0.11	2727	1.71	0.30	7.84	5.88
PT-13	75278	10630	0.14	0.34	0.90	0.09	2821	1.81	0.34	7.73	6.01
PT-14	75686	12161	0.16	0.27	0.83	0.12	3255	1.67	0.35	8.04	6.10
PT-15	75905	11473	0.15	0.33	0.89	0.11	2951	1.74	0.29	7.84	6.13
PT-16	78988	16670	0.21	0.24	0.79	0.13	3517	1.56	0.26	8.14	6.06
PT-17	79480	11987	0.15	0.28	0.83	0.11	3414	1.72	0.33	7.94	5.93
PT-18	92162	13624	0.15	0.33	0.89	0.11	3639	1.76	0.33	8.16	6.26
PT-19	95452	12693	0.13	0.34	0.90	0.11	4570	1.73	0.33	8.18	6.42
PT-20	105093	14127	0.13	0.33	0.89	0.11	4704	1.77	0.35	8.25	6.35
PT-21	106092	12057	0.11	0.35	0.89	0.10	4373	1.81	0.32	7.89	6.24
PT-22	117460	14854	0.13	0.32	0.87	0.10	4721	1.76	0.30	8.07	6.30
PT-23	138504	16582	0.12	0.37	0.92	0.10	5753	1.77	0.30	8.15	6.40
PT-24	151027	23894	0.16	0.31	0.86	0.11	6462	1.67	0.32	8.75	7.04
PT-25	212991	21931	0.10	0.36	0.93	0.09	7747	1.82	0.33	8.15	6.56

