Universidade Federal Rural de Pernambuco
Departamento de Física
Programa de Pós-graduação em Física Aplicada

José Ilberto Fonceca Junior

Estudo da correlação entre propriedades estatísticas de verbetes

Recife - PE

2017

José Ilberto Fonceca Junior

Estudo da correlação entre propriedades estatísticas de verbetes

Dissertação apresentada ao Programa de Pós-graduação em Física Aplicada do Departamento de Física da UFRPE, como requisito para a obtenção do grau de MESTRE em Física Aplicada.

Orientador: Pedro Hugo de Figueirêdo

Doutor

Recife - PE

2017

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife - PE, Brasil

F673e Fonceca Junior, José Ilberto
Estudo da correlação entre propriedades estatísticas de verbetes / José Ilberto Fonceca Junior - 2017
118.f.:il.
Orientador: Pedro Hugo de Figueirêdo
Dissertação (Mestrado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Física Aplicada, Recife, BR-PE, 2017.
Inclui Referências e apêndice(s).
1. Entropia 2. Linguística quantitativa 3. Mecânica estatística 4. Métodos estatísticos aplicados I. Figueirêdo, Pedro Hugo de, orient. II. Título
CDD 621

José Ilberto Fonceca Junior

Estudo da correlação entre propriedades estatísticas de verbetes

Dissertação apresentada ao Programa de Pós-graduação em Física Aplicada do Departamento de Física da UFRPE, como requisito para a obtenção do grau de MESTRE em Física Aplicada.

Aprovado em 19 de abril de 2017

BANCA EXAMINADORA

Pedro Hugo de Figueirêdo

Doutor

Adauto José Ferreira de Souza

Doutor

Ramón Enrique Ramayo González

Doutor

A todos que participam da minha caminhada aleatória nos domínios científicos, em especial meu pai e Karina.

Agradecimentos

As modestas contribuições que serão expostas nesse trabalho não seriam possíveis sem a orientação do professor Pedro Hugo. Seus *insights*, conselhos, empenho em participar o tanto quanto seja necessário no processo produtivo e sua humanidade direcionada aos seus orientandos e alunos são características únicas e inspiradoras para a próxima geração de profissionais que ele tem ajudado a formar. Quando esse trabalho estiver entregue, espero que comer uma tapioca ou um pastel enquanto conversamos sobre a vida e o trabalho seja sempre uma possibilidade.

Sou grato também a minha família, especialmente a Sr. Ilberto e Rodolfo, pelo apoio e incentivo durante a minha formação profissional. Com eles sempre foi possível dividir um pouco das conquistas intelectuais, angústias e felicidades pessoais e acadêmicas. Em especial, sou grato ao apoio do meu pai, a pessoa que talvez se sinta mais feliz e orgulhosa com cada um dos passos que tenho dado.

A Karina Mavignier por ser a pessoa que talvez mais precise aturar as angústias, ausência e minha incapacidade de administrar o tempo, esse trabalho talvez deva mais a sua compreensão e motivação do que ao meu empenho. Não poderia esquecer também de Dona Bernadete, Sr. Ronald (que diria que essa dissertação é coisa de um “rapaz porreta”) e Rafinha que tanto tiveram que ouvir que eu tinha uma lista pendente e precisava estudar, sou muito grato por ter vocês por perto.

Aos amigos da graduação sou grato pelas conversas, risadas, conselhos e rodízios, vocês têm ajudado a tornar as piores partes do caminho um pouco mais felizes e bem humoradas. Fico muito agradecido aos companheiros da pós pelas horas discutindo listas, ideias e sugestões, com vocês pude dividir parte dos problemas que surgiram e espero continuar dividindo.

Agradeço também a coordenação da Pós-Graduação por sua disponibilidade em solucionar os problemas que surgiram durante o período e a CAPES pelo financiamento em forma de bolsa de estudo, sem essas contribuições, a realização desse trabalho e também da Pós-Graduação seria bem mais árduas.

"Of all the possible pathways to disorder,
nature favors just a few."

James Gleick

Resumo

As investigações das línguas naturais através da aplicação de métodos matemáticos e estatísticos que buscam caracterizar propriedades de textos literários têm sido objeto de intensa investigação nas últimas décadas, constituindo uma área denominada de linguística quantitativa. Os primeiros trabalhos nessa área surgiram entre as décadas de 1930 e 1950, com os trabalhos de George Zipf no estudo da distribuição de frequências e Claude Shannon com seu trabalho em previsão de letras e palavras e entropia como medida de redundância em língua inglesa. Nesta dissertação serão investigadas a autocorrelação e correlações cruzadas das séries temporais utilizando técnicas comuns ao estudo de séries temporais não-estacionárias. Discutiremos também quais propriedades emergem dessas correlações e suas implicações no processo de escrita. Ao longo dessa análise, todos os resultados foram obtidos para um conjunto de 250 textos literários escritos em 10 línguas distintas. No momento final desse trabalho, analisaremos as propriedades de textos genéricos obtidos através de dois modelos de distribuições de distância: uma que leva em consideração as distâncias entre os números primos consecutivos e outra que utiliza a distribuição de Weibull. Exploraremos as características que surgem em cada um dos modelos comparando-as com seus equivalentes nos textos em linguagem natural.

Palavras-Chave: entropia, linguística quantitativa, mecânica estatística e métodos estatísticos aplicados.

Abstract

The application of mathematical and statistical methods to exploit properties in natural languages has a recent and prolific history. These methods and the quantitative tecnhiques adapted and created through the study of languages are part of an area usually called quantitative linguistics. The first work on such area was performed by George Zipf from 1930 to 1950 in which the distribution of word frequencies were studied. His works were followed by Claude Shannon's analysis on entropy and letters prediction as a measure of redundancy in written english. In this work, we firstly present a study on correlation and cross-correlation through the time series extracted from texts by using common approaches to investigate non-stationary time series. To perform the required analysis we have used a *corpora* as large as 250 literary texts from 10 different languages. The properties emerging from these correlations will also be discussed and properly explained. Secondly, we move to the description of the distance distribution responsible for the long-range structure observed on written language. We devise those distributions by assuming the distance distribution from consecutive prime numbers and distances taken from a Weibull distributed process. The revenues from such models will be put under scrutiny by using the techniques presented during the work and comparing them to properties emerging in natural language.

Keywords: entropy, quantitative linguistics, statistical mechanics and applied statistical methods.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

O estudo de sistemas complexos utilizando conceitos advindos da física tem possibilitado a inserção de modelos físicos em diversas áreas do conhecimento. Desde a década de 1970, essa interação entre a física e demais áreas originou ramos como econofísica, psicofísica, sociofísica e biofísica, assim como possibilitou o uso de modelos físicos em áreas tão diversas quanto linguística e urbanismo [1]. Nesses campos, a interação entre áreas permitiu a criação de novos modelos descritivos e de previsão dos fenômenos de interesse, assim como um aprofundamento da compreensão dos conceitos e processos físicos utilizados.

Sistemas complexos são comumente compostos de vários elementos que podem possuir diversos graus de liberdade e interagir através de mecanismos não-lineares e, muitas vezes, não triviais. Em geral, tais sistemas compartilham características fundamentais como heterogeneidade, adaptabilidade, frustração e memória [2].

Linguística quantitativa

A linguagem permitiu o ser humano criar representações e interpretações distintas para a realidade e também para entes abstratos. Inicialmente através da tradição oral em que todo discurso se mantinha restrito ao espaço e tempo, ou seja, o discurso proferido não iria além do círculo social que o orador e seus ouvintes compartilhavam e seu conteúdo se perderia tão logo as palavras fossem ditas [3]. No entanto, estima-se que existam entre 5000 e 7000 línguas vivas e todas elas são preservadas pela tradição oral [4].

Posteriormente, por volta de 3500 A.C., houve o desenvolvimento da linguagem escrita, ou linguagem natural, permitindo o registro e transmissão da linguagem de maneira não-local e atemporal uma vez que o instrumento de registro seja preservado [5]. A linguagem natural permitiu a evolução de conceitos cada vez mais abstratos o que tornou possível o estudo da própria linguagem em termos de elementos puramente abstratos, fenômeno que tem sido estudado pela psicologia e neurologia [6, 7].

Com o registro escrito da língua, torna-se possível a sua análise quantitativa e o estudo de suas propriedades estruturais. Esta característica tem feito com que físicos, estatísticos, matemáticos, cientistas da computação e da informação explorem as propriedades na busca de modelos capazes de descrever o comportamento quantitativo da língua [8], a esse estudo costuma-se dar o nome de *linguística quantitativa*. Os primeiros trabalhos relevantes na área foram feitos por George Zipf [9, 10], Claude Shannon [11, 12] e Benoit Mandelbrot [13] entre as décadas de 1930 e 1950.

Em relação aos desenvolvimentos na linguística quantitativa, a década de 2000 foi marcada pela presença cada vez mais comum de publicações com modelos físicos que descrevessem propriedades dos textos [14–16]. Parte desses modelos tenta explicar leis empíricas como:

i) a lei de Zipf que sugere que a distribuição do número de verbetes $n(k)$ com frequência k obedece a seguinte relação:
$$n(k) \sim k^{-\beta} ;$$

ii) a lei de Herdan-Heaps [17–19] que também relaciona o número total de verbetes V com o número total de palavras T através de uma lei de potência:
$$V \sim T^{\lambda}$$
em que os expoentes β e λ são característicos da língua estudada.

Apesar da diversidade de temas que encontram intersecções com diversas áreas, podemos classificar os tópicos da linguística quantitativa em três categorias:

1) **Cognição, evolução e caracteristícas universais da linguagem escrita**: Nessa categoria, os métodos quantitativos são empregados para estudar os fenômenos ligados à psicologia, neurologia e teoria da informação. Pode-se citar os estudos sobre a relação entre o comprimento das palavras e a eficiência da comunição [20], das evoluções sintática e lexical de idiomas distintos [21, 22] e os impactos da tradução nas propriedades estruturais de um texto [23, 24] como exemplos de trabalhos nessa área. Em uma das frentes dessa categoria, encontra-se trabalhos sobre as relações entre propriedades dos verbetes e o estudo de suas séries temporais [25–27].

2) **Extração de palavras-chave**:

Com o aumento dos recursos computacionais das últimas décadas, tornou-se possível a extração e armazenamento de quantidades progressivamente maiores de dados [28]. A extração de dados relevantes possui uma longa linha de pesquisa em linguística quantitativa, em especial a extração de palavras-chave em textos. Nessa subárea, o trabalho inaugural foi feito por Hans Luhn [29] em que ele propõe um método para a criação automática de resumos utilizando as frequências dos verbetes como medida de relevância. Trabalhos recentes [16, 30–33] sugerem que o uso da distribuição espacial dos verbetes como métrica para detecção de palavras-chave é uma abordagem com resultados significativos e mais eficientes computacionalmente.

3) **Distribuição espacial e diversidade linguística**:

Essa área tem uma concentração de estudos voltados à criação de modelos que expliquem de que maneira surge a distribuição espacial de verbetes [34, 35], assim como essa distribuição é capaz de moldar a diversidade de uma língua [36].

Para os estudos desenvolvidos na área, é comum adotar uma coletânea de textos de interesse em uma dada língua, chamado de *corpus*, ou até mesmo em muitas línguas (*corpora*). Na literatura, encontra-se trabalhos que usam somente um texto [16] a trabalhos que usam conjuntos de milhões de textos [22, 37]. Nossas análises serão feitas sobre diversos textos em idiomas distintos.

Corpora

Seguindo a metodologia adotada por Rolim [38], utilizaremos aqui um total de 485 textos em linguagem escrita de 10 línguas distintas. Sendo o *corpora* dividido em dois grandes conjuntos com 250 obras literárias (Apêndice A) e 235 artigos da wikipedia (Apêndice B). Justifca-se o uso desse *corpora* pela região de tamanhos de texto que ele possui, variando de centenas de palavras a centenas de milhares.

Os idiomas selecionados fazem parte de três grupos linguísticos distintos:

i) família germânica: alemão, dinamarquês, inglês e sueco;

ii) família latina: espanhol, francês, italiano e português;

iii) família urálica: finlandês e húngaro.

O critério de separação se dá pelo fato de famílias serem grupos linguísticos que possuem origem comum e que dividem correspondências sistemáticas em forma e significado não atribuíveis a mudanças ou apropriações [5]. Ou seja, elas possuem características únicas compartilhadas somente entre o próprio grupo, sendo algumas delas detectáveis através das métricas utilizadas nesse trabalho.

Todos os caracteres alfanuméricos foram considerados válidos em nossas análises independente da capitalização. Palavras compostas separadas por (-) ou (') foram consideradas válidas. De maneira mais geral, palavras foram definidas como qualquer combinação de caracteres alfanuméricos contida entre dois espaços em branco. Dessa forma, foi necessária a remoção da pontuação dos textos e quebras de linhas, processo conhecido como atomização, assim como todas as letras foram transformadas em minúsculas. Ambos os procedimentos são comuns à análise e mineração de textos em áreas como processamento de linguagem natural e linguística quantitativa [39, 40].

Uma vez que executamos esse mecanismo de mineração de palavras do texto, extraímos todas as palavras e seus parâmetros relevantes ao presente estudo, tais como: comprimento, frequência, séries temporais, parâmetro de relevância de um verbete, etc. Nosso trabalho propõe a análise de propriedades das séries temporais e da distribuição espacial de verbetes e, para tornar mais fluída a leitura, ele está estruturado como se segue.

Estrutura da dissertação

O estudo das séries temporais que podem ser extraídas de textos em linguagem natural possui abordagens distintas que tentam revelar a presença de correlação nas séries [27, 41]. No entanto, a análise das séries temporais de alguns dos parâmetros capazes de detectar palavras-chave e o estudo das correlações cruzadas das séries não possuem registro na literatura.

De maneira semelhante, avanços significativos foram feitos no estudo da distribuição espacial de verbetes e as propriedades que podem ser extraídas a partir dessas distribuições [34]. Porém o uso das distribuições espaciais num processo de produção de textos genéricos e o estudo das propriedades desses textos não está presente na literatura conhecida. Este trabalho, portanto, propõe uma nova análise para esses problemas e para tal objetivo, o texto foi dividido em 4 capítulos.

O capítulo 2 traz uma apresentação dos principais conceitos utilizados durante esse texto e que são relevantes ao estudo da linguística quantitativa. Serão discutidos temas como lei de Zipf, lei de Heaps, detecção de palavras-chave, entropia e informação.

O capítulo 3 apresenta o estudo das séries temporais extraídas de textos e a presença de autocorrelações e correlações cruzadas. Discutimos a existência de regimes distintos de correlação nos textos obtidos através das técnicas *Detrended Fluctuation Analysis* (DFA) e *Detrended Cross-Correlation Analysis* (DCCA).

O capítulo 4 é voltado ao estudo da distribuição espacial dos verbetes e sua aplicação na produção de textos genéricos. Nele é apresentado um algoritmo capaz de construir estruturas que possuem propriedades semelhantes aos textos em linguagem natural em que estudamos suas características e as comparamos com os padrões encontrados em linguagem escrita. Por fim, o capítulo 5 apresenta as conclusões do trabalho e perspectivas futuras.

Os próximos capítulos serão voltados para explicação e análise dos dados extraídos através das características dos textos. Essas características são estudadas através das propriedades de palavras e verbetes, portanto é fundamental a separação desses conceitos. Verbetes são as distintas combinações de caracteres em um texto, enquanto as palavras são as diversas ocorrências dessas combinações.

Conceitos

O estudo da linguística quantitativa está fundamentado na análise de propriedades que possam ser quantificadas nas línguas. As análises de George Zipf sobre a relação entre o número de ocorrências $n(k)$ de uma certa frequência k dos verbetes de um texto são consideradas o ponto inicial da área [9, 10]. A relação do tipo lei de potência, como a encontrada por Zipf, não foi a única estudada entre parâmetros que podem ser extraídos dos textos.

Nas duas décadas que seguiram a primeira publicação de Zipf surgiram diversos estudos com modelos capazes de quantificar propriedades relevantes dos textos e da língua. Dois desses modelos se destacam:

i) Hans Luhn propõe o uso da frequência como parâmetro classificatório para as palavras que devem compor o resumo de um texto [29];

ii) Claude Shannon propõe o uso de dois parâmetros: a entropia e a informação para quantificar propriedades em diversas escalas, variando de letras a uma língua [12].

Esses conceitos são centrais às análises que serão feitas nos capítulos que seguem, portanto esse capítulo é voltado a fundamentar e exemplificar esses conceitos utilizando, sempre que possível, as abordagens mais comuns e recentes na área.

Leis de Potência

Muitas quantidades empíricas se agrupam em torno de um valor típico [42]. Alturas e massas de pessoas selecionadas aleatoriamente em uma população são exemplos de quantidades que possuem uma escala característica. Apesar dessas quantidades variarem, suas distribuições apresentam valores muito pequenos de probabilidade ao se distanciar do valor típico (ou médio), ou seja, a variância é finita. Isso significa que a distribuição de variáveis desse tipo é bem caracterizada pela média e desvio padrão, indicando a existência de uma escala característica no sistema estudado.

Na física, sistemas que não possuem as características expostas acima encontram paralelo em fênomenos críticos. O interesse da criticalidade na física tem origem nos estudos dos pontos críticos em transições de fase [43]. Um exemplo desse tipo de sistema pode ser ilustrado pelo diagrama de fase para um fluído simples da Figura 2.1.

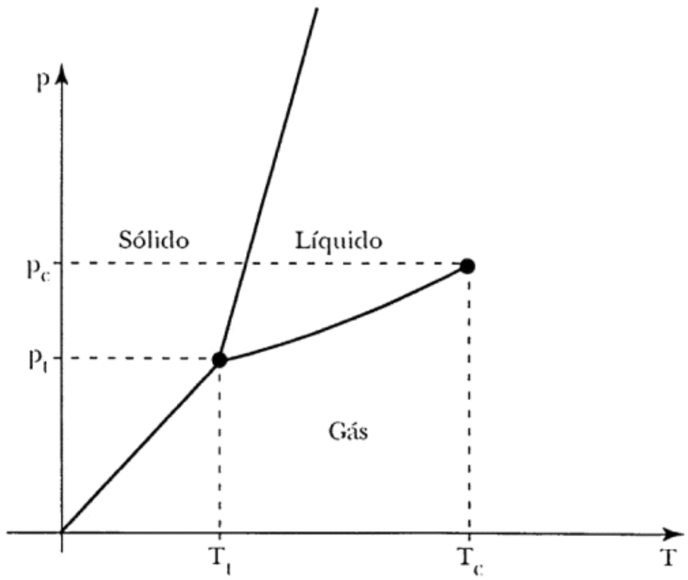


Figura 2.1: Diagrama de fase P-T para um fluído simples. O ponto (T_t, P_t) é o ponto triplo de coexistência das três fases e o ponto (T_c, P_c) é o ponto crítico. Figura extraída de [44].

Nele, vemos a coexistência de fases, delimitadas pelas linhas sólidas, o ponto triplo em (T_t, P_t) que indica a coexistência de três fases e o ponto crítico em (T_c, P_c) que possui as seguintes características:

i) ao se percorrer a curva de coexistência entre o gás e o líquido, quanto mais próximo o valor da temperatura T for do valor crítico T_c a diferença entre as respectivas densidades diminui e anula-se quando $T = T_c$;

ii) nas vizinhanças do ponto crítico, as derivadas de quantidades termodinâmicas (funcões resposta), tais como compressibilidade ou calor específico, podem apresentar um comportamento singular ou anômalo que caracterizam um estado crítico da matéria [44].