Referência	T	V	D	η	$\bar{\sigma}$	γ	k_{max}	$\overline{H}$	θ	$\bar{l}_v$	$\bar{l}_{vr}$
SEW-01	3259	1347	0.41	0.25	0.82	0.12	138	1.47	0.04	7.17	5.63
SEW-02	2762	1165	0.42	0.25	0.81	0.10	104	1.45	0.09	7.64	5.61
SEW-03	2136	887	0.42	0.29	0.86	0.11	57	1.48	0.14	7.22	4.12
SEW-04	1456	632	0.43	0.26	0.84	0.09	65	1.46	0.00	7.54	5.48
SEW-05	1937	718	0.37	0.28	0.80	0.13	57	1.53	0.06	7.12	5.22
SEW-06	2459	937	0.38	0.20	0.76	0.14	110	1.44	0.04	7.79	5.60
SEW-07	2516	914	0.36	0.29	0.83	0.08	93	1.48	0.19	7.32	5.46
SEW-08	2594	1125	0.43	0.25	0.80	0.11	75	1.44	0.03	7.67	5.39
SEW-09	2862	1209	0.42	0.26	0.78	0.11	94	1.46	0.10	7.74	3.98
SEW-10	3469	1201	0.35	0.29	0.87	0.12	141	1.48	0.07	7.61	5.51
SEW-11	4336	1521	0.35	0.35	0.89	0.12	142	1.47	0.03	7.18	4.75
SEW-12	1418	659	0.46	0.21	0.75	0.12	66	1.40	0.08	7.65	5.90
SEW-13	2169	855	0.39	0.36	0.88	0.12	76	1.48	0.04	7.16	4.35
SEW-14	686	385	0.56	0.22	0.79	0.10	28	1.25	0.00	7.30	4.27
SEW-15	1381	588	0.43	0.29	0.85	0.07	61	1.56	0.00	7.20	4.46
SEW-16	1134	549	0.48	0.16	0.70	0.19	55	1.49	0.00	6.79	3.36
SEW-17	964	471	0.49	0.17	0.75	0.14	34	1.33	0.00	6.90	4.46
SEW-18	554	323	0.58	0.08	0.65	0.15	26	1.28	0.00	7.13	4.33
SEW-19	4993	1870	0.37	0.32	0.88	0.10	141	1.56	0.03	7.34	6.24
SEW-20	1133	546	0.48	0.24	0.72	0.15	103	1.39	0.00	6.75	3.25
SEW-21	516	344	0.67	0.17	0.68	0.12	36	1.38	0.07	7.70	4.00
SEW-22	3613	1620	0.45	0.24	0.80	0.12	159	1.36	0.04	7.12	4.77
SEW-23	5006	1935	0.39	0.26	0.83	0.11	166	1.40	0.13	7.96	5.54
SEW-24	5423	1847	0.34	0.34	0.87	0.10	223	1.50	0.07	7.95	6.32
SEW-25	5699	1991	0.35	0.35	0.88	0.13	185	1.47	0.14	7.82	5.38
SE-01	17743	2974	0.17	0.32	0.86	0.09	716	1.79	0.17	6.37	4.81
SE-02	16604	4079	0.25	0.33	0.90	0.08	616	1.67	0.22	7.29	5.04
SE-03	25490	4116	0.16	0.32	0.87	0.09	983	1.79	0.26	6.82	4.98
SE-04	32445	5152	0.16	0.33	0.88	0.10	1153	1.79	0.26	7.07	5.16
SE-05	34439	7333	0.21	0.32	0.87	0.11	1434	1.68	0.22	7.45	5.34
SE-06	36731	6251	0.17	0.34	0.90	0.09	1398	1.72	0.31	7.22	5.18
SE-07	37296	6871	0.18	0.36	0.91	0.09	1528	1.72	0.25	7.08	5.21
SE-08	39873	6527	0.16	0.30	0.86	0.10	1850	1.75	0.33	7.36	5.21
SE-09	43383	6734	0.16	0.40	0.95	0.09	1748	1.74	0.33	6.79	5.30
SE-10	48669	5556	0.11	0.39	0.94	0.08	2284	1.92	0.31	6.90	5.15
SE-11	45680	9457	0.21	0.33	0.88	0.10	1917	1.66	0.27	7.52	5.10
SE-12	45028	9868	0.22	0.24	0.79	0.13	1567	1.56	0.33	7.74	5.36
SE-13	46068	11259	0.24	0.29	0.86	0.12	1665	1.58	0.26	7.60	5.33
SE-14	52282	7730	0.15	0.37	0.92	0.08	2135	1.79	0.33	7.10	5.26
SE-15	53421	7968	0.15	0.31	0.85	0.10	1897	1.77	0.31	7.49	5.52
SE-16	62271	8728	0.14	0.33	0.86	0.11	2130	1.75	0.34	7.61	5.42
SE-17	64463	9031	0.14	0.37	0.93	0.08	2689	1.83	0.34	7.22	5.27
SE-18	66213	9155	0.14	0.36	0.90	0.10	2423	1.80	0.34	7.55	5.36
SE-19	65143	7858	0.12	0.48	1.03	0.07	1938	1.89	0.26	7.64	6.09
SE-20	68765	10823	0.16	0.34	0.90	0.10	3402	1.73	0.31	7.52	5.74
SE-21	76031	7841	0.10	0.39	0.94	0.09	3110	1.90	0.40	7.30	5.52
SE-22	79451	11804	0.15	0.37	0.92	0.10	3602	1.79	0.31	7.60	5.74
SE-23	85625	14428	0.17	0.33	0.89	0.10	4634	1.71	0.32	7.71	5.46
SE-24	93009	9738	0.10	0.43	1.00	0.08	4037	1.86	0.40	7.69	5.89
SE-25	130641	14536	0.11	0.30	0.85	0.10	6482	1.81	0.37	7.58	5.77

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Nino Boccara. *Modeling Complex Systems*. Springer Verlag, 2004.

[2] Sunny Y. Auyang. *Foundations of Complex-system Theories: In Economics, Evolutionary Biology, and Statistical Physics*. Cambridge University Press, 1998.

[3] *Ethnologue - Languages of the World http://www.ethnologue.com/world*.

[4] Charles Higounet. *História Concisa da Escrita*. Parábola, 2011.