Nas proximidades do ponto crítico, portanto, as quantidades mencionadas possuem comportamentos divergentes descritos por leis de potência com expoentes característicos, denominados expoentes críticos. Os resultados empíricos indicam que sistemas físicos distintos podem compartilhar um mesmo conjunto de expoentes formando classes de universalidade [44]. Tipicamente, propriedades estruturais como a distribuição de tamanhos de domínios magnéticos podem apresentar distribuições do tipo lei de potência [45], caracterizando a ausência de uma escala típica de seus valores.

Quantidades descritas por esse tipo de distribuição não são bem caracterizadas pela média ou desvio padrão [46] e são representadas pela equação (2.1):

$$p(x) \sim x^{-\alpha} \tag{2.1}$$

em que x é a quantidade de interesse e α é o expoente, por vezes chamado de fator de escala. Esse tipo de distribuição é observada no estudo de intensidades de terremotos [47], de populações de cidades [48] e na linguística quantitativa para a análise da distribuição de frequências e comprimentos de vocabulários.

Lei de Zipf

O interesse de Zipf no estudo quantitativo da linguagem era voltado a compreender os processos da fala e seus desdobramentos na dinâmica social. Para isso, ele propôs que o princípio de menor esforço regia os processos de comunicação. Seu modelo assumia que havia uma competição entre o interesse do orador em reduzir o vocabulário ao mínimo, isso implicaria em utilizar uma única palavra com todos os significados possíveis, enquanto o ouvinte deseja que a quantidade de palavras e significados se igualasse.

De acordo com Zipf [10], essa competição era responsável pelas frequências e diversidade dos verbetes nos textos. Ao utilizar o livro *Ulysses* de James Joyce e definindo o parâmetro *rank* $r = 1$ para o verbete mais frequente, *rank* $r = 2$ para o segundo mais frequente e assim sucessivamente, ao multiplicar o *rank* pela frequência k, obtinha-se um valor constante C. Isso pode ser equacionado da seguinte maneira:

$$r \times k = C\,. \tag{2.2}$$

Ainda de acordo com Zipf, esse resultado sugeria a existência de um equilíbrio no vocabulário ao longo do texto. Em seu livro, Zipf ilustra esse resultado para *Ulysses* comparando-o com dados de textos jornalísticos (análise feita por R. C. Eldridge) e uma lei de potência de expoente $\alpha = 1$, o resultado apresentado em [10] é reproduzido em um gráfico duplo logarítmico (log-log) na Figura 2.2.

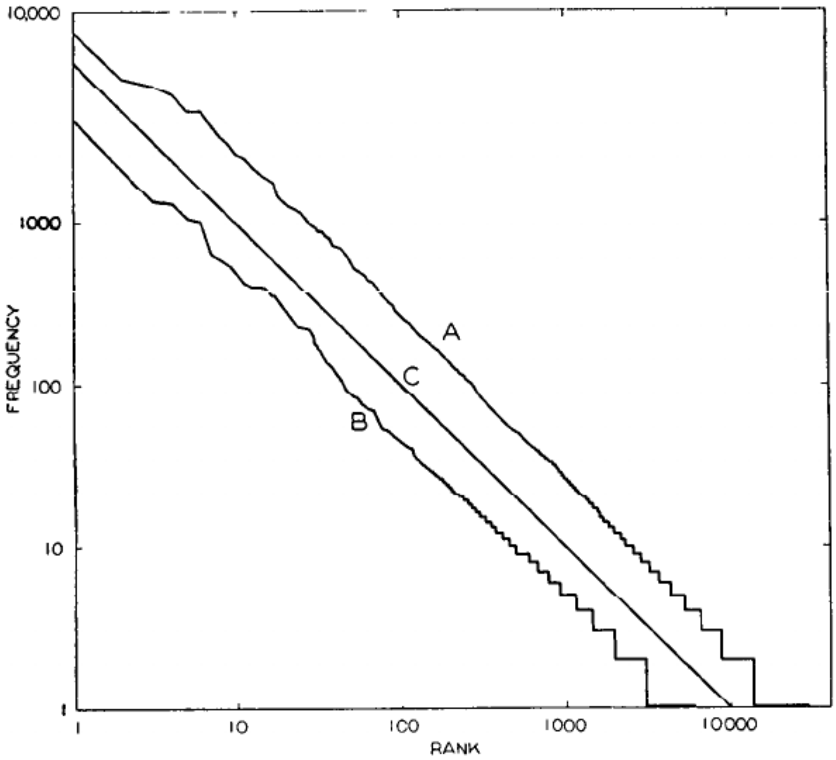


Figura 2.2: Distribuição do rank com a frequência. A curva A representa o texto *Ulysses*, B uma coleção de jornais americanos e C é a função r^{-1}. Figura extraída de [10].

As curvas A e B na Figura 2.2 sugerem a presença de uma lei de potência relacionando frequência e rank:

$$k(r) \sim r^{-z}\,. \tag{2.3}$$

Essa equação é conhecida como a formulação de *rank* da lei de Zipf. Outra formulação, essa apresentada em [9] e conhecida como frequencista, utiliza o número de verbetes n que ocorrem k vezes. Essa será a formulação utilizada no decorrer do texto e é descrita por:

$$n(k) \sim k^{-\beta}\,. \tag{2.4}$$

De maneira análoga, podemos ilustrar essa formulação utilizando novamente o livro *Ulysses* [49] através de um gráfico log-log como exibido na Figura 2.3. Vale a pena ressaltar que as versões do livro utilizadas aqui e por Zipf podem diferir, assim como o que se considera ser um verbete válido.

O expoente da lei de potência para essa formulação é $\beta = 2$ e isso pode ser

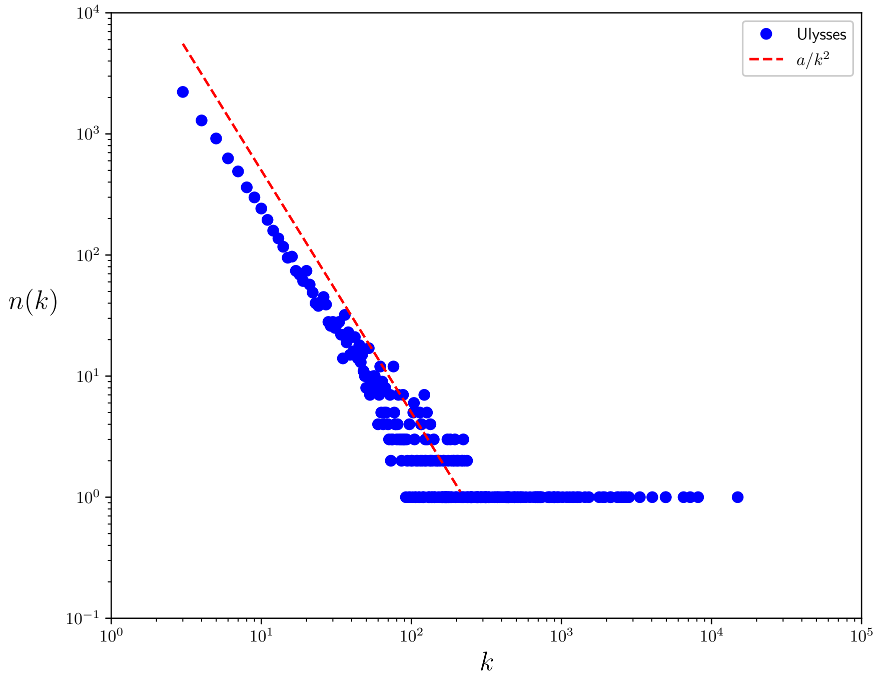


Figura 2.3: Distribuição do número de verbetes $n(k)$ com a frequência k para *Ulysses*. A linha tracejada em vermelho representa a função a/k^2 em que $a = 50.000$ foi escolhido tal que os valores de $n(k)$ fossem da mesma ordem de grandeza que aqueles encontrados no texto.

demonstrado se considerarmos que o *rank* r de um verbete que ocorre n vezes pode ser considerado como o número de verbetes que ocorrem pelo menos n vezes:

$$r = \sum_{k'=k}^{\infty} n(k') \approx \int_{k}^{\infty} n(k')dk' \tag{2.5}$$

utilizando as equações (2.3) e (2.4), resolvendo a integral e isolando os termos que contêm k, descobrimos que:

$$z = \frac{1}{\beta - 1} \tag{2.6}$$

e é através dessa relação que estabelecemos o valor de β utilizado na Figura 2.3 para construir a curva tracejada. Apesar da distribuição sugerida por Zipf ser observada em textos em linguagem natural, a justificativa dada em seu trabalho para a origem dos resultados não é a única encontrada na literatura [50].

O fato de um processo bastante intricado como a escrita originar uma distribuição de frequências matematicamente simples e invariante em relação a textos, língua e *corpus* despertou dúvidas sobre a validade desse estimador [51]. Muitas abordagens para demostrar a lei de Zipf através de primeiros princípios foram criadas, entre elas vale a pena citar a optimização da entropia [13], processos estocásticos multiplicativos [46], reúso preferencial [52] e optimização na comunicação [9, 10, 14].

Os processos sugeridos, em geral, são voltados a ter como produto final a lei de Zipf, pouco tem sido feito para verificar a validade das hipóteses utilizadas nos modelos, problema também comum aos sistemas que apresentam propriedades com distribuições do tipo lei de potência [53]. Porém, não é somente a veracidade da origem da lei de Zipf que costuma ser questionada na literatura, textos aleatórios também apresentam a lei de Zipf [54, 55] utilizando a formulação de *rank*, indicando a possibilidade da lei se tratar somente de um artefato estatístico [50].

Ferrer i Cancho e Elvevåg [56] mostraram recentemente que a estatística detalhada de textos aleatórios não produz distribuições compatíveis com textos humanos. Experimentos comportamentais [50] em que os participantes precisavam redigir um breve conto envolvendo personagens fictícios que teriam sua classificação de relevância dada pelo autor mostraram um comportamento zipfiano do *rank* com a frequência do verbete.

Esses resultados sugerem que não há consenso sobre a relevância da lei de Zipf na língua ou de qual processo a origina, no entanto, eles indicam que o seu uso como ferramenta para o estudo de textos em linguagem natural é válido. E como ferramenta de análise da distribuição de frequências em texto, a lei de Zipf será posteriormente utilizada, assim como a lei de Herdan-Heaps que nos fornece uma relação entre o tamanho do vocabulário do texto e seu tamanho.

Lei de Herdan-Heaps

Em 1960, Gustaf Herdan propôs que o crescimento do vocabulário V de um texto segue uma lei de potência com o tamanho do texto T, lei que seria sistematizada *a posteriori* por Harold Heaps [18]. Para justificar tal modelo, Herdan [57] utiliza uma única suposição: a taxa relativa de aumento de novos verbetes (dV/V) é proporcional à taxa relativa de aumento do número de palavras (dT/T):

$$\frac{dV}{V} = \lambda \frac{dT}{T} \tag{2.7}$$

resolvendo a equação acima, chega-se a lei de Herdan-Heaps:

$$V(T) \sim T^{\lambda} \tag{2.8}$$

os valores de λ se encontram entre 0 e 1 [18, 57, 58], em que 1 representaria textos com vocabulário tão grande quanto o texto ($V = T$). O resultado encontrado para capítulos de dois romances: *Eugen Onegin* e *The Captain's Daughter* ambos de Alexander Pushkin está ilustrado na Figura 2.4. Para gerar o gráfico log-log apresentado na Figura 2.4, Herdan considerou o vocabulário acumulado até certo comprimento do capítulo utilizado, medindo o vocabulário acumulado a cada 100 palavras.

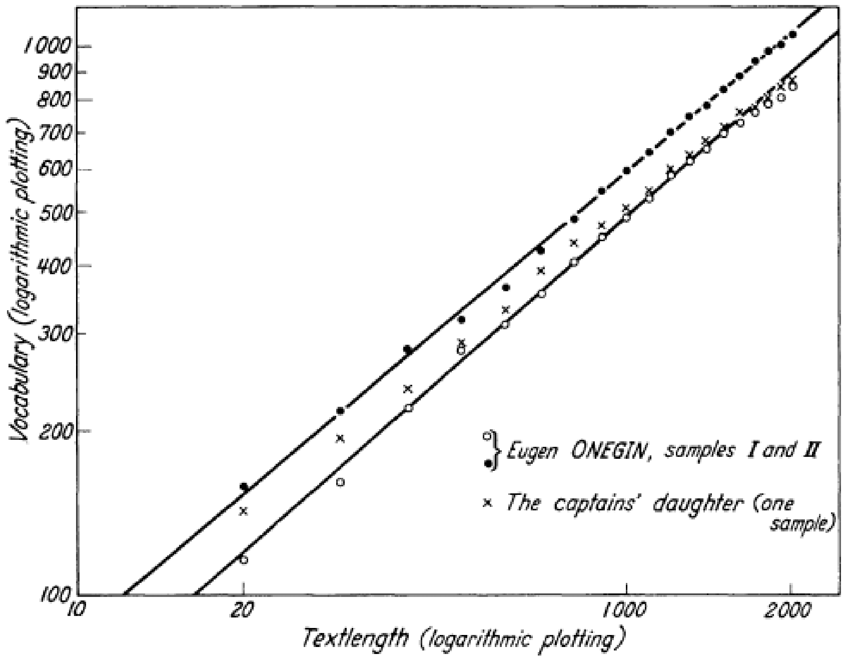


Figura 2.4: Vocabulário acumulado em função do comprimento corrente do texto para capítulos de *Eugen Onegin* e *The Captain's Daughter*. Figura extraída de [57].

Em 2013, Pola [59] utilizou a metodologia de Herdan no livro *War and Peace* de Leo Tolstoy em quatro idiomas e mostrou que a Lei de Herdan-Heaps é melhor descrita por uma lei de potência com dois regimes, sendo inalterada pela tradução do texto, contrariando os resultados de Gelbukh e Sidorov [60].

Gerlach e Altmann [22] também mostraram a existência desses dos dois regimes da lei de potência para um *corpus* de 5,2 milhões de livros publicados ao longo dos últimos cinco séculos e digitalizados pelo Google. No entanto, os autores consideraram o vocabulário e o tamanho de cada livro como parâmetros para se calcular os expoentes da lei de Herdan-Heaps.

Seguindo essa metodologia, porém utilizando um único regime para a lei de potência, Rolim [38] determinou que o expoente médio de Herdan-Heaps é $\lambda = 0{,}71 \pm 0{,}05$ para o *corpora* que será compartilhado também por nossas análises. Esse valor de λ assegura que o conjunto de textos selecionados são representativos em relação às propriedades estatísticas da linguagem escrita [60].

Utilizando o mesmo *corpora* e metodologia, ilustramos a lei de Herdan-Heaps para três línguas de família linguísticas distintas na Figura 2.5. Extraímos também os valores para os expoentes de Zipf dos textos literários β e de Herdan-Heaps λ para os 485 textos, os resultados são mostrados na Tabela 2.1. Esses resultados são semelhantes aos encontrado por Rolim [38] que estendeu a hipótese [60] do uso dos valores dos expoentes de Herdan-Heaps para a classificação de grupos linguísticos.

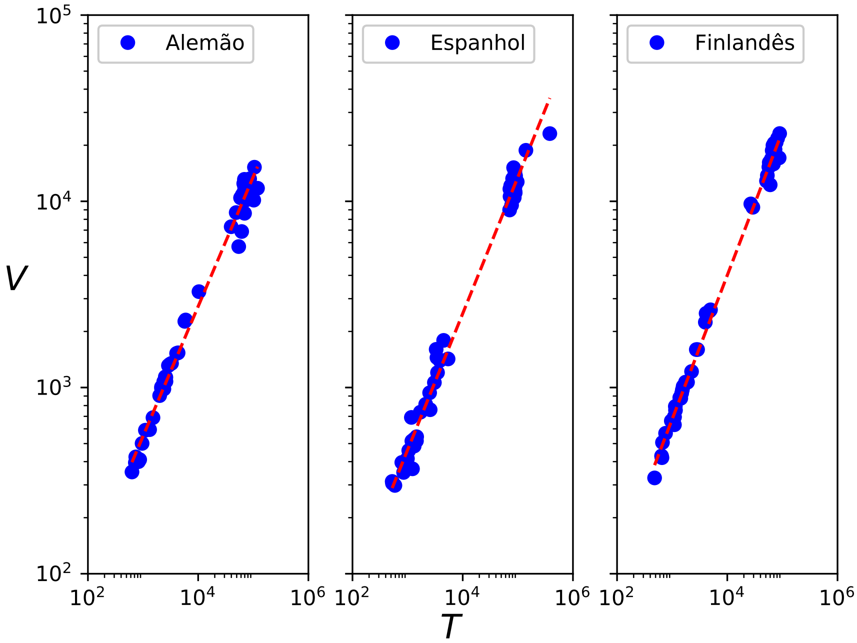


Figura 2.5: Gráficos log-log do vocabulário V como função do tamanho do texto T para 3 línguas de grupos linguísticos distintos e suas respectivas regressões do tipo lei de potência.

Algumas observações sobre os métodos e escolhas do *corpus* nas análises são necessárias:

i) para determinar o expoente médio de Zipf para cada língua, os textos da wikipedia não foram considerados devido ao pequeno no número de verbetes nos textos o que torna sua análise pouco representativa do ponto de vista estatístico;

ii) também assumimos que existe uma frequência k_{zipf} em que occore a menor distância entre o vetor $(k, n(k))$ e a origem. Essa é a frequência limite em que a lei de Zipf apresenta resultados representativos. Esse parâmetro foi adotado para termos um

	Língua	β	λ
1	Alemão	$1{,}96 \pm 0{,}09$	$0{,}70 \pm 0{,}01$
2	Dinamarquês	$1{,}96 \pm 0{,}10$	$0{,}69 \pm 0{,}01$
3	Espanhol	$2{,}02 \pm 0{,}10$	$0{,}73 \pm 0{,}01$
4	Finlandês	$2{,}08 \pm 0{,}08$	$0{,}77 \pm 0{,}01$
5	Francês	$1{,}96 \pm 0{,}11$	$0{,}70 \pm 0{,}01$
6	Húngaro	$2{,}18 \pm 0{,}12$	$0{,}78 \pm 0{,}01$
7	Inglês	$1{,}74 \pm 0{,}12$	$0{,}59 \pm 0{,}02$
8	Italiano	$2{,}02 \pm 0{,}10$	$0{,}73 \pm 0{,}01$
9	Português	$2{,}02 \pm 0{,}12$	$0{,}73 \pm 0{,}01$
10	Sueco	$1{,}99 \pm 0{,}11$	$0{,}70 \pm 0{,}01$
	Média	$2{,}00 \pm 0{,}11$	$0{,}71 \pm 0{,}01$

Tabela 2.1: Tabela com os valores médios dos expoentes de Zipf dos textos literários e de Herdan-Heaps (todos os textos) para as línguas analisadas, obtidos a partir das equações (2.4) e (2.8), respectivamente.

valor limitante da frequência, acima dele a curva passa a exibir um ruído na sua extremidade levando a valores espúrios para o coeficiente de Zipf;

iii) o espectro de tamanhos dos textos utilizados na presente dissertação $10^2 < T < 10^5$ corresponde a valores intermediários em relação à literatura [22, 37]. Uma vez que os tamanhos variam entre 4 décadas distintas, os resultados encontrados para o expoente de Herdan-Heaps tornam-se mais robustos, ou seja, há uma redução no erro do coeficiente da regressão.