[5] Ferdinand de Saussure. *Curso de Linguística Geral*. Cultrix, 2013.

[6] Noam Chomsky. *Linguagem e Pensamento*. Vozes, 1971.

[7] Noam Chomsky. *Sobre natureza e linguagem*. Martins Fontes, 2006.

[8] Steven Pinker. *Do Que É Feito o Pensamento*. Companhia das Letras, 2008.

[9] Thomas L. Griffiths. Rethinking language: How probabilities shape the words we use. *Proceedings of the National Academy of Sciences*, 108(10):3825–3826, 03 2011.

[10] James Gleick. *A Informação*. Companhia das Letras, 2011.

[11] Igor Zolnerkevic. A vida das palavras. *Pesquisa FAPESP*, 185, Julho 2011.

[12] Gabriel Altmann. On the symbiosis of physicists and linguists. *Romanian Reports in Physics*, 60(3):417–422, 2008.

[13] George Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, 1949.

[14] George Zipf. *The Psycho-biology of Language*. MIT Press, 1936.

[15] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal,*, 27(3):379–423, 1948.

[16] B. Mandelbrot. An informational theory of the statistical structure of language. *Communication Theory*, 1953.

[17] J. P. Herrera and P. A. Pury. Statistical keyword detection in literary corpora. *The European Physical Journal B*, 63(1):135–146, 2008.

[18] Ramon Ferrer i Cancho and Ricard V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

[19] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA, 1978.

[20] Steven T. Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 01 2011.

[21] Marcelo A. Montemurro and Damián H. Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2013/08/21 2010.

[22] Martin Gerlach and Eduardo G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, May 2013.

[23] Eduardo G. Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 07 2012.

[24] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 05 2009.

[25] M. Ausloos. Measuring complexity with multifractals in texts. translation effects. *Chaos, Solitons & Fractals*, 45(11):1349–1357, 11 2012.

[26] Marzio Cassandro, Pierre Collet, Antonio Galves, and Charlotte Galves. A statistical-physics approach to language acquisition and language change. *Physica A: Statistical Mechanics and its Applications*, 263(1–4):427–437, 2 1999.

[27] Harmen J. Bussemaker, Hao Li, and Eric D. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences*, 97(18):10096–10100, 08 2000.

[28] Alexander M. Petersen, Joel N. Tenenbaum, Shlomo Havlin, H. Eugene Stanley, and Matjaž Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, 2, 12 2012.

[29] Kosmas Kosmidis, Alkiviadis Kalampokis, and Panos Argyrakis. Statistical mechanical approach to human language. *Physica A: Statistical Mechanics and its Applications*, 366(0):495–502, 7 2006.

[30] Michael W. Berry and Jacob Kogan, editors. *Text Mining*. Wiley, 2010.

[31] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 01 2011.

[32] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April 1958.

[33] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. M. Somoza. Keyword detection in natural languages and dna. *EPL (Europhysics Letters)*, 57(5):759, 2002.

[34] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado, and J. L. Oliver. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*, 79(3):035102–, 03 2009.

[35] Hongding Zhou and Gary W. Slater. A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, 329(1–2):309–327, 11 2003.

[36] Zhen Yang, Jianjun Lei, Kefeng Fan, and Yingxu Lai. Keyword extraction by entropy difference between the intrinsic and extrinsic mode. *Physica A: Statistical Mechanics and its Applications*, 392(19):4523–4531, 10 2013.

[37] Ali Mehri and Amir H. Darooneh. Keyword extraction by nonextensivity measure. *Physical Review E*, 83(5):056106–, 05 2011.

[38] C. Carretero-Campos, P. Bernaola-Galván, A. V. Coronado, and P. Carpena. Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications*, 392(6):1481–1492, 3 2013.

[39] Viviane M. de Oliveira, M. A. F. Gomes, and I. R. Tsang. Theoretical model for the evolution of the linguistic diversity. *Physica A: Statistical Mechanics and its Applications*, 361(1):361–370, 2 2006.

[40] Ricard V. Solé, Bernat Corominas-Murtra, and Jordi Fortuny. Diversity, competition, extinction: the ecophysics of language change. *Journal of The Royal Society Interface*, 06 2010.

[41] Steven Roger Fischer. *Uma Breve História da Linguagem*. Novo Século, 2009.

[42] Nitin Indurkhya and Fred J. Damereu, editors. *Handbook of natural language processing*. CRC Press, 2010.

[43] José Fernando M. Rocha, editor. *Origens e Evolução das Ideias da Física*. EDUFBA, 2011.

[44] Enrico Fermi. *Thermodynamics*. Dover, 1936.

[45] Rudolf Clausius. *The Mechanical Theory of Heat:*. J. Van Voorst, 1867.

[46] Linda E. Reichl. *A Modern Course in Statistical Physics*. Wiley, 2 edition, 1998.

[47] R. P. Feynman, R. B. Leighton, and M. Sands. *Lições de Física*, volume 1. Bookman, 2009.

[48] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

[49] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–170, 03 1992.

[50] Alain Schenkel, Jun Zhang, and Yi-Cheng Zhang. Long-range correlations in human writings. *Fractals*, 01(01):47–57, 2014/04/09 1993.

[51] Werner Ebeling and Alexander Neiman. Long-range correlations between letters and sentences in texts. *Physica A: Statistical Mechanics and its Applications*, 215(3):233–241, 5 1995.

[52] Marcelo A. Montemurro and Pedro A. Pury. Long-range fractal correlations in literary corpora. *Fractals*, 10(04):451–461, 2014/04/09 2002.

[53] T. A. Brody, J. Flores, J. B. French, P. A. Mello, A. Pandey, and S. S. M. Wong. Random-matrix physics: spectrum and strength fluctuations. *Rev. Mod. Phys.*, 53:385–479, Jul 1981.

[54] Berryman M. J., A. Allison, and D. Abbott. Statistical techniques for text classification based on word recurrence intervals. *Fluctuation and Noise Letters*, 03(01):L1–L10, 2014/06/13 2003.

[55] Diego Raphael Amancio, Eduardo G Altmann, Osvaldo N Oliveira Jr, and Luciano da Fontoura Costa. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, 13(12):123024, 2011.

[56] Michael Hackenberg, Antonio Rueda, Pedro Carpena, Pedro Bernaola-Galván, Guillermo Barturen, and José L. Oliver. Clustering of dna words and biological function: A proof of principle. *Journal of Theoretical Biology*, 297(0):127–136, 3 2012.

[57] Marcelo A. Montemurro and Damián H. Zanette. Entropic analysis of the role of words in literary texts. *Advances in Complex Systems*, 05(01):7–17, 2014/06/19 2002.

[58] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988.

[59] Ali Mehri and Amir H. Darooneh. The role of entropy in word ranking. *Physica A: Statistical Mechanics and its Applications*, 390(18–19):3157–3163, 9 2011.

[60] Marcelo A. Montemurro and Damián H. Zanette. Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. *PLoS ONE*, 8(6):e66344 EP –, 06 2013.

[61] Popescu Ioan-Iovitz. *Word Frequency Studies*. De Gruyter Mouton CY - Berlin, Boston, 2009.

[62] Bernat Corominas-Murtra and Ricard V. Solé. Universality of zipf's law. *Physical Review E*, 82(1):011102–, 07 2010.

[63] Damián H. Zanette and Susanna C. Manrubia. *Multiplicative processes in social systems*, volume Volume 7, pages 129–158. World Scientific, 2014/07/22 2007.

[64] G. Herdan. 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):222–228, 1958.

[65] Francesc Font-Clos, Gemma Boleda, and Álvaro Corral. A scaling law beyond zipf's law and its relation to heaps' law. *New Journal of Physics*, 15(9):093033, 2013.