O fato dos expoentes de Zipf e Herdan-Heaps possuírem valores bem definidos motivou a criação de modelos que fossem capazes de gerar ambas as distribuições. Simon propôs em 1955 [52] um modelo que utiliza processos estocásticos multiplicativos capaz de exibir o comportamento zipfiano. Em 2013, Gerlach e Altmann [22] propuseram uma generalização para o modelo de Simon. Utilizando ambos os modelos Pola [59] criou textos artificais que exibiam as duas distribuições com expoentes $z = 0{,}96$ e $\lambda = 0{,}99$.

Outras duas abordagens relevantes são os sistemas descritos pela fórmula de Turing [61] e efeitos de dicionários finitos [62, 63]. Esses modelos mostram que sistemas que possuem diversas classes de objetos são capazes de reproduzir ambas as leis. Portanto, um modelo para a distribuição de verbetes num texto deve ser capaz de reproduzir essas propriedades, assim como apresentar uma distribuição de parâmetros de relevância, entropia e outras propriedades semelhantes a de um texto real.

Detecção de palavras relevantes

Os primeiros estudos quantitativos da linguagem escrita adotaram abordagens frequencistas [9, 10, 13] e não foi diferente para o primeiro modelo proposto para a extração de palavras-chave em um texto. Em 1958, Luhn [29] propôs um método para criação automática de resumos de artigos utilizando as sentenças do texto e frequências dos verbetes.

Modelo de Luhn

A escolha de quais sentenças estariam em um resumo seria feita a partir de seus conteúdos informacionais, uma vez que todas as sentenças fossem comparadas e classificadas através de uma medida de informação. Para a quantificação da informação, Luhn sugeriu o uso da frequência dos verbetes. Em sua proposta, os verbetes com maiores e menores frequências seriam excluídos, criando uma região de verbetes de frequência intermediária que poderiam ser utilizados.

A justificativa para tal escolha se encontra no fato de que verbetes raros ou os muito frequentes apresentam pouca relevância em textos. A tabela 2.2 traz uma seleção de verbetes extraídos do livro *On the origin of species* de Charles Darwin (referência IN-20 no Apêndice A), nela é possível observar o comportamento descrito por Luhn.

Verbete	**Frequência**	***Rank***	**Verbete**	**Frequência**	***Rank***
the	10.199	1	*formations*	90	225
a	2.496	6	*always*	90	226
as	1.593	10	*absence*	20	905
species	1.488	11	*whale*	5	2.345
hybrids	119	169	*abnormally*	1	7.422

Tabela 2.2: Tabela com uma escolha de verbetes do texto *On the origin of species* apresentando seus respectivos frequências e *rank*. O maior valor do *rank* para esse texto é $r_{max} = 7.475$.

O leitor familiarizado com o tópico do qual trata o livro de Darwin, ao observar a Tabela 2.2 atentamente, verificará que a proposta de Luhn pode levar à seleção de verbetes não importantes como *always* e exclusão de verbetes como *species*. Esse modelo mostra uma dependência muito grande de quais frequências mínimas e máximas são adotadas para definir a região de frequências dos verbetes relevantes. Por não levar em consideração a distribuição espacial dos verbetes no texto, o modelo de Luhn é pouco eficiente para detectar correlações espaciais.

Correlação espacial e unidade fundamental da escrita

O tipo de modelo frequencistas proposto por Luhn não é capaz de revelar como se dá a distribuição espacial dos verbetes no texto. Essa afirmação se torna mais evidente quando analisamos textos embaralhados, ou seja, textos que tiveram todas as suas palavras trocadas aleatoriamente de posição. Nesses textos, as frequências dos verbetes seriam preservadas, apesar de sua distribuição espacial ter sido profundamente alterada (a depender do número de embaralhamentos, teríamos um texto em que todos os verbetes estão uniformemente distribuídos [38]).

Uma vez considerado o exemplo acima, a estrutura organizacional das palavras se torna fundamental para a detecção de palavras relevantes. Os estudos na área de distribuição espacial de símbolos revelaram a presença de correlações de longo alcance em sequências de nucleotídeos [64], em textos [65] e letras e sentenças [66].

Ebeling e Neiman [66] analisaram a contribuição de símbolos, palavras e sentenças para a estrutura de textos em inglês. Para tal, eles adotaram três formas de embaralhar o texto obtendo resultados distintos:

i) os caracteres válidos foram embaralhados (as 26 letras do alfabeto, os sinais de pontuação "." e ",", parênteses "(" e ")", o símbolo "#" e o espaço em branco " "). Os autores afirmaram que todas as correlações foram destruídas no texto embaralhado;

ii) as palavras foram embaralhadas (palavras são qualquer conjunto de caracteres válidos entre dois espaços em branco);

iii) as sentenças foram embaralhadas (qualquer conjunto de caracteres válidos entre dois pontos finais). Por construção, tanto a forma *ii* como a presente não devem apresentar correlações em escalas maiores do que as respectivas sentenças.

Essas formas de embaralhar o texto sugerem que as correlações de ordens superiores são baseadas na estrutura do texto e nas relações semânticas, ambas ligadas a distribuição de letras, palavras e sentenças.

Stephens e Bialek [67] consideram as palavras como uma rede de letras interagentes que reduzem o número de configurações acessíveis pela rede. Utilizando um modelo de máxima entropia, eles mostram que as correlações entre as letras geram uma aproximação da estatística completa das palavras. Seguindo a proposta que as palavras são as unidades fundamentais da escrita, Montemurro e Pury [68] analisaram a estrutura fractal de grandes registros de linguagem escrita, mapeando os textos utilizados em séries temporais. Os resultados encontrados por eles apontam que além de correlação de curto alcance que é produto da regras sintáticas da língua atuando no comprimento de sentenças, estruturas com correlações de longo alcance surgem em grandes amostras de linguagem escrita.

A combinação desses resultados [66–68] indicam que a distribuição de letras influencia as palavras que compõem a língua. Essas, por sua vez, criam as estruturas responsáveis pelas correlações em escalas que variam de sentenças ao comprimento do texto. Portanto, no presente trabalho, consideraremos a palavra como unidade fundamental da escrita, dessa maneira todos os parâmetros utilizados serão definidos a partir dessa unidade.

Intermitência

Utilizando a palavra como unidade fundamental, Ortuño e colaboradores [30] propuseram que através da análise da distribuição espacial dos verbetes seria possível criar um mecanismo de detecção de palavras-chave em textos, sem o uso de qualquer informação disponível *a priori*. A abordagem apresentada consistia em determinar a distribuição $p(x)$ das distâncias entre sucessivas ocorrências de um verbete.

Essa abordagem é comum no estudo de estatística de níveis do espectro de sistemas quântico desordenados [69]. Em geral, a metodologia utilizada nessa área é construir o espectro dos espaçamentos consecutivos s_n, entre os níveis E_1, E_2, etc, em que s_n é dado por:

$$s_n = \frac{(E_{n+1} - E_n)}{\bar{d}(E)}$$

em que $\bar{d}(E)$ é o espaçamento médio entre os níveis até a energia E [70]. A partir dos valores de s_n, é possível analisar seu histograma e compreender as características da distribuição de níveis. Em sua aplicação linguística, uma vez extraído o conjunto das distâncias $\{x_i\}$, define-se $p(x)$ como a frequência relativa de ocorrência de uma determinada separação $x \in \mathbb{N}$ e $P_1(x) = \sum_{x'=1}^{x} p(x')$ como a distribuição acumulada.

A dependência com a frequência pode ser eliminada se normalizarmos as separações pela média $\bar{x}$, definindo a distância normalizada s como:

$$s = \frac{x}{\bar{x}} \tag{2.9}$$

repetindo esse processo para todos os verbetes, criamos o conjunto $\{s_i\}$ de distâncias normalizadas. Utilizando essa métrica, se as palavras estiverem aleatoriamente distribuídas, $P_1(s)$ seguirá uma distribuição de Poisson no limite de altíssimas frequências:

$$P_1(s) = 1 - e^{-s}\,. \tag{2.10}$$

A Figura 2.6, extraída de [30], apresenta a distribuição acumulada $P_1(s)$ para quatro verbetes distintos do livro *The Quijote* de Miguel de Cervantes e a distribuição dada pela equação 2.10. Os verbetes escolhidos ("Quijote", "Sancho", "the" e "and") representam bem esse tipo de abordagem, os dois primeiros citados são relevantes no texto, mostrando distribuições acumuladas que se distanciam da curva poissoniana em distâncias relativas intermediárias, enquanto os dois últimos mostram regimes muito próximos do comportamento poissoniano. Esse fato pode ser interpretado como se os primeiros verbetes fossem utilizados somente em contextos específicos e suas ocorrências concentradas em regiões bem delimitadas, enquanto os últimos estariam distribuídos aleatoriamente ao longo do texto.

O cálculo de $P_1(s)$ para cada um dos verbetes do texto possui um alto custo computacional. Para contornar esse problema, Ortunño propôs o uso do desvio padrão de s, frequentemente nomeado de intermitência [38] (o equivalente ao desvio padrão relativo à média de x):

$$\sigma = \sqrt{\bar{s^2} - \bar{s}^2}\,. \tag{2.11}$$

Apesar da distribuição acumulada e o desvio padrão serem parâmetros distintos, ambos carregam informação sobre a distribuição associada, ou seja, através de ambos conseguimos acessar características da distribuição. Portanto o uso de σ como métrica para o

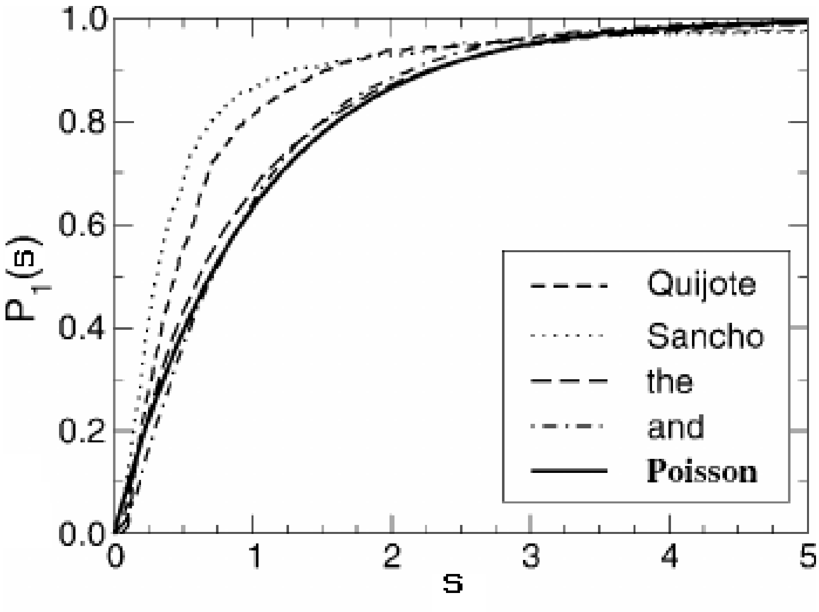


Figura 2.6: Distribuição acumulada das distâncias normalizadas entre as sucessivas ocorrências para quatro verbetes do texto *The Quijote*. A linha sólida corresponde a distribuição acumulada para a distribuição de Poisson. Figura adaptada de [30].

estudo da distribuição espacial seria mais razoável se considerado o custo computacional associado à extração de ambos os parâmetros.

Ortuño defende o uso da intermitência argumentando que ele apresenta um crescimento rápido conforme aumenta a heterogeneidade da distribuição de espaçamentos. Ao mesmo tempo em que ele é capaz de capturar a diferença entre diferentes regimes da distribuição: regime poissoniano ($\sigma = 1$), efeitos de aglomeração ($\sigma > 1$) e de repulsão ($\sigma < 1$).

Essa métrica ajuda a compreender os espectros das posições observados na Figura 2.7 para dois verbetes em IN-20. Nela temos o espectros das posições absolutas de dois verbetes do texto (o intervalo até a posição 100.000 foi considerado para uma melhor visualização do espectro, o tamanho do texto é $T = 152.383$). Apesar da pequena diferença entre as frequências dos verbetes *species* ($k = 1.488$) e *as* ($k = 1.593$), eles apresentam distribuições espaciais distintas, enquanto o primeiro possui aglomerações em certas regiões, o segundo possui uma distribuição aproximadamente uniforme. Isso implica em graus de relevância diferentes para esses verbetes e utilizando a intermitência somos capazes de quantificar suas relevâncias.

A Tabela 2.3 traz os valores de σ para *species*, *as* e os 10 verbetes de maiores

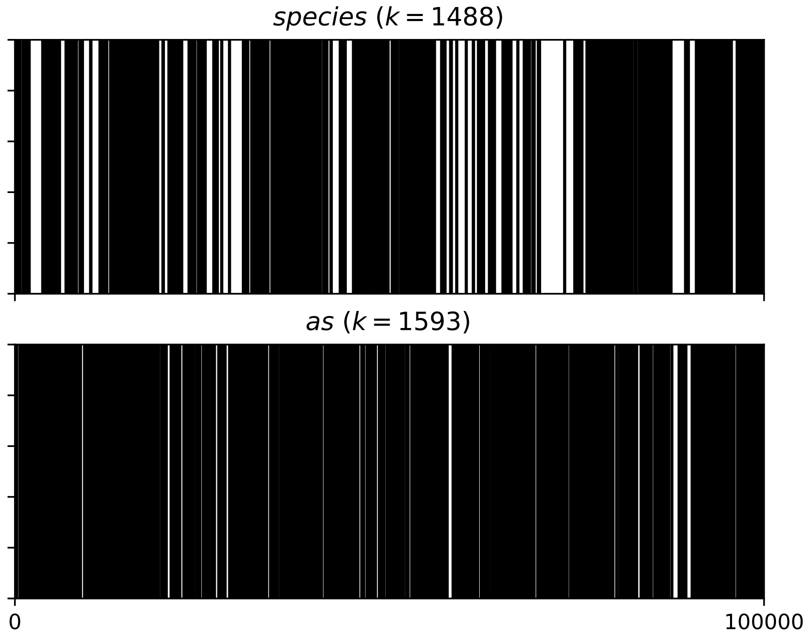


Figura 2.7: Espectros das posições absolutas para os verbetes *species* e *as* em IN-20. Os espectros são mostrados do início do livro à posição 100.000 ($T = 152.383$).

Ranking σ	**Verbete**	**Frequência**	σ	**Ranking σ**	**Verbete**	**Frequência**	σ
1	*formations*	90	5,41	6	*workers*	27	4,48
2	*bees*	65	5,33	7	*slaves*	34	4,45
3	*sterility*	78	4,99	8	*ants*	28	4,29
4	*hybrids*	119	4,96	9	*diagram*	41	4,22
5	*instincts*	72	4,80	10	*instinct*	58	4,20
304	*species*	1.488	2,00	1.557	*as*	1.593	1,17

Tabela 2.3: Tabela com os 10 verbetes de maiores intermitência σ e os verbetes *species* e *as* no texto IN-20.

σ para IN-20. Pode-se afirmar que 9 dos verbetes presentes entre os 10 de maiores intermitência são relevantes no texto (excetuando-se *diagram*). A tabela ainda traz outras duas informações importantes: *species* é uma palavra bastante relevante no texto, apesar do ranking de σ apontá-la na posição 304 e *as* possui $\sigma_{as} = 1{,}17$ indicando um regime próximo ao poissoniano em sua distribuição de distâncias.

Podemos, a partir do σ, construir um gráfico como na Figura 2.8 que ilustra a relação entre intermitência e frequência de todos os verbetes em preto, assim como os verbetes presentes na Tabela 2.3 são diferenciados com cores diferentes. A forma dessa distribuição da intermitência com a frequência é comum a todos os textos no *corpora* e se

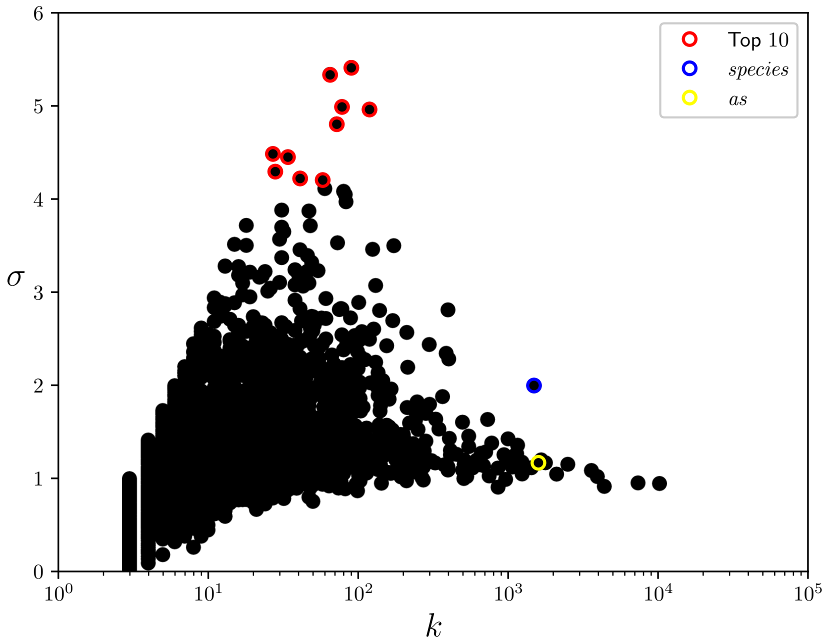


Figura 2.8: Gráfico log-linear da intermitência σ e o número de ocorrências k para os verbetes com $k \geq 3$ em IN-20. Os círculos em vermelho destacam os verbetes com os maiores valores de σ, enquanto azul e amarelo indicam os verbetes *species* e *as*, respectivamente.

torna mais evidente quanto maior for o texto como discutido por Rolim [38]. Esse gráfico revela algumas características peculiares da distribuição de σ com k:

i) utilizando o valor de σ como o parâmetro para a determinação de palavras-chave, vemos que existem verbetes com $\sigma > 1$ entre $10^0 < k < 10^3$, porém os verbetes de maiores intermitência estão contidos, em sua grande parte, na região em que $10 < k < 10^2$;

ii) σ nos revela as características particulares da distribuição espacial dos verbetes e pode ser utilizado não somente como ferramenta de extração de palavras-chave, mas também como parâmetro para avaliar as propriedades da distribuição de distâncias dos verbetes;

iii) existe uma região limitante superior que se apresenta em todos os textos utilizados no *corpora* e sugere a existência de uma distribuição super-poissoniana de distâncias ($\sigma > 1$) que gere esse caso limitante [38];

iv) a região inferior também possui contornos limitantes que pode ser identificado como um regime sub-poissoniano ($\sigma < 1$) em baixas frequências que se aproxima de uma distribuição poissoniana ($\sigma = 1$) para altas frequências [38]. Essas duas regiões limitantes serão adequadamente analisadas na seção 4.1.

Ortuño e colaboradores [30] estenderam o uso da intermitência para sequências de DNA na busca de "palavras relevantes" compostas por nucleotídeos. Outras abordagens utilizando essa métrica foram feitas para o estudo do intervalo de recorrência entre verbetes para classificação textual [71], identificação autoral [71, 72], análises de partes funcionais do genoma [73] e da distribuição espacial de verbetes em linguagem escrita [38].

No entanto, como apontado por Zhou e Slater [31], essa abordagem possui limitações na extração de palavras-chave. Eles citam a dificuldade para identificar verbetes relevantes em regimes de baixas e altas frequências (como é o caso de *species* mostrado na Tabela 2.3 em IN-20). Outra limitação do modelo está no regime de baixas frequências, quando há a mudança de uma única posição de uma das ocorrências, presenciamos uma mudança significativa no valor da intermitência, evidenciando uma instabilidade da métrica.