[66] Tommaso Pola. Statistical analysis of written languages. Master's thesis, Università di Bologna, 13 Dicembre 2013 2013.

[67] Alexander F. Gelbukh and Grigori Sidorov. Zipf and heaps laws' coefficients depend on language. In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '01, pages 332–335, London, UK, UK, 2001. Springer-Verlag.

[68] Sebastian Bernhardsson, Luis Enrique Correa da Rocha, and Petter Minnhagen. The meta book and size-dependent properties of written language. *New Journal of Physics*, 11(12):123015, 2009.

[69] Xiao-Yong Yan and Petter Minnhagen. Maximum entropy, word-frequency, chinese characters, and multiple meanings. *preprint (arXiv:1402.1939v1)*, 2014.

[70] R. D. Lord. Studies in the history of probability and statistics. viii. de morgan and the statistical study of literary style. *Biometrika*, 45(1-2):282, 1958.

[71] T. C. Mendenhall. The characteristic curves of composition. *Science*, ns-9(214S):237–246, 1887.

[72] Peter Grzybek, editor. *Contributions to the Science of Text and Language*. Springer, 2006.

[73] Constantinos Papadimitriou Konstantinos Karamanos Fotis K. Diakonos Maria Kalimeri, Vassilios Constantoudis and Harris Papageorgiou. Word-length entropies and correlations of natural language written texts. *arXiv:cond-mat/1401.6224*, 2014.

[74] M. L. Metha. *Random Matrices*. Academic Press, 2004.

[75] Pedro Carpena, Pedro Bernaola-Galván, and Plamen Ch. Ivanov. New class of level statistics in correlated disordered chains. *Phys. Rev. Lett.*, 93:176804, Oct 2004.

[76] N. Argaman, F. M. Dittes, E. Doron, J. P. Keating, A. Yu. Kitaev, M. Sieber, and U. Smilansky. Correlations in the actions of periodic orbits derived from quantum chaos. *Physical Review Letters*, 71(26):4326–4329, 12 1993.

[77] Eric Goles, Oliver Schulz, and Mario Markus. Prime number selection of cycles in a predator-prey model. *Complexity*, 6(4):33–38, 2001.

[78] Haret C. Rosu. Quantum hamiltonians and prime numbers. *Modern Physics Letters A*, 18(18):1205–1213, 2014/08/01 2003.

[79] Z. Gamba, J. Hernando, and L. Romanelli. Are prime numbers regularly ordered? *Physics Letters A*, 145(2–3):106–108, 4 1990.

[80] S. R Dahmen, S. D Prado, and T Stuermer-Daitx. Similarity in the statistics of prime numbers. *Physica A: Statistical Mechanics and its Applications*, 296(3–4):523–528, 7 2001.

[81] Boon Leong Lan and Shaohen Yong. Power spectrum of the difference between the prime-number counting function and riemann's function: 1/f2? *Physica A: Statistical Mechanics and its Applications*, 334(3–4):477–481, 3 2004.

[82] George G. Szpiro. Peaks and gaps: Spectral analysis of the intervals between prime numbers. *Physica A: Statistical Mechanics and its Applications*, 384(2):291–296, 10 2007.

[83] Marek Wolf. Nearest-neighbor-spacing distribution of prime numbers and quantum chaos. *Phys. Rev. E*, 89:022922, Feb 2014.

[84] C. F. Gauss. *Werke*, volume II. 1849.

[85] A. M. Ledengre. *Essai sur la Theorie des Nombres*. Duprat, 1798.

[86] D. W. DeTemple. The non-integer property of sums of reciprocals of consecutive inte- gers. *Math. Gaz*, (75):193–194, 1991.

[87] Julian Havil. *Gamma: Exploring Euler's Constant*. Princeton University Press, 2009.

[88] H. Eugene Stanley. *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press, 1971.

[89] *Project Gutenberg: www.gutenberg.org.*

[90] *Domínio Público: www.dominiopublico.gov.br.*

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).