Na tentativa de aprimorar a proposta de Ortuño e colaboradores, eles propuseram incluir as primeiras e últimas palavras do texto na contagem de distâncias e os resultados encontrados mostram que a intermitência se tornou mais eficaz para detectar palavras relevantes nos regimes de altas e baixas frequências [31]. A justificativa para esse modelo seria que as palavras comuns, por possuírem uma distribuição aproximadamente poissoniana ($\sigma \sim 1$), deveriam ser encontradas em praticamente todas as regiões do texto e as extremidades não seriam excessão. Porém, as palavras relevantes seriam raramente encontradas nas extremidades.

Também baseado na métrica da intermitência, Carpena e colaboradores [32] sugeriram uma abordagem que combina a frequência de cada verbete com a estrutura de agrupamento das palavras. Eles apontaram que o comportamento poissoniano seria observado somente para uma distribuição contínua de distâncias o que não ocorre com as palavras que são separadas por distâncias inteiras. Então, faz-se necessário o uso da distribuição geométrica (o equivalente discreto da distribuição de Poisson):

$$P_{geo}(x) = p(1-p)^{x-1} \tag{2.12}$$

em que $p = k/T$ é a probabilidade de encontrar o verbete no texto, k é sua frequência, T o comprimento do texto e x é a distância entre as sucessivas ocorrências.

A distribuição geométrica possui média e variância dadas por:

$$\bar{x} = 1/p$$
$$\bar{x^2} - \bar{x}^2 = \frac{1-p}{p^2}$$

logo sua intermitência em relação à média σ_{geo} é:

$$\sigma_{geo} = \sqrt{1-p}\,. \tag{2.13}$$

Os autores afirmam que a distribuição geométrica seria um bom modelo para descrever o comportamento de palavras irrelevantes e sugerem o uso de um desvio padrão normalizado por σ_{geo} (chamado de medida de agrupamento):

$$\sigma_{nor} = \frac{\sigma}{\sigma_{geo}} = \frac{\sigma}{\sqrt{1-p}}\,. \tag{2.14}$$

A distribuição de σ_{nor} ($P(\sigma_{nor})$) mostra uma dependência forte com k, consequência da definição dada na equação (2.14) [32], indicando uma melhor detecção de palavras-chave nos regimes de baixas e altas frequências.

Apesar das abordagens existentes que aprimoram a intermitência (σ) proposto por Ortuño e colaboradores [30] como métrica de extração de verbetes relevantes, manteremos seu uso nesse trabalho para a análise da autocorrelação e correlação cruzada de séries temporais e no estudo da distribuição espacial dos verbetes. Defendemos seu uso pelas razões que seguem:

1. na análise de séries temporais, discutida Capítulo 3, do nosso estudo será voltado a compreender as possíveis correlações de curto e longo alcance das séries. Assim, o uso da intermitência como métrica de relevância de um verbete não deve alterar qualitativamente os resultados, uma vez que ele é utilizado somente como um rótulo que indica a relevância daquele verbete;

2. o estudo da distribuição espacial, discutido no Capítulo 4, será voltado a compreender quais distribuições reproduzem as propriedades encontradas nos textos como a lei de Zipf, lei de Heaps, regiões limitantes em σ e na entropia estrutural (ver seção 2.3). Como o σ é um parâmetro capaz de revelar as propriedades da distribuição de distâncias dos verbetes, ele possui a robustez necessária para as análises feitas;

3. a determinação de σ para cada um dos verbetes do texto apresenta baixo custo computacional, em relação às outras alternativas, sendo essa a principal justificativa prática para seu uso nesse trabalho.

Entropia

A segunda metade do século XIX e o período que antecedeu a Primeira Guerra Mundial foram marcados, do ponto de vista tecnológico, pela Segunda Revolução Industrial, uma época de amplo desenvolvimento socio-tecnológico na Europa e Estados Unidos. Nesse período, a humanidade presenciou a criação de máquinas de vapor mais eficientes e sua rápida substituição pelo maquinário elétrico, motores a vapor e de combustão interna [74]. No entanto, os fenômenos envolvidos nessa nova tecnologia ainda eram pouco compreendidos por volta de 1850.

Entropia na termodinâmica

Motivados pela fenomenologia dos processos térmicos e pela tecnologia já utilizada, começaram a surgir os trabalhos que fundamentam a termodinâmica. Um dos objetos de estudo dessa área é a irreversibilidade dos processos naturais, estudada através de fundamentos termodinâmicos pela primeira vez por Sadi Carnot [75]. A partir desses fundamentos introduzidos por Carnot, William Thomson (o Lord Kelvin) e Rudolf Clausius formalizaram os estudos em irreversibilidade. Clausius, então, propôs a existência de uma grandeza S que possui variação dada por:

$$dS \equiv \frac{đQ}{T} \tag{2.15}$$

em que Q seria a quantidade de calor fornecida a um sistema de temperatura T. Essa equação é conhecida como a formulação de Clausius da Segunda Lei da Termodinâmica para sistemas fechados em equilíbrio térmico e sua igualdade é verificada em processos reversíveis.

Essa grandeza, segundo Clausius [76], está ligada ao conteúdo transformacional de um sistema, portanto ele a nomeou como entropia (palavra de origem grega que significa "transformação"). Posteriormente Ludwig Bolzmann estendeu o conceito de entropia para sistemas fora do equílibrio utilizando o número de microestados Ω acessíveis para as partículas em um macroestrado, definindo assim a equação (2.16):

$$S = -k_B \ln \Omega \tag{2.16}$$

em que k_B é uma constante que ajuda a definir a unidade de entropia e é conhecida como constante de Boltzmann.

A formulação de Boltzmann supõe que todos os microestados acessíveis são equiprováveis [77]. Ao estudar sistemas com microestrados com probabilidades distintas, Josiah W. Gibbs desenvolveu uma expressão mais generalizada para a entropia [78]:

$$S = -k_B \sum_i p_i \ln p_i \tag{2.17}$$

em que k_B é a constante de Boltzmann e p_i é a probabilidade de um sistema está no microestrado i. É importante notar que se $p_i = \frac{1}{\Omega}$ em que Ω é novamente o número de microestados acessíveis, a equação (2.17), com limite superior da soma igual a Ω, se resume à equação (2.16).

Ambas as equações (2.16) e (2.17) possuem validades que excedem a fenomenologia termodinâmica (podemos, por exemplo, utilizar a equação (2.16) para descrever modelos de urna [79]). Tal fato implica em cautela na hora de fazer a associação entre a entropia da fenomenologia termodinâmica, como proposta por Clausius, e as formulações de Boltzmann e Gibbs.

No entanto, até o ponto discutido aqui, o significado da entropia ainda está associado à transformação e reversibilidade, ela mede o quanto um sistema foi alterado ao passar por um processo físico em relação a seu estado inicial, uma vez que assim como pressão, volume e temperatura, a entropia também é uma variável de estado. Uma nova associação foi feita por Paul e Tatyana Ehrenfest através de um modelo de urna, eles propuseram que a entropia estava associada à ordem (ou desordem) das partículas no sistema. Sendo o significado de ordem expresso pelo número de maneiras distintas na qual o sistema possa ser configurado preservando suas características exteriores [80]. Essa interpretação é a que nos ajudará a compreender a utilidade da entropia no estudo da linguagem natural.

Entropia de Shannon

O significado dado pelos Ehrenfest à entropia, o modelo de Gibbs e a busca pela definição do que seria informação guiaram Claude Shannon no desenvolvimento de uma teoria da informação apresentada em 1948. Em seu artigo [11], Shannon demonstra a existência de uma grandeza H em que dada a distribuição de probabilidades $p(r_i)$ $(i = 1, 2, ..., N)$:

$$H = -K_S \sum_{i=1}^{N} p(r_i) \log_b [p(r_i)] \tag{2.18}$$

em que N é o número de eventos disjuntos e K_S é uma constante positiva.

De acordo com Shannon, K_S é somente uma constante utilizada na escolha de uma unidade de medida para H, portanto podemos escolher $K_S = 1$ e a equação acima se torna:

$$H = -\sum_{i=1}^{N} p(r_i) \log_b [p(r_i)] \ . \tag{2.19}$$

Para o caso em que escolhemos a base do logaritmo $b = 2$ da equação (2.19), Shannon nomeou a unidade para a entropia de *bit*. Sistemas que possuem dois estados possíveis equiprováveis (como um relé ou uma moeda não viciada) são capazes de armazenar 1 *bit* de informação.

A função H satisfaz as seguintes propriedades comuns à definição de incerteza associada a uma distribuição $p(r_i)$:

1) H é uma função contínua de $p(r_i)$;

2) se todas as probabilidades $p(r_i) = 1/N$, H será uma função monotonicamente crescente com N;

3) se a probabilidade de um evento $p(r_i)$ pode ser tomada como a probabilidade conjunta de dois eventos independentes sucessivos r_i^I e r_i^{II}, H para $p(r_i)$ pode ser escrito como a soma dos valores da função H de cada um dos eventos: $H = H_I + H_{II}$.

A essa grandeza (H) foi dado o nome de entropia da distribuição de probabilidade $p(r_i)$ e posteriormente chamada de entropia de Shannon. É importante mencionar que apesar da semelhança entre as equações (2.17) e (2.19), os seus significados são bastante distintos. A primeira se refere a uma grandeza associada ao estado organizacional de um sistema de partículas que compartilham o mesmo macroestado, mas que possuem microestados com probabilidades distintas de ocorrências. A última é uma propriedade de uma distribuição de probabilidades e esse fato justifica seu uso no estudo de diversos sistemas, em especial no estudo quantitativo da linguagem escrita [81].

Porém existe uma conexão entre os conceitos feita através da interpretação de Ehrenfest para a entropia e a grandeza que Shannon definiu como informação I. Shannon propôs que a taxa de transmissão de informação R seria dado pela diferença entre a entropia da fonte de de ruído H_0 (sendo essa a maior entropia possível para a distribuição de probabilidade do fenômeno, uma vez que o conjunto de eventos $\{r_i\}$ é indefinido) e a entropia redundante H (associada a um processo com conjunto de eventos $\{r_i\}$ finito e que podem ser acessados):

$$R = H_0 - H\,. \tag{2.20}$$

Posteriormente, essa taxa de transmissão de informação ficou conhecida simplesmente como informação I [81] e também ganhou uma nova interpretação: informação é a medida do ganho (ou perda) de certeza durante um processo no qual o nosso conhecimento acerca dos eventos acessíveis varia. No lançamento de uma moeda não viciada, as probabilidades de cada um dos eventos são $p_{N-V} = 0.5$ e $q_{N-V} = 0.5$ e a entropia calculada através da equação (2.19) com $b = 2$ é $H_0 = 1$ *bit*. Já para uma moeda desonesta em que $p_D = 0.9$ e $q_D = 0.1$, a entropia é $H \sim 0{,}47$ *bit*, portanto temos que a informação perdida no processo é $I \sim 0.53$ *bit*.

Dessa forma, a entropia de Shannon seria capaz de revelar o quão ordenado e previsível é uma configuração do sistema, esse é exatamente o ponto de intersecção conceitual entre as entropias de Shannon e Gibbs. E esse fato motivará o uso da formulação de Shannon com base $b = e$ dada por:

$$H = -\sum_{i=1}^{N} p(r_i) \ln\left[p(r_i)\right]. \tag{2.21}$$

No decorrer desse trabalho utilizaremos a equação (2.21) para estudar a distribuição de distâncias entre as sucessivas ocorrências dos verbetes em linguagem escrita.

Análises da entropia em textos

Explorando o conceito de entropia definido por Shannon, Montemurro e Zanette [82] propuseram um modelo para estudar a distribuição de palavras de acordo com sua função linguística. Ao dividir o texto em P partições definidas naturalmente em escalas distintas, como por exemplo: os capítulos de um livro, as seções, parágrafos e sentenças. Definem-se, então, N_i como o número total de palavras da i-ésima partição e $k_i(w)$ o número de ocorrências do verbete w nessa partição, temos que a fração de verbetes w na partição i ($f_i(w)$) é dada por:

$$f_i(w) = \frac{k_i(w)}{N_i}\,.$$

Podemos assim definir a distribuição de probabilidades $p_i(w)$ de ocorrência de uma dada fração $f_i(w)$ para todas a partições:

$$p_i(w) = \frac{f_i(w)}{\sum\limits_{j=1}^{P} f_j(w)} \,. \tag{2.22}$$

A entropia associada à distribuição $p_i(w)$ é dada pela equação 2.18 com $K_S = \frac{1}{\ln P}$, assim escolhido para que o valor máximo de $S_{max}(w) = 1$, e $b = e$:

$$S(w) = -\frac{1}{\ln P} \sum_{i=1}^{P} p_i(w) \ln \left[p_i(w) \right] . \tag{2.23}$$

Dois casos limites de interesse da equação (2.23) são: um verbete uniformemente distribuído entre as partições ($p_i(w_1) = 1/P$) nos fornece $S(w_1) = 1$ e um verbete presente em somente uma das partições ($p_i(w_2) = 1$ e $p_j(w_2) = 0$ para $i \neq j$), tem-se $S(w_2) = 0$.

Dessa maneira, Montemurro e Zanette argumentam que a entropia dada pela equação (2.23) é capaz de classificar verbetes, uma vez que aqueles de uso frequente (como preposições, artigos e conjunções) apresentam altos valores de entropia, enquanto as palavras relevantes do texto apresentariam valores baixos de entropia [21,82,83]. Essas conclusões são ilustradas no gráfico da Figura 2.9 em que os verbetes mais comuns possuem entropia mais próxima de 0, enquanto verbetes na região $2 \cdot 10 < k < 3 \cdot 10^2$ possuem entropias próximas ao valor unitário.

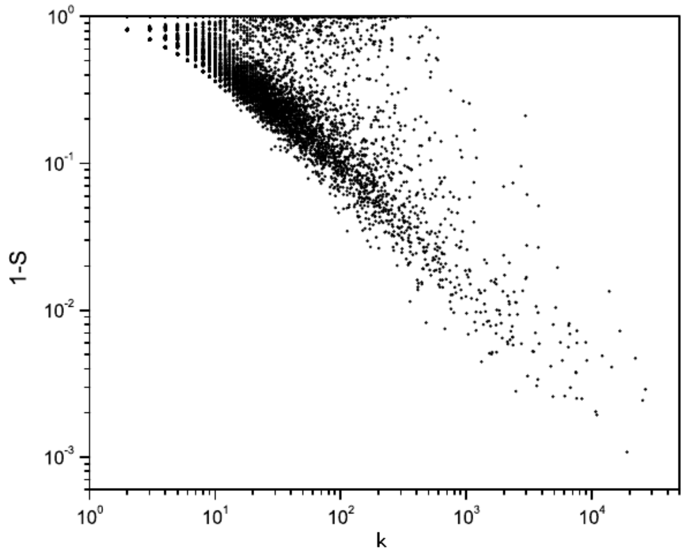


Figura 2.9: Gráfico log-linear da entropia $S(w)$ (ver equação (2.23)) e o número de ocorrências k para os verbetes de 36 obras de William Shakespeare. Figura adaptada de [82].

Abordagens frequencistas semelhantes para a distribuição de verbetes foram adotadas na extração de palavras-chave [16], quantificação da informação do conteúdo semântico língua [21], estudo da informação nas estruturas de longo alcance [83] e extração da informação mútua entre verbetes [84].

Propondo uma abordagem que leva em consideração a distribuição das distâncias entre as sucessivas ocorrências dos verbetes, Rolim [38], inspirado pelo trabalho em entropia de sistemas bosônicos unidimensionais de Mehri e Darooneh [85], definiu a entropia $H(w)$:

$$H(w) = -\sum_{\{d\}} p_w(d) \ln\left[p_w(d)\right] \tag{2.24}$$

em que $p_w(d)$ representa a probabilidade de ocorrência da distância d associada ao verbete w e $\{d\}$ é o conjunto de distâncias entre as sucessivas ocorrências do verbete. A construção do conjunto $\{d\}$ é feito de maneira análoga ao cálculo da intermitência σ visto na seção 2.2.

O comportamento típico dessa entropia com a frequência dos verbetes é ilustrado nos gráficos da Figura 2.10 para IN-20 e HU-25 (ver Apêndice A para referências). A concavidade para baixo presente na curva é comum a todos os textos no *corpora* e seu valor de máximo foi estudado por Rolim [38] que também investigou dois modelos de distribuição de distâncias para os verbetes (exponencial e geométrico), encontrando expressões para a entropia de ambas as distribuições.

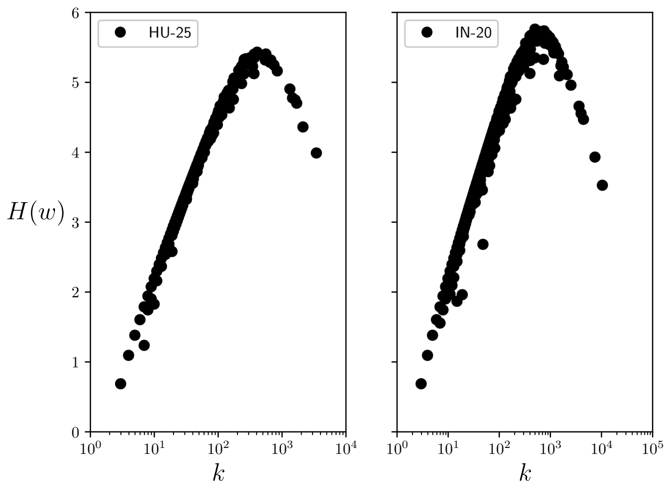


Figura 2.10: Gráficos log-linear da entropia $H(w)$ dada pela equação (2.24) para a distribuição de distâncias e o número de ocorrências k para os verbetes de IN-20 e HU-25.

Os modelos utilizados por Rolim e suas implicações serão mais detalhadamente estudados no capítulo 4, em que exploraremos também outros modelos para a distribuição de distâncias como o proposto por Altmann e colaboradores [34]. No entanto, antes de nos aprofundarmos no estudo da distribuição espacial de verbetes e sua utilização na produção de textos genéricos, estudaremos o comportamento das séries temporais de textos em linguagem natural e suas correlações. Essa ordem se faz necessária para que possamos utilizar os métodos de análise de séries temporais no estudo de propriedades de textos genéricos construídos a partir de uma distribuição de distâncias.

Séries temporais

Muitas observações da natureza consistem de registros no tempo que geralmente exibem variações nos períodos de observação [86]. Aos registros organizados em ordem temporal, dá-se o nome de série temporal. A exemplo desse tipo de observação podem ser citados o registro da temperatura de uma cidade, o volume de água de um rio ou as frequências de verbetes em textos.

Textos escritos podem ser interpretados como um sinal físico, pois podem ser decompostos em multiplos níveis distintos [26]. A Figura 3.1 ilustra uma série temporal para as frequências (STF), comprimentos (STC) e intermitência[1] (STI) das 500 primeiras posições no livro PT-13 ($T = 77.192$). Esses são exemplos típicos de séries temporais em textos em que mapeamos o parâmetro desejado (frequência ou comprimento do verbete) referente à palavra naquela dada posição em um valor. Se considerarmos as quantidades observadas como variáveis aleatórias, uma série temporal nada mais é que uma sequência de váriaveis aleatórias y_1, y_2, y_3 ... y_N, em que y_1 é a medida tomada na primeira observação, y_2 na segunda observação e assim por diante. De maneira mais geral, uma coleção de variáveis aleatórias $\{y_t\}$ indexada pelo intervalo de tempo t é chamada de processo estocástico [87].

A aplicação de modelos estocásticos para descrição de propriedades dos textos em linguagem natural já foi realizada para compreender quais mecanismos levam a lei de Zipf [52], a criação de textos aleatórios [54, 59] e o crescimento do vocabulário [22] são alguns dos exemplos que encontram paralelo com o presente trabalho. Na seção 4.2, um modelo estocástico será utilizado para geração de textos aleatórios através da distribuição de Weibull para determinar o espaçamento entre ocorrências de um mesmo verbete.

[1] A série temporal da intermitência é construída considerando que seu valor para os verbetes com frequência $k = 1$ é $\sigma_{k=1} = -1$ e $\sigma_{k=2} = 0$ para os verbetes de frequência $k = 2$. Justifica-se a escolha desses valores para que a série temporal não possua descontinuidades.

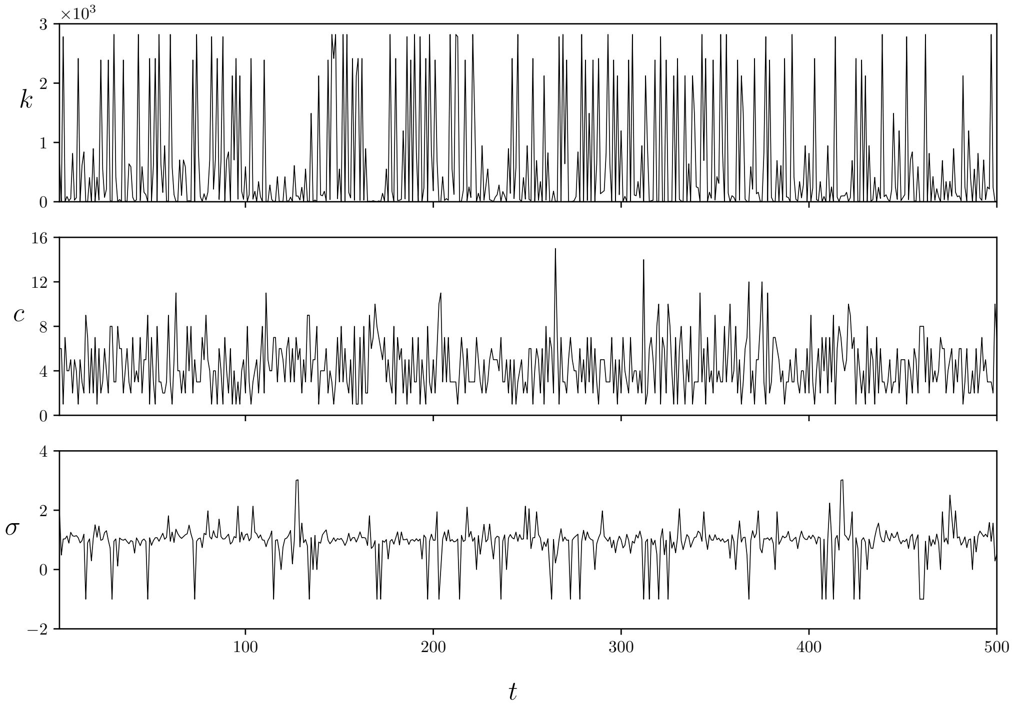


Figura 3.1: Séries temporais das frequências k, comprimentos c e intermitência σ em PT-13 para as 500 primeiras posições do texto.

Correlação em séries temporais

Uma das áreas de interesse no estudo das séries temporais é a presença de memória de longo alcance, sua existência indica qual a relevância das autocorrelações na série. Uma das medidas mais comumente utilizadas para a determinação da existência de memória de longo alcance é o expoente de Hurst $\mathcal{H}$ [26]. Ele é uma medida da tendência relativa de uma série temporal ter um retorno à média ou aglomerar em uma direção [88] e costuma ser obtido através da análise de intervalo re-escalado de Hurst (R/S Analysis).

R/S Analysis

Para determinar o valor do expoente de Hurst através de R/S Analysis, seguimos o algoritmo em 7 passos abaixo [86, 89]:

1) determinamos a média (μ) da série temporal:

$$\mu = \frac{1}{N_t} \sum_{t=1}^{N_t} y_t \tag{3.1}$$

em que N_t é o tamanho da série;

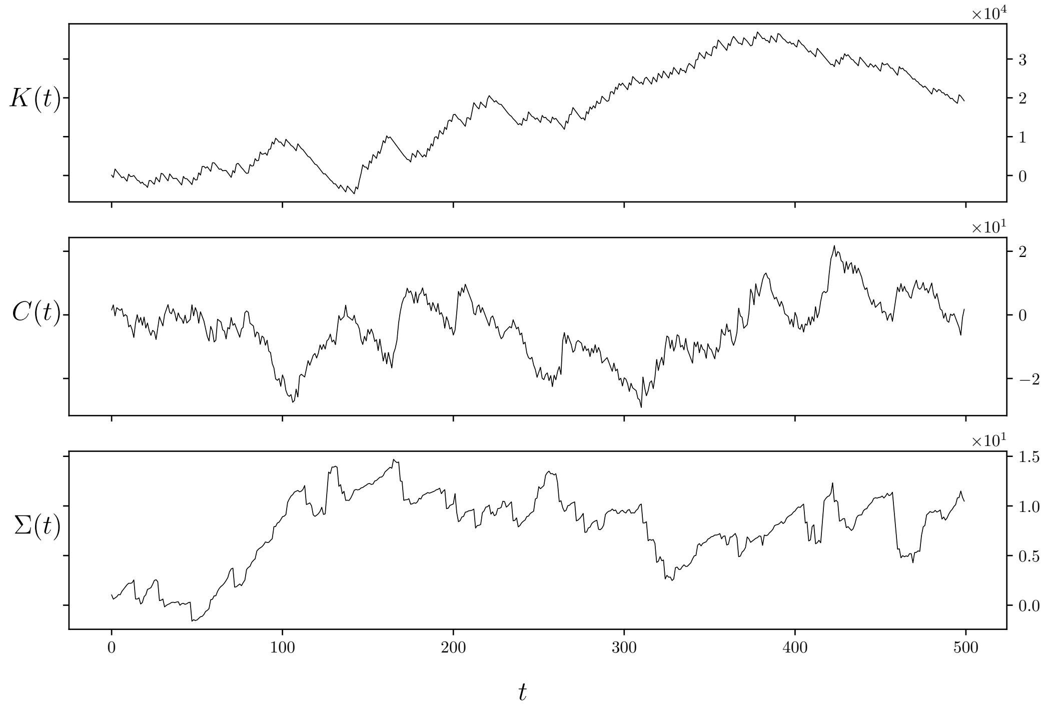


Figura 3.2: Perfis de STF $K(t)$, STC $C(t)$ e STI $\Sigma(t)$ em PT-13 para as 500 primeiras posições do texto. A escolha das 500 primeiras palavras foi feita para melhor visualização da série.

2) a partir da média, construímos o perfil $Y(t)$ da série:

$$Y(t) = \sum_{u=1}^{t} (y_u - \mu) \ . \tag{3.2}$$

O perfil nos indica a soma das diferenças entre os valores da série e a média. A Figura 3.2 ilustra o perfil das séries mostradas na Figura 3.1 para o texto PT-13;

3) dividimos o perfil da série em s subséries de tamanho τ. Para evitar a existência de subséries com tamanhos menores que τ (situação comum quando a divisão do tamanho da série por τ não é exata), uma das técnicas utilizadas é o espelhamento da série em sua última posição para criação de uma nova série $y_{fin} = \{y_1, y_2, y_3, ..., y_N, y_{N-1}, y_{N-2}, ..., y_3, y_2, y_1\}$ que será divididas em $s_{fin} = 2s$ subséries de tamanho τ [90];

4) calcula-se então o valor do intervalo $R(n)$ dado pela diferença entre o maior e menor valor do perfil da série $Y(t,n)$ dentro da região $n = \{1, 2, ..., s_{fin}\}$:

$$R(n) = \max_{\tau_n \leq t < \tau_{n+1}} Y(t,n) - \min_{\tau_n \leq t < \tau_{n+1}} Y(t,n)\,;$$

5) encontra-se o desvio padrão S_H da série acumulada:

$$S_H = \left[\frac{1}{N_t} \sum_{t=1}^{N_t} (y_t - \mu)^2 \right]^{1/2} \quad ;$$

6) a partir de $R(n)$ e S_H, determinamos a razão R/S_H para cada uma das s subséries e calculamos sua média $\left\langle \frac{R}{S_H} \right\rangle_\tau$, chamada também de intervalo re-escalado, para todas as subséries de tamanho τ;

7) ao repetir os passos 1 a 6 para distintos valores de τ, Hurst [91] observou que o intervalo re-escalado é dado por uma lei de potência:

$$\left\langle \frac{R}{S_H} \right\rangle \sim \tau^{\mathcal{H}} . \tag{3.3}$$

O gráfico típico ao fazer a R/S Analysis para valores dos comprimentos das subséries variando entre $8 \leq \tau \leq T/8$, em que T é o tamanho do texto, é mostrado na Figura 3.3 para as séries da frequência e comprimento de PT-13.

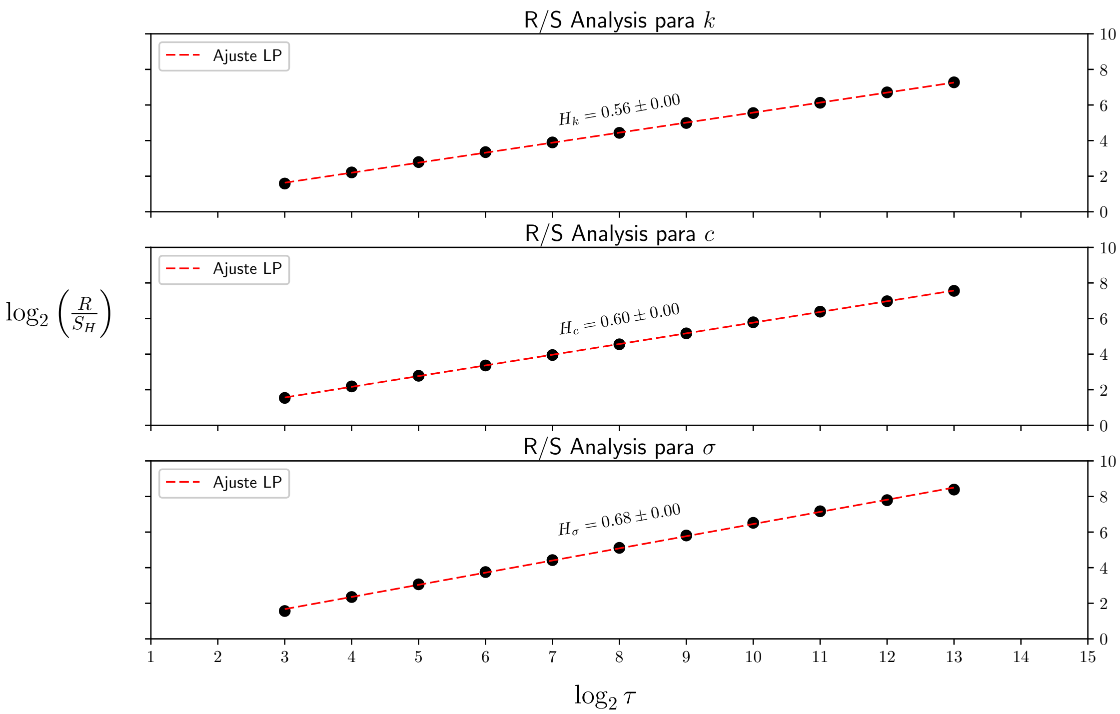


Figura 3.3: R/S Analysis para as séries temporais das frequências, comprimentos e intermitência em PT-13. O comprimento das subséries variam entre $8 \leq \tau \leq T/8$. Os expoentes da regressão para frequência k, comprimento c e intermitência σ são $\mathcal{H}_k = 0{,}56 \pm 0{,}00$, $\mathcal{H}_c = 0{,}60 \pm 0{,}00$ e $\mathcal{H}_\sigma = 0{,}68 \pm 0{,}00$, respectivamente. A linha tracejada em vermelho corresponde ao ajuste do tipo de lei de potência para os dados.

O expoente de Hurst

Os valores do expoente de Hurst se encontram entre $0 < \mathcal{H} < 1$ em que seu valor para uma série temporal pode indicar 3 cenários:

i) $0 < \mathcal{H} < 0.5$ indica anti-correlação na série, ou seja, o sinal representado possui uma tendência de longo alcance em que os valores alternam entre altos e baixos em pares adjacentes. Isso indica que os valores maiores possuem um probabilidade maior de serem seguidos por valores menores e vice-versa, criando um estrutura anti-correlacionada em longa escala;

ii) $0.5 < \mathcal{H} < 1$ sugere correlação na série. Valores nesse intervalo caracterizam sinais em que valores maiores da série costumam ser seguidos por valores maiores da série, assim como valores menores tendem a ser seguidos por valores menores, criando um estrutura correlacionada de longo alcance;

iii) $\mathcal{H} \sim 0.5$ pode indicar que a série é completamente descorrelacionada. No entanto, o expoente de Hurst é uma medida da memória de longo alcance das séries, o valor do expoente próximo a meio pode indicar que estruturas de curto alcance correlacionadas (ou anti-correlacionadas) podem existir, mas suas autocorrelações decaem rapidamente para zero [86, 88].

Utilizando a equação (3.3) e variando o comprimento do intervalo analisado $8 \leq \tau \leq T/8$, encontramos que os expoentes de Hurst, como indicado na Figura 3.3, para as séries temporais da frequência, comprimento e intermitência para PT-13 são respectivamente $\mathcal{H}_k = 0{,}56 \pm 0{,}00$, $\mathcal{H}_c = 0{,}60 \pm 0{,}00$ e $\mathcal{H}_\sigma = 0{,}68 \pm 0{,}00$, ao supor a não estacionaridade das séries. Esses valores indicam a existência de correlações de longo alcance na distribuição de frequências e comprimentos das palavras no texto.

O expoente de Hurst foi inicialmente utilizado para estudos relacionados ao armazenamento de água [91, 92], posteriormente teve seu uso ampliado em diversas áreas tão distintas quanto as análises de correlação em séries temporais financeiras [89, 93] e o estudo dos genes essenciais/não-essencias no genoma de bactérias [94].

A versão generalizada de $\mathcal{H}$ também foi utilizado por Ausloos [26] no estudo das séries temporais de textos originais e traduzidos. Mapeando as séries temporais da frequência e comprimento para três textos de maneira análoga ao que foi feito no presente trabalho, ele encontrou valores do expoente generalizado de Hurst $\mathcal{H}_G$ que indicam a presença de correlação nas séries.

É importante uma distinção entre o expoente de Hurst e sua versão generalizada: o expoente de Hurst é definido para séries temporais estacionárias gerada por um processo estocástico auto-similar [95]. Essas séries seriam, portanto, originadas através de um processo estocástico em que a distribuição de probabilidade conjunta das variáveis aleatórias envolvidas não mudam quando sofrem uma translação temporal ou mudança de escala. Isso implica que essas séries possuem média e variância independentes do tempo (as s subséries geradas para um valor de τ) e da escala adotada (como o tamanho τ escolhido para o intervalo) [87, 95].

A versão generalizada do expoente pode ser utilizada para séries não-estacionárias, como as encontradas em textos em linguagem natural e traduçoes em linguagens artificiais [26, 68, 96]. Porém, ao não considerar as tendências locais da série que podem apresentar características não-estacionárias e não conseguir distinguí-las das correlações de longo alcance, o método pode nos fornecer valores espúrios para o expoente [95].

Para contonar as limitações do expoente de Hurst (e de sua forma generalizada), a análise de flutuações sem tendência - *Detrended Fluctuation Analysis* em inglês - ou simplesmente DFA, tem sido adotada como técnica padrão para o estudo de séries temporais não-estacionárias que podem possuir tendências locais e correlações de longo alcance [97].

Detrended Fluctuation Analysis (DFA)

Em 1994, Peng e colaboradores [98] propuseram o uso de uma técnica para a análise de séries temporais não-estacionárias capaz de distinguir entre tendências locais e correlações de longo alcance. Essa técnica, chamada de *Detrended Fluctuation Analysis* (DFA), tem sido, desde então, utilizada em séries temporais de origens distintas: em nucleotídeos [98], séries de chuvas e vazões de rios [99] e preço do petróleo [100].

As propostas de aplicação do DFA existentes na literatura para textos em linguagem natural são voltadas ao estudo das séries temporais da frequência e comprimento de verbetes [27, 41, 101]. Motivados por esses trabalhos, propomos o uso do DFA para as séries temporais da frequência, comprimento e intermitência σ (discutido na subseção 2.2.3), assim como o estudo das suas correlações de curto e longo alcance em textos escritos.

Descrição do DFA

O DFA possui 7 passos em seu algoritmo [90,98], sendo os 3 primeiros idênticos aos da R/S Analysis (ver subseção 3.1.1) em que determinamos a média da série, seu perfil e a dividimos em s subséries de tamanho τ, a partir de então fazemos:

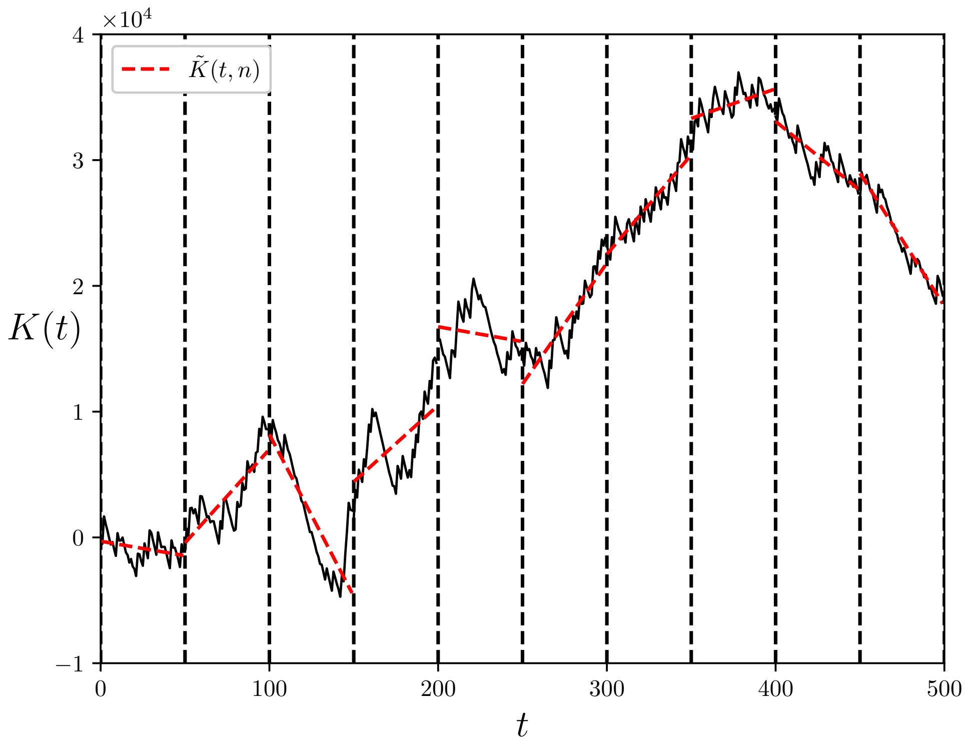


Figura 3.4: Tendências locais de STF em PT-13 para uma regressão polinomial de primeira ordem. O tamanho das subséries foi escolhido como $\tau = 50$ e foram consideradas as 500 primeiras posições.

1) com o perfil da série determinado, tomamos sua regressão polinomial de ordem l em sua subsérie n (para essa ordem, o método é chamado de DFA-l) para encontrar sua tendência local $\tilde{Y}(t,n)$ dada pela curva de ajuste polinomial para a posição t na subsérie n. A Figura 3.4 ilustra a tendência local para o perfil de STF de PT-13 para uma regressão de primeira ordem. A partir da tendência local, construímos a sequência

residual:

$$\epsilon_n(t) = Y(t,n) - \tilde{Y}(t,n)\,; \tag{3.4}$$

2) da sequência residual podemos escrever a função de flutuação $F(n,\tau)$ que mede a flutuação acumulada em torno da regressão polinomial para a subsérie n de tamanho τ e é dada por:

$$[F(n,\tau)]^2 = \frac{1}{\tau}\sum_{u=1}^{\tau}[\epsilon_n(u)]^2\ ; \tag{3.5}$$

3) repete-se os passos de 1) a 5) do algoritmo acima para as s subséries e calcula-se o valor total da função de flutuação, dado por:

$$[F(\tau)]^2 = \frac{1}{s}\sum_{n=1}^{s}[F(n,\tau)]^2\ ; \tag{3.6}$$

4) ao variar τ, determinamos a relação do tipo lei de potência entre as funções de flutuação $F(\tau)$ e o comprimento da subsérie τ:

$$F(\tau) \sim \tau^D \tag{3.7}$$

em que D é o expoente de escala para o DFA.

Após executado o algoritmo descrito acima, temos os valores das flutuações $F(\tau)$ e comprimentos τ como ilustrado na Figura 3.5 para o DFA-1 das séries temporais das frequências e comprimentos em PT-13. No gráfico, vemos a existência de duas regiões com diferentes expoentes, esse comportamento das flutuações já foi observado para as séries temporais de frequência e comprimento presentes em linguagem natural [27, 41].

Resultados DFA

A partir da série podemos executar o algoritmo do DFA e obtemos as flutuações como mostra a Figura 3.6 novamente temos a separação em duas regiões[2] com expoentes distintos. Os valores para o coeficiente D possuem a mesma interpretação dada para o expoente $\mathcal{H}$ discutida na subseção 3.1.2. Portanto, os resultados para as séries temporais mostradas na Figura 3.5 indicam a existência de dois tipos de regimes:

[2] O comprimento τ de separação entre os dois regimes foi obtido ao minimizar a soma do erro das regressões nas duas regiões quando varia-se τ. Essa metodologia foi proposta e adequadamente fundamentada em [27].

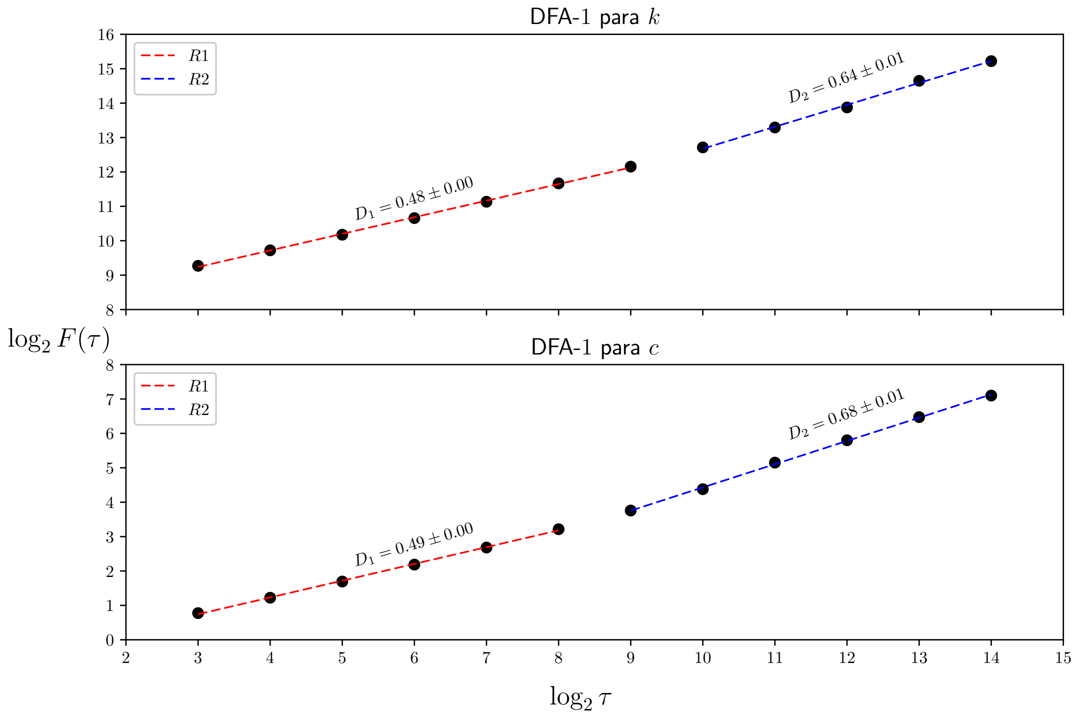


Figura 3.5: Flutuações em função do comprimento τ para as séries das frequências e comprimentos utilizando o DFA-1 no texto PT-13. Observa-se a presença de dois regimes ($R1$ e $R2$) com expoentes distintos (D_1 e D_2).

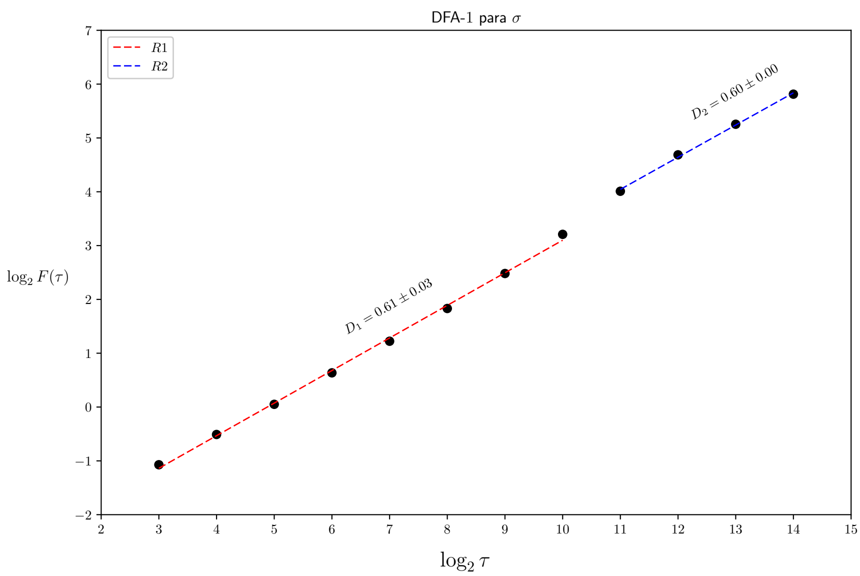


Figura 3.6: Flutuações em função do comprimento τ para STI utilizando o DFA-1 no texto PT-13.

i) regime $R1$ aproximadamente descorrelacionado quando analisamos regiões relativamente pequenas $2^3 \leq \tau \leq 2^9$ comparadas ao tamanho do texto ($T = 77.192$). Isso pode indicar que em curtas escalas, o uso de verbetes independe de sua frequência e tamanho;

ii) regime $R2$ correlacionado presente em regiões grandes do texto $2^9 \leq \tau \leq 2^{14}$, indicando a presença de correlação de longo alcance na escolha de verbetes que possuam certos tamanhos e frequências.

No entanto, a Figura 3.6 indica que apesar da existência de dois regimes correlacionados da intermitência para o texto em análise, seus valores são bastante próximos. Esse resultado sugere que os valores da intermitência possuem correlação de curto e longo alcance e que, ao menos para PT-13, a existência de um único regime seria suficiente para justificar as flutuações observadas. É importante observar também que esse resultado pode ser afetado pela ordem do polinômio l adotada ao executar o algoritmo.

A escolha da ordem do polinômio afeta os valores obtidos para a sequência residual, modificando os valores das flutuações. Portanto, costuma-se analisar qual a menor ordem de l que possui o melhor ajuste da lei de potência. A Figura 3.7 ilustra o valor do expoente médio $\tilde{D}$ nos dois regimes $R1$ e $R2$ para todos os textos literários em português das séries temporais da frequência, comprimento e intermitência. Observa-se que o valor de $l_{min} = 4$ indica o menor expoente em que se obtém o melhor ajuste para o português, resultados semelhantes são encontrados para as demais línguas em nosso *corpora*. Uma vez conhecido o valor l_{min}, visando diminuir o tempo computacional despendido para executar o algoritmo, faremos o uso dos polinômios até essa ordem.

Uma característica importante presente nos gráficos da Figura 3.7 é a distinção entre o comportamento dos valores de $\tilde{D}$ para as ordens do polinomîio maiores que l_{min} em relação às médias $\left\langle \tilde{D}_1 \right\rangle$ e $\left\langle \tilde{D}_2 \right\rangle$. Esse comportamento, típico para os textos em todos os idiomas do nosso *corpora*, sugere no regime $R2$ que a ordem l necessária para encontrar o melhor ajuste seja maior que l_{min}. Portanto para justificar a escolha de $l_{min} = 4$ na segunda região, argumentamos que:

i) a escolha de ordens superiores de regressão pode enviesar o valor $\tilde{Y}(t, n)$, uma vez que há o aumento do número de parâmetros utilizados na regressão, aumentando a possibilidade de sobre-ajuste para valores pequenos de τ [102];

ii) o tempo computacional necessário para executar o algoritmo escala rapidamente com a escolha da ordem do polinômio de regressão, tornando inviável escolhas de polinômios de ordens superiores a $l = 10$.

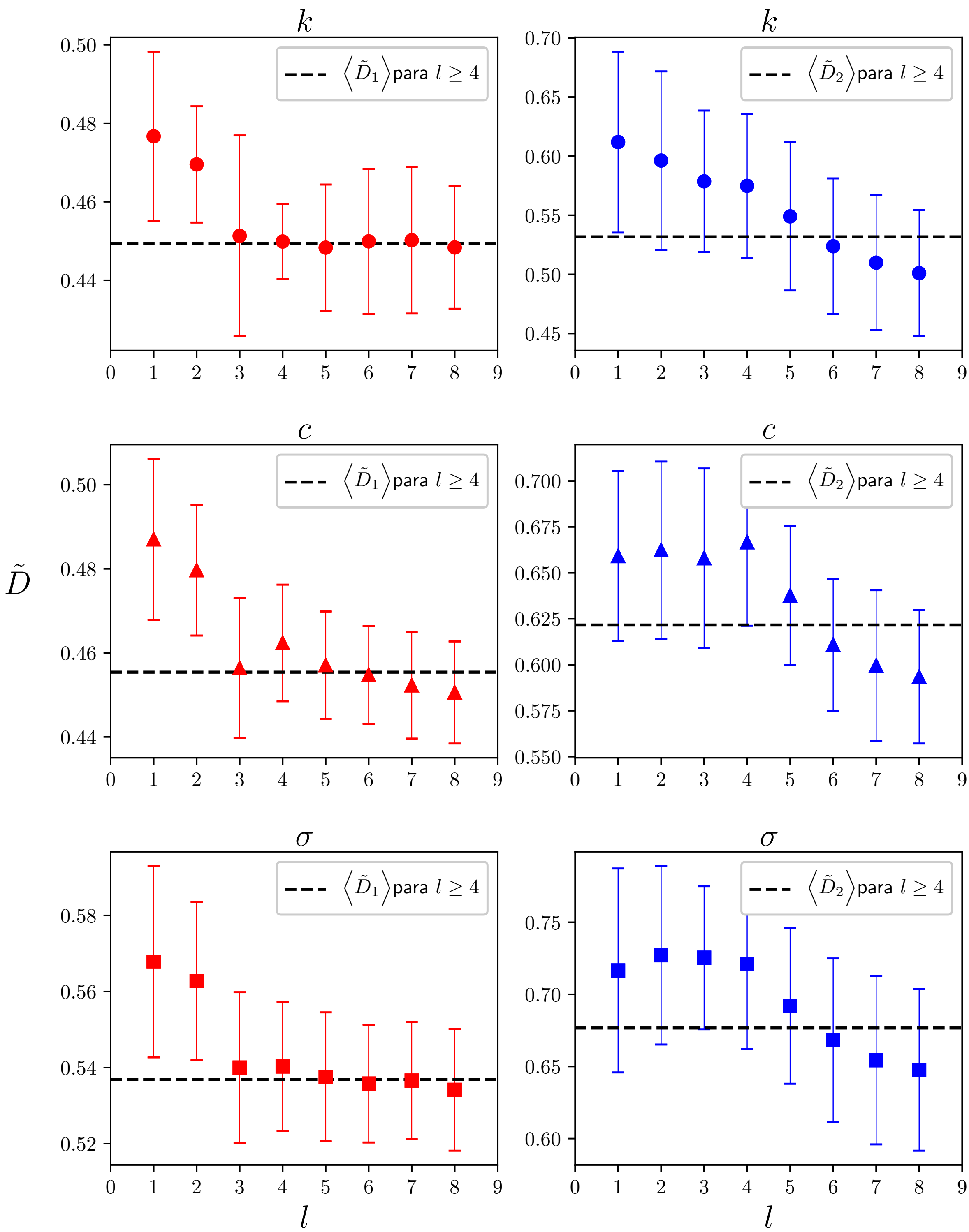


Figura 3.7: Expoentes médios de todos os textos em português em ambos os regimes $R1$ e $R2$ para 8 ordens da regressão polinomial do DFA. Foi calculado o valor médio de $\tilde{D}_1$ e $\tilde{D}_2$ entre os valores $4 \leq l \leq 8$ para melhor visualização do menor polinômio de melhor ajuste l_{min}.

Ao aplicar os DFA-1 e DFA-4 para as séries temporais da frequência, comprimento e σ para todos os textos literários no *corpora*, encontramos seus expoentes médios por regimes e língua como mostrado na Figura 3.8.

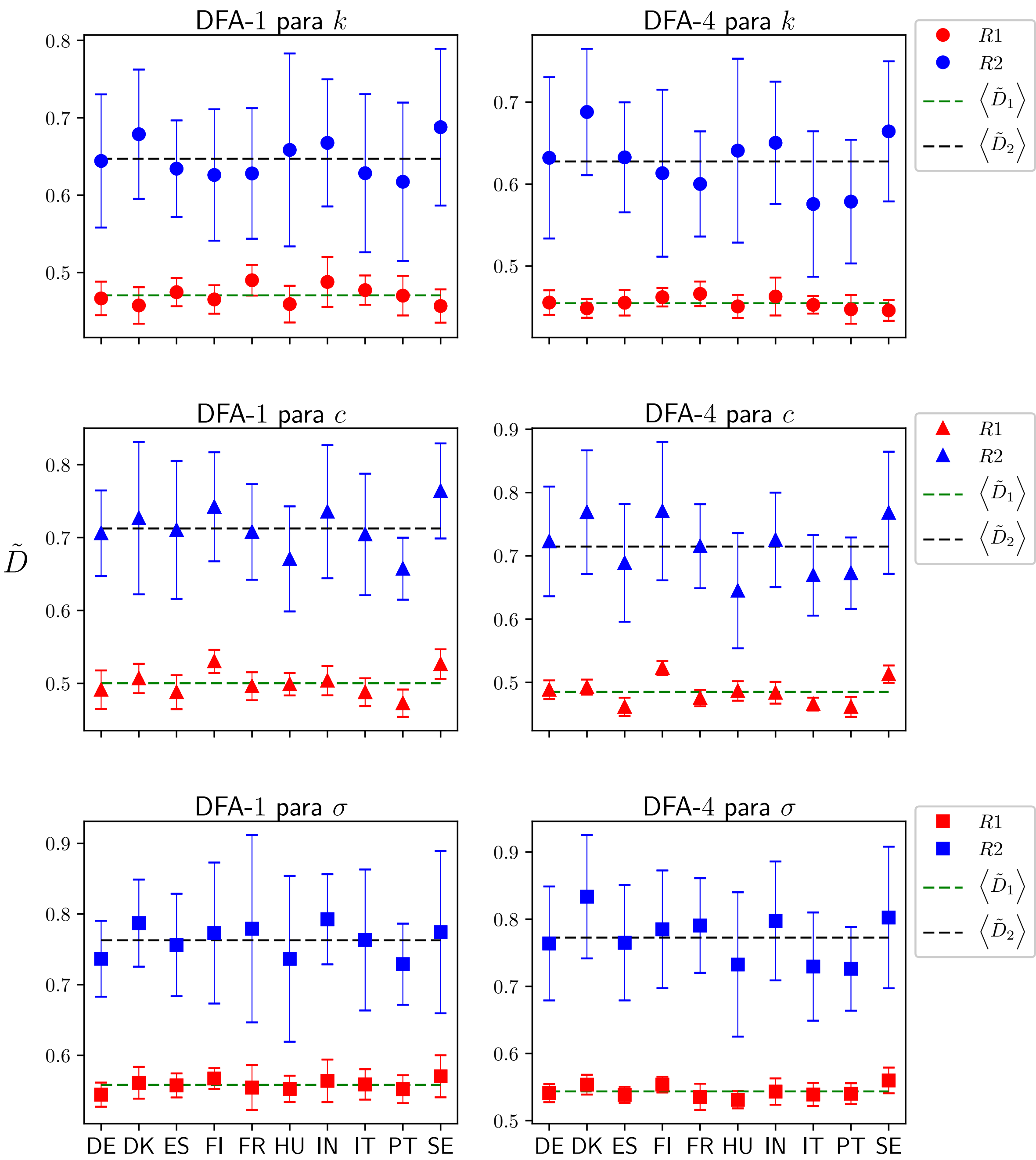


Figura 3.8: Expoentes médios dos dois regimes $R1$ e $R2$ de todos os textos de cada língua para DFA-1 e DFA-4. As linhas tracejadas em verde e preto representam respectivamente o valor médio do expoente das regiões $R1$ e $R2$ para todas as línguas. As línguas foram representadas por suas siglas de referência (para todos os textos utilizados e suas referências, ver o Apêndice A).

Os resultados do DFA para STF e STC sugerem as seguintes interpretações:

1) no regime $R1$ em que os comprimentos das subséries variam de uma frase até seções e capítulos, as séries temporais da frequência e comprimento são respectivamente anticorrelacionadas e descorrelacionadas, resultados semelhantes foram encontrados na literatura [27, 41]. Tais resultados implicam que a distribuição de verbetes em curtas

escalas por suas frequências segue um padrão de verbetes de menor frequência seguidos de verbetes de frequências mais altas, enquanto que a distribuição de verbetes por comprimento se assemelha a um processo browniano;

2) o regime $R1$ para as séries temporais da intermitência também indicam a presença de correlações até mesmo em curtas escalas. Indicando que as palavras relevantes e irrelevantes, em geral, se encontram próximas entre si;

3) no regime $R2$ em que temos escalas muito maiores que seções e capítulos e lidam com estruturas que se aproximam do tamanho do texto, as séries temporais estudadas são correlacionadas. Esse resultado indica que existem estruturas de longo alcance fortemente correlacionadas nessas séries;

4) a presença de dois regimes distintos para as séries temporais da intermitência em todos os textos e línguas na Figura 3.8 sugere que o resultado ilustrado na Figura 3.6 é atípico para esse tipo de série.

Os textos literários utilizados possuem tamanhos que variam entre $10^4 < T < 3 \cdot 10^5$, variando assim o tamanho máximo das subséries $\tau_{max} = T/8$ para cada texto. A variação em τ_{max} é um dos fatores responsáveis pela mudança de D_2 nos textos e seus efeitos podem ser notados na barra de erro para $\tilde{D}$ em $R2$ tanto para o DFA-1 quanto para o DFA-4.

Os métodos R/S Analysis e DFA apresentados até o presente momento lidam com as autocorrelações presentes em séries temporais estacionárias e não-estacionárias, respectivamente. Para medir a influência entre as interações das propriedades dos verbetes é necessária uma técnica capaz de analisar as correlações entre séries, ou correlação cruzada. A análise de correlação cruzada sem tendências, proposta em 2008 [103], visa exatamente o estudo desse tipo de correlação para séries temporais não-estacionárias.

Detrended Cross-Correlation Analysis (DCCA)

Podobnik e Stanley [103] propuseram o uso de uma técnica para estudar correlações cruzadas em séries temporais não-estacionárias chamada por eles de *Detrended Cross-Correlation Analysis* (DCCA). Segundo os autores, ela é capaz de encontrar o nível de correlação entre séries distintas ao utilizar as flutuações da covariância entre as séries.

Por ser uma técnica capaz de lidar com séries não-estacionárias, ela já foi aplicada no estudo de séries temporais financeiras [104, 105]. Até o presente trabalho, não foram encontrados resultados na literatura da aplicação do DCCA para análise de textos em linguagem natural e o estudo de suas propriedades. Propomos, assim, o uso da técnica para o estudo da correlação cruzada entre as séries STF-STC ($k - c$), STF-STI ($k - \sigma$) e STC-STI ($c - \sigma$).

Descrição do DCCA

O algoritmo do DCCA possui 7 passos [103], sendo os 3 primeiros idênticos aos da R/S Analysis (ver subseção 3.1.1), ou seja, calculamos a média de cada série, construímos seus perfis e as dividimos em s subséries de tamanho τ. Logo após executamos o passo 1) do DFA (ver subseção 3.2.1) em que construímos as sequências residuais a partir dos valores da tendência local. Uma vez que temos as sequências residuais $\epsilon_n^1(t)$ e $\epsilon_n^2(t)$ para as séries 1 e 2, executamos os seguintes passos:

1) determinamos a função das flutuações entre as séries $F_{CC}(n, \tau)$ que mede a flutuação entre as séries acumuladas em torno da regressão polinomial para a subsérie n de tamanho τ e é dada por:

$$[F_{CC}(n,\tau)]^2 = \frac{1}{\tau}\sum_{u=1}^{\tau}\left[|\epsilon_n^1(u)|\right]\left[|\epsilon_n^2(u)|\right] \; ; \tag{3.8}$$

2) repetem-se os passos anteriores para todas as subséries e calcula-se o valor total da função de flutuação $F_{CC}(\tau)$ dada por:

$$[F_{CC}(\tau)]^2 = \frac{1}{s}\sum_{n=1}^{s}[F_{CC}(n,\tau)]^2 \; ; \tag{3.9}$$

3) variando τ, determinamos o tipo de lei de potência entre as funções de flutuação $F_{CC}(\tau)$ e o comprimento da subsérie τ:

$$F_{CC}(\tau) \sim \tau^{D_{CC}} \tag{3.10}$$

em que D_{CC} é o expoente de escala para o DCCA.

Ao executar o algoritmo descrito acima para as séries $k - c$, $k - \sigma$ e $c - \sigma$ para o DCCA-1 em PT-13, obtemos os valores das flutuações $F_{CC}(\tau)$ e comprimentos τ como ilustrado na Figura 3.9. Assim como no caso do DFA, há dois regimes $R1$ e $R2$ com expoentes distintos D^1_{CC} e D^2_{CC}.

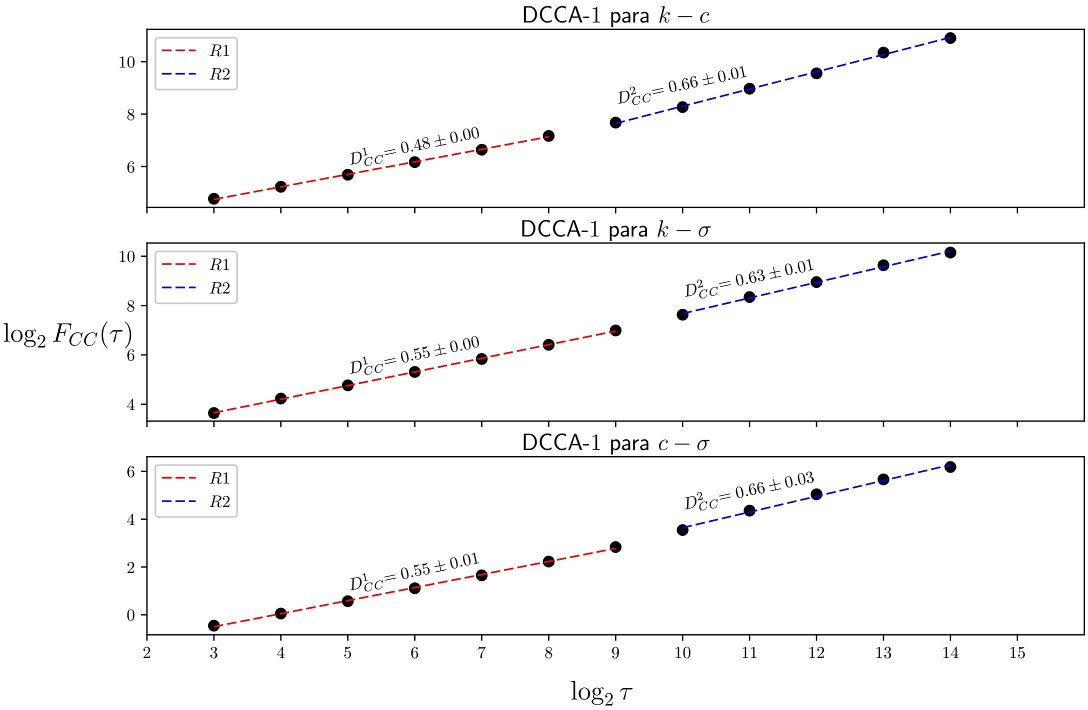


Figura 3.9: Flutuações das séries cruzadas em função do comprimento τ para $k-c$, $k-\sigma$ e $c-\sigma$ utilizando o DCCA-1 no texto PT-13. Novamente observa-se a existência de dois regimes $R1$ e $R2$ de correlação e seus respectivos expoentes D^1_{CC} e D^2_{CC}.

Resultados DCCA

Os valores dos expoentes D_{CC} obtidos através do DCCA possuem a mesma interpretação dada para $\mathcal{H}$ (discutida na subseção 3.1.2) e D no DFA. Isso significa que os gráficos na Figura 3.9 apontam a presença de um regime aproximadamente descorrelacionado entre $k-c$ enquanto os demais regimes para todas as análises cruzadas são correlacionados.

Precisamos novamente determinar qual a menor ordem l_{min} do polinômio de melhor ajuste para o DCCA em nosso *corpora*, para isso repetimos a estratégia adotada na subseção 3.2.2 em que calculamos o expoente médio $\tilde{D}_{CC}$ em $R1$ e $R2$ para todos os textos literários em português para as séries $k-c$, $k-\sigma$ e $c-\sigma$.

Os resultados são mostrados na Figura 3.10 indicando que podemos considerar $l_{min} = 4$ nos dois regimes (ver a discussão sobre a escolha de l_{min} na subseção 3.2.2). Uma vez definido l_{min}, podemos então aplicar o DCCA-1 e DCCA-4 para as séries $k-c$, $k-\sigma$ e $c-\sigma$ em todo o *corpora* e obtemos os resultados ilustrados na Figura 3.11.

A Figura 3.11 mostra que em $R1$ ambas as ordens do DCCA apontam que

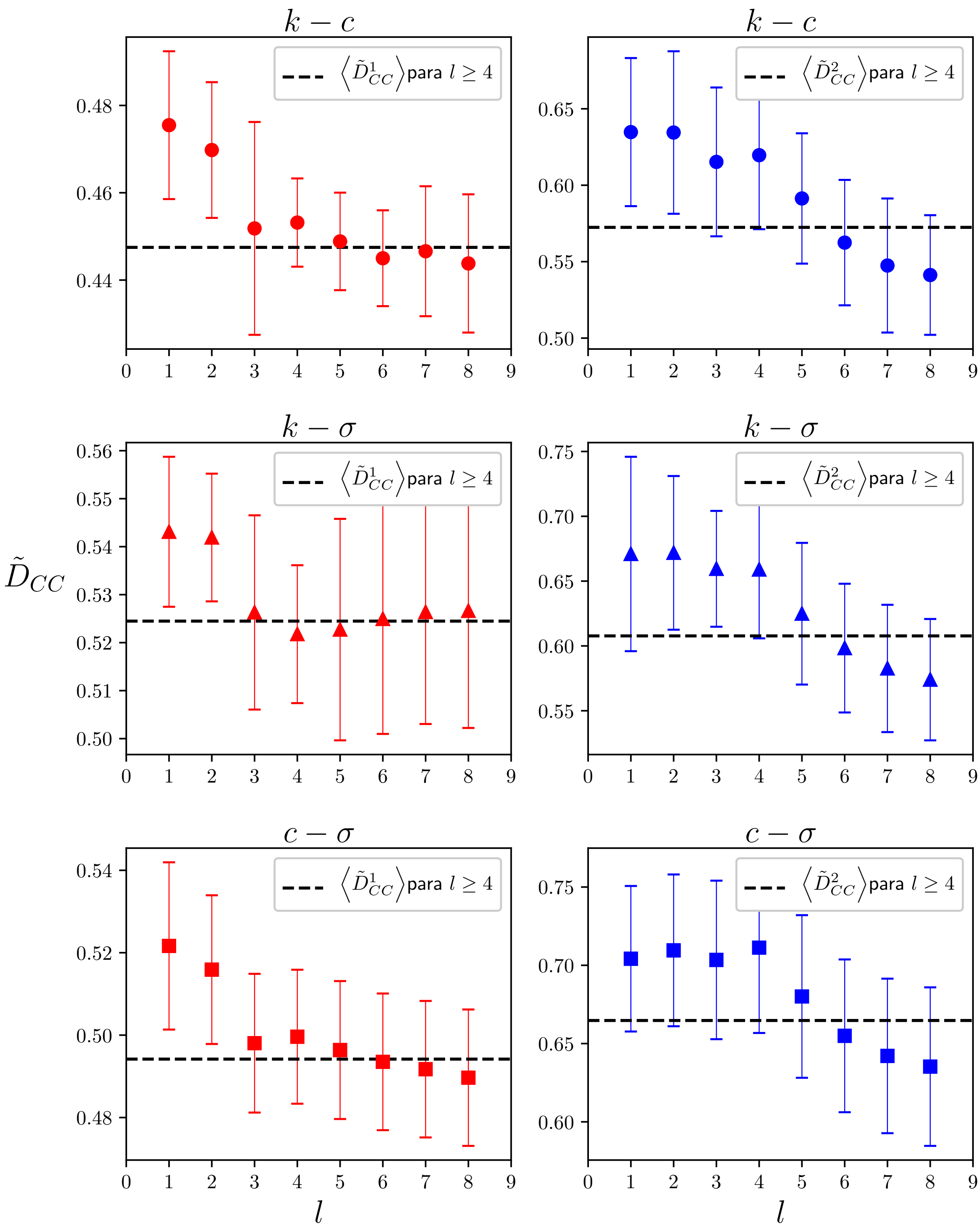


Figura 3.10: Expoentes médios de todos os textos em português em ambos os regimes $R1$ e $R2$ para 8 ordens da regressão polinomial do DCCA. Foi calculado o valor médio de $\tilde{D}^1_{CC}$ e $\tilde{D}^2_{CC}$ entre os valores $4 \leq l \leq 8$ para melhor visualização do menor polinômio de melhor ajuste l_{min}.

$k-c$ e $c-\sigma$ são séries sem correlações cruzadas, significando uma independência na ocorrência dessas propriedades dos verbetes em escalas de comprimento que variam entre frases até capítulos. Para as séries $k-\sigma$, temos que ambas as ordens do DCCA apontam a

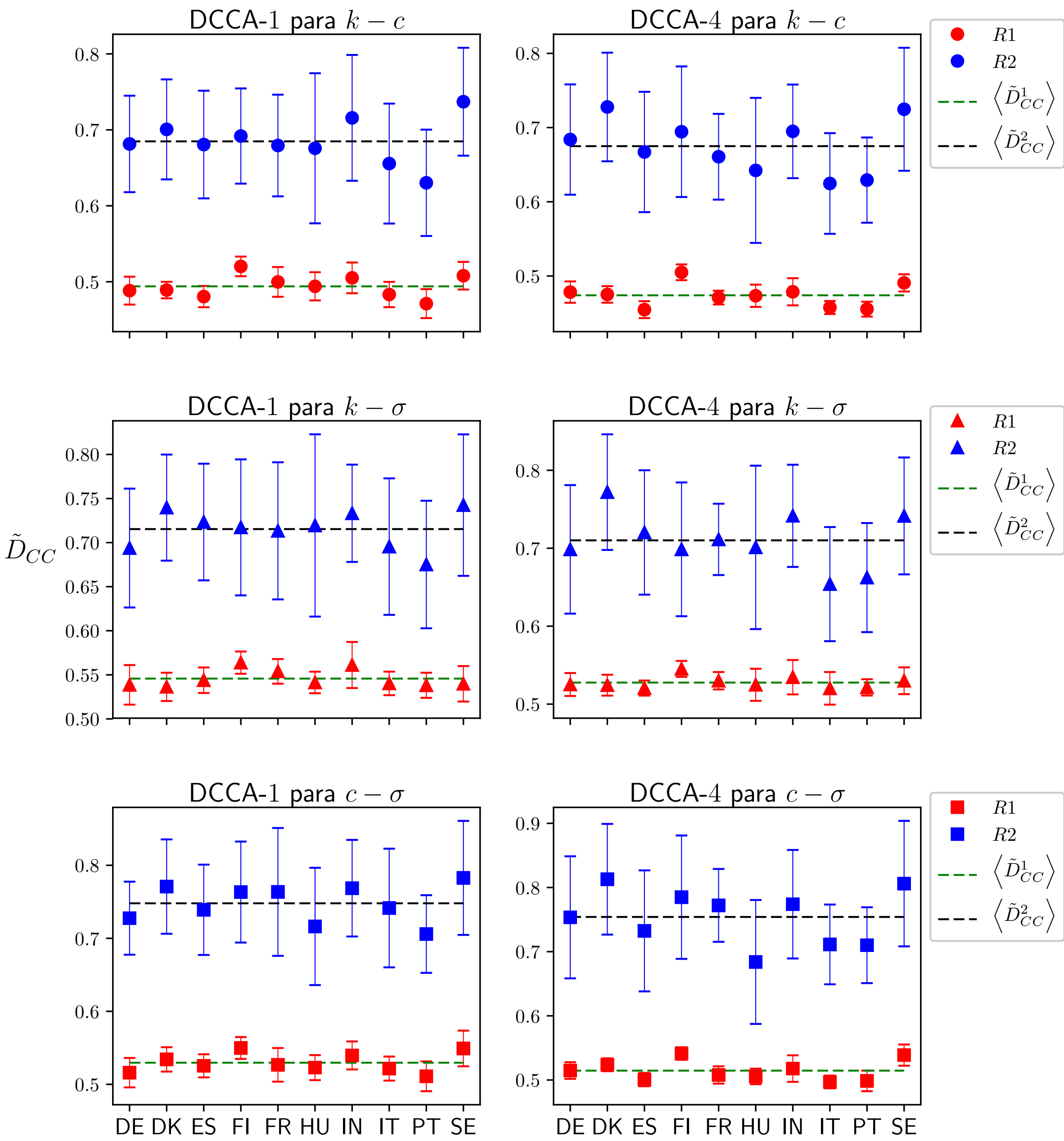


Figura 3.11: Expoentes médios dos dois regimes $R1$ e $R2$ de todos os textos de cada língua para DCCA-1 e DCCA-4. As médias em cada regime para os expoentes são representadas pelas retas tracejadas. As línguas foram representadas por suas siglas de referência (para todos os textos utilizados e suas referências, ver o Apêndice A).

existência de uma pequena correlação em curtas escalas. Temos também que em $R2$ todas as séries possuem correlações cruzadas independente da ordem do DCCA mostrada. Esse resultado revela a existência de correlações de longo alcance entre as escolhas de verbetes por suas frequências, comprimentos e relevância nos textos.

Apesar do DCCA ser capaz de detectar correlações existentes entre as séries temporais, ele não é capaz de revelar a estrutura de correlação entre as propriedades. No propósito de compreender os mecanismos responsáveis pelas correlações faz-se necessária a busca por um modelo da distribuição espacial de verbetes que possua as características reveladas pelo DFA e DCCA.

Distribuição espacial

Todas as propriedades estudadas até esse ponto do trabalho são resultados de processos estocásticos como indicado pela literatura [22, 46, 52]. Esses processos seriam responsáveis pela distribuição de frequências e comprimentos dos verbetes, assim como as suas distribuições espaciais. Rolim [38] propôs o uso da intermitência como ferramenta para estudar as distribuições espaciais dos textos em linguagem natural.

Em seu trabalho, o autor investiga as regiões limitantes das curvas típicas da intermitência σ como função da frequência k, essa curva típica foi apresentada na Figura 2.8 e as características gerais foram discutidas na subseção 2.2.3. Seus resultados serão discutidos e reavaliados ao longo desse capítulo. E, por fim, apresentaremos um modelo estocástico para criação de textos genéricos que possui propriedades semelhantes às observadas em textos reais.

Distribuições espaciais limitantes

A Figura 4.1 ilustra a distribuição da intermitência com a frequência para os textos ESW-15 (T = 2.510) e ES-25 (T = 380.331, sendo esse o maior texto em todo o *corpora*), as referências se encontram respectivamente nos Apêndices B e A. Como discutido na subseção 2.2.3, essas curvas são típicas para os textos do *corpora* e se tornam mais bem definidas quanto maior for o texto.

Para o estudo das regiões limitantes da distribuição, Rolim [38] propôs o uso de três processos distintos:

1) considerando que as palavras estão geometricamente distribuídas ao longo do texto, ou seja, as posições são determinadas por um fator multiplicativo q em relação a sua ocorrência anterior;

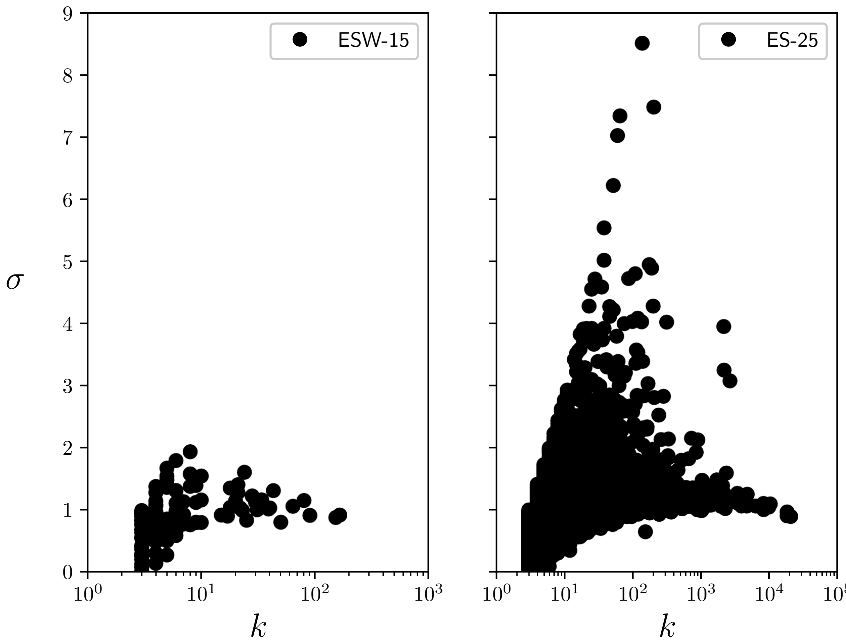


Figura 4.1: Gráfico log-linear da intermitência σ e o número de ocorrências k para ESW-15 e ES-25.

2) construindo uma sequência de números primos consecutivos e estudando a distribuição de suas distâncias;

3) adotando um modelo hamiltoniano que descreve as posições das palavras e suas correlações de longo alcance.

Modelo geométrico

Para se obter a região limitante superior, Rolim sugeriu uma distribuição em que as posições dos verbetes são determinadas por um fator multiplicativo q em relação a sua ocorrência anterior, seguindo uma sequência geométrica. Constrói-se então o conjunto $\{x\}$ das posições para um verbete com k ocorrências:

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

Supondo que a primeira ocorrência é na posição inicial do texto $x_1 = 1$, a posição x_i será dada por:

$$x_i = q^{i-1} .$$

Escrevemos assim as distâncias entre as sucessivas ocorrências s_i:

$$s_i = x_{i+1} - x_i = q^i \left(1 - q^i\right)$$

Uma vez que conhecemos as distâncias s_i e utilizando o resultado da soma dos n primeiros termos de uma série geométrica, podemos calcular a distância média:

$$\langle s \rangle = \frac{1}{k-1} \sum_{i=1}^{k-1} s_i = \frac{q^{k-1} - 1}{k-1} \tag{4.1}$$

e o segundo momento da distribuição:

$$\langle s^2 \rangle = \frac{1}{k-1} \sum_{i=1}^{k-1} s_i^2 = \frac{(q-1)}{(q+1)} \frac{q^{2(k-1)} - 1}{k-1} \, . \tag{4.2}$$

A partir das equações (4.1) e (4.2), determina-se a intermitência σ:

$$\sigma \equiv \sqrt{\frac{\langle s^2 \rangle}{\langle s \rangle^2} - 1} = \sqrt{(k-1) \frac{(q-1)}{(q+1)} \frac{q^{k-1} + 1}{q^{k-1} - 1} - 1} \, . \tag{4.3}$$

No limite em que $q \gg 1$, a equação (4.3) se torna:

$$\sigma = \sqrt{k-2} \tag{4.4}$$

Ao pensar na distribuição de distâncias que emerge desse modelo, nota-se que uma das características deste é que todas as distâncias s_i possuem a mesma probabilidade $p_g(s_i)$ dada por:

$$p_g(s_i) = \frac{1}{k-1} \tag{4.5}$$

uma vez que todas as distâncias s_i são distintas. Como conhecemos a distribuição de distâncias, podemos utilizar a definição de entropia dada pela equação (2.24) e calcular $H_g(w)$:

$$H_g(w) = -\sum_{i=1}^{k-1} \frac{1}{k-1} \ln \left(\frac{1}{k-1} \right) = \ln(k-1) \, . \tag{4.6}$$

Modelo dos primos consecutivos

Na literatura, encontra-se sistemas físicos descritos por formulações de números primos em áreas diversas, variando dos estudos de correlação em órbitas periódicas [106] a modelos de predador-presa [107]. Em especial, a sequência de números primos consecutivos segue uma distribuição de Poisson, resultado evidenciado por Wolf [70], tornando-se, dessa maneira, uma possível escolha à distribuição descorrelacionada que descreva a região limitante inferior.

Rolim estudou as propriedades da distribuição espacial dos primos consecutivos. Para tal, ele considerou um texto como sendo uma reta e as posições dos verbetes como números inteiros. Isto posto, para um verbete com frequência k, a primeira ocorrência é posta na posição 2 e as $k-1$ ocorrências nas demais nas posições dos números primos.

Computa-se, então, as distâncias entre as sucessivas ocorrências do verbete e determina-se a intermitência σ da distribuição. Ao fazer esse processo para verbetes com frequências entre $3 \leq k \leq k_{max}$, pode-se gerar um gráfico como na Figura 4.2 que apresenta os resultados para os dois modelos propostos até aqui.

Percebe-se que para frequências $k \geq 10$, o modelo dos primos consecutivos nos dá uma aproximação razoável para a relação intermitência com frequência, enquanto o modelo de distâncias geometricamente separadas nos informa o valor limitante superior.

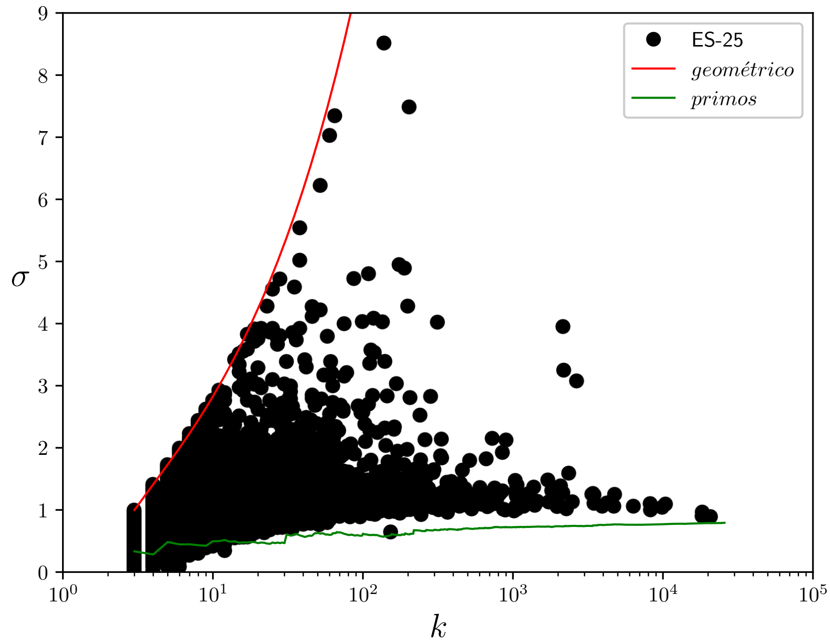


Figura 4.2: Gráfico log-linear entre σ e k para ES-25 e as curvas limitantes inferior e superior. A curva em vermelho obtida através da equação (4.4) nos dá o comportamento limitante superior, enquanto a curva em verde ilustra a intermitência de números primos consecutivos, mostrando-se uma aproximação adequada da curva limitante inferior para $k \geq 10$.

A distribuição de distâncias entre números primos consecutivos foi estudada também por Wolf, em seu trabalho ele propôs que o comportamento assintótico da distribuição de probabilidade dos espaçamentos pode ser escrita como:

$$p_p(d) = \frac{1}{\bar{d}(T)} e^{-d/\bar{d}(T)} \tag{4.7}$$

em que $\bar{d} = T/(k-1)$ representa a distância média entre ocorrências quando consideramos um texto de tamanho T e um verbete de frequência k. Assim, a entropia como dada pela equação (2.24) será:

$$H_p(w) = 1 + \ln\left(\frac{T}{k-1}\right) . \tag{4.8}$$

Para obtermos a equação (4.8) acima, assumimos a versão contínua da equação (2.24) e a integramos nos limites $0 \leq d < \infty$.

Podemos, portanto, representar a dependência de $H_p(w)$ com k no gráfico da Figura 4.3 na qual observamos que o comportamento desse modelo nos dá o limite superior da curva para altas frequências. Assim como esses resultados sugerem que o processo responsável por gerar as distribuições espaciais em um texto deve ter uma distribuição geométrica como geradora da região limitante superior. Um olhar mais atentos indica que os regimes são bem separados na região próxima da frequência em que a entropia é máxima k_0.

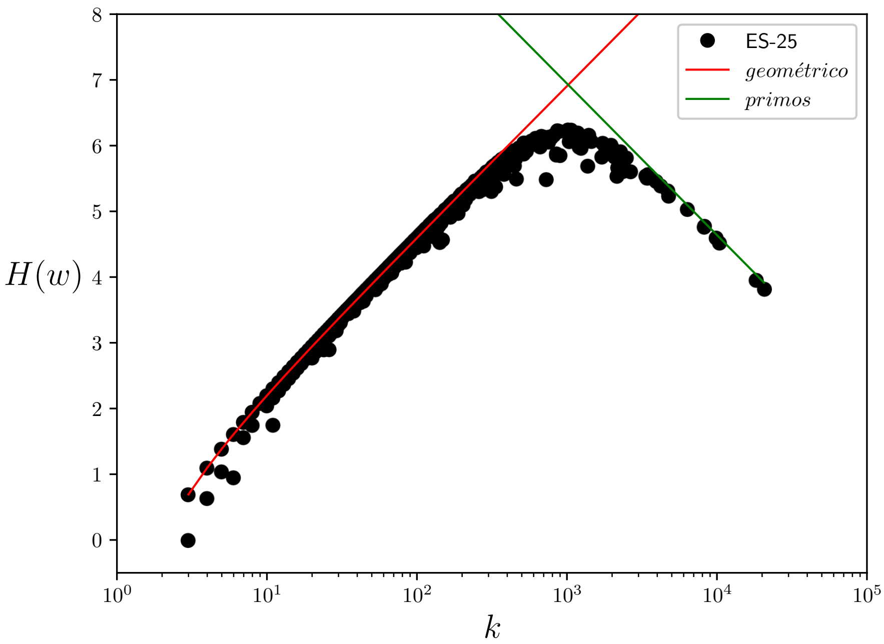


Figura 4.3: Gráfico log-linear da entropia $H(w)$ da distribuição de distâncias e o número de ocorrências k para ES-25 mostrando as curvas limitantes superior e inferior dadas pelas equações (4.6)) e (4.8), respectivamente.

Rolim estudou as propriedades da frequência k_0 e definiu que para $T \gg 1$, esse valor é:

$$k_0 = \sqrt{eT} \tag{4.9}$$

em que e é o número de Euler. Assim, a entropia máxima $H_{max}(w)$ seria dada por:

$$H_{max}(w) = \frac{1}{2}\ln(eT) . \tag{4.10}$$

A partir da análise de $H_{max}(w)$ e H_{max}, a última sendo a entropia máxima medida no texto, com o comprimento T para os textos do *corpus* de cada língua, Rolim detectou a existência de uma relação logarítmica do tipo:

$$H_{max} \sim v \ln T$$

em que o valor médio da constante é $\bar{v} \sim 0{,}49 \pm 0{,}01$ para as línguas dos textos presentes no *corpora*.

Modelo hamiltoniano

Carpena e colaboradores [32] criaram um modelo com motivação física no qual eles propuseram o uso de um hamiltoniano para a distribuição dos níveis de energia em cadeias com desordem correlacionada. Essa relação seria expressa pelo hamiltoniano $\mathscr{H}$ abaixo:

$$\mathscr{H} = \sum_{i} \xi_i \left|i\right\rangle \left\langle i\right| + \sum_{\langle i,j \rangle} U \left|i\right\rangle \left\langle j\right| \tag{4.11}$$

em que ξ_i representa a energia do sítio i, U é a energia de acoplamento entre os sítios i e seus primeiros vizinhos $\{j\}$, e $i = \{1, 2, ..., N\}$ em uma cadeia linear de tamanho N.

Para esse modelo, os sítios apresentam correlações de longo alcance segundo uma lei de potência [32]. Para introduzir as correlações entre os sítios, Carpena e colaboradores propuseram utilizar:

$$\xi_i = \sum_{u=1}^{N/2} \left[u^{-\alpha} \left(\frac{2\pi}{N} \right)^{1-\alpha} \right]^{1/2} \cos\left(\frac{2\pi i u}{N} + \phi_u \right) \tag{4.12}$$

em que ϕ_u são as $N/2$ fases aleatórias uniformemente distribuídas no intervalo $[0, 2\pi]$. A equação (4.12) é obtida quando se toma a transformada discreta inversa de Fourier para a variável u. Assim, por construção, o espectro de potências das séries $\{\xi_i\}$ é do tipo $u^{-\alpha}$. Dessa forma, o parâmetro α regula o grau de correlação espacial. Escolhendo-se os valores $\alpha = 0$ obtemos a descrição de um sistema descorrelacionado, $\alpha < 0$ o sistema possui anti correlação e $\alpha > 0$ representa correlações na série das energias $\{\xi_i\}$.

Os resultados obtidos por Carpena e colaboradores para a intermitência como função de α foram reavaliados por Rolim, no qual o parâmetro $\sigma_\alpha(N)$ revelou ter as seguintes características: para todos os valores de α e N, o modelo não é capaz de gerar distribuições espaciais super-poissonianas ($\sigma > 1$). Essa característica é fundamental para uma distribuição que gere os padrões de σ com k comuns à linguagem natural (as Figuras 2.8 e 4.1 explicitam esses padrões).

Porém, como discutido na subseção 2.2.3, os verbetes na região inferior possuem um regime sub-poissoniano $\sigma < 1$ para baixas frequências e em altas frequências encontramos um regime poissoniano $\sigma = 1$. Sendo a última uma região com verbetes que possuem distribuições espaciais descorrelacionadas. Assim, Rolim propôs o uso do modelo descrito acima para $\alpha = 0$ em um sistema de tamanho $N = k$, ao fazer isso ele obteve a Figura 4.4 que mostra os resultados da intermitência como função da frequência.

A Figura 4.4 mostra que o modelo proposto descreve de maneira adequada a região limitante inferior da distribuição da intermitência com a frequência. Porém, como evindenciado por Rolim, esse modelo não é o único capaz de reproduzir uma distribuição descorrelacionada e que apresente o comportamento limitante inferior.

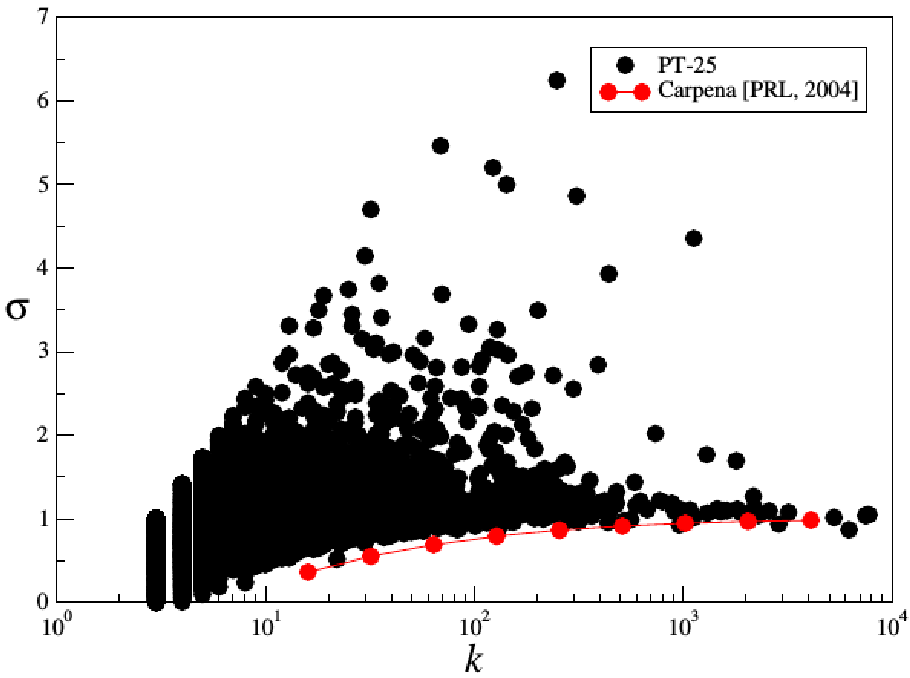


Figura 4.4: Gráfico log-linear da intermitência em função da frequência k para o modelo hamiltoniano e o texto PT-25. Para se obter os pontos em vermelho, Rolim considerou $k = N$ e $\alpha = 0$. A Figura foi extraída de [38].

Os resultados dos três modelos propostos por Rolim nos revelam as seguintes propriedades para o processo que origina distribuição espacial dos verbetes em textos:

1) ele possui características que fazem com que certos verbetes tenham uma distribuição geométrica das posições de suas ocorrências formando o conjunto de distâncias de

maior intermitência (com valores sempre maiores que 1) para todos os valores de k;

2) ao mesmo tempo, esse processo gera um conjunto de distâncias para outros verbetes com valores de $\sigma < 1$ para baixas frequências e $\sigma \sim 1$ quando $k \to \infty$, tendo os modelos hamiltonianos e de primos consecutivos como boas aproximações desses processos nessa região.

No entanto, pouco pode ser afirmado sobre como seria o processo em si e qual seria a distribuição de probabilidade que possui as características necessárias para apresentar os dois regimes. A partir desse momento, o trabalho se voltará à análise desses processos, suas propriedades e equivalências com textos em linguagem natural.

Modelos estocásticos para criação de textos genéricos

Como visto na seção anterior, a distribuição de números primos consecutivos possui uma distribuição sub-poissoniana $\sigma < 1$ no regime de frequências observados nos textos utilizados em liguagem natural. No entanto, um processo que considere as distâncias entre os números primos consecutivos pode gerar distribuições super-poissonianas. Para isso, é suficiente que existam diversidade de distâncias e regiões de agrupamento entre as ocorrências sucessivas. Assim, propomos um modelo para a distribuição espacial de verbetes na qual suas ocorrências estão separadas por distâncias que separam os números primos consecutivos.

Outra abordagem bastante distinta, mas justificada também pela presença de comportamentos super-poissoniano, poissoniano e sub-poissoniano, é o uso da distribuição de Weibull para as distâncias entre as sucessivas ocorrências. O uso desse tipo de distribuição foi proposto por Altmann e colaboradores [34] e exibe resultados coerentes para a distribuição espacial de verbetes como será apropriadamente discutido na subseção 4.2.2. Dito isto, sugerimos o uso da distribuição de Weibull na criação de um texto genérico.

Distância entre primos consecutivos

A distribuição de probabilidade para os espaçamentos entre números primos sucessivos dada pela equação 4.7 se refere a um limite assintótico quando possuímos uma sequência de números primos suficientemente grande [70]. Contudo, sequências menores podem gerar regimes super-poissonianos $\sigma > 1$.

Assim, pode-se propor o uso das distâncias entre primos sucessivos como distribuição de distâncias num processo de criação de textos genéricos. Um dos possíveis algoritmos para criação de textos dessa maneira tem seus passos detalhados abaixo:

1) determinamos o comprimento T do texto;

2) encontra-se quantos números primos existem até T e cria-se uma lista das distâncias entre as sucessivas ocorrências;

3) a primeira ocorrência de um verbete se dá na primeira posição livre do texto e as demais serão postas em posições não ocupadas separadas por distâncias sorteadas aleatoriamente na lista obtida no passo 2);

4) quando não houver mais posições possíveis dentro do tamanho especificado do texto, determina-se a frequência desse verbete e repete-se o passo 3) com um novo verbete até que o texto esteja completamente preenchido.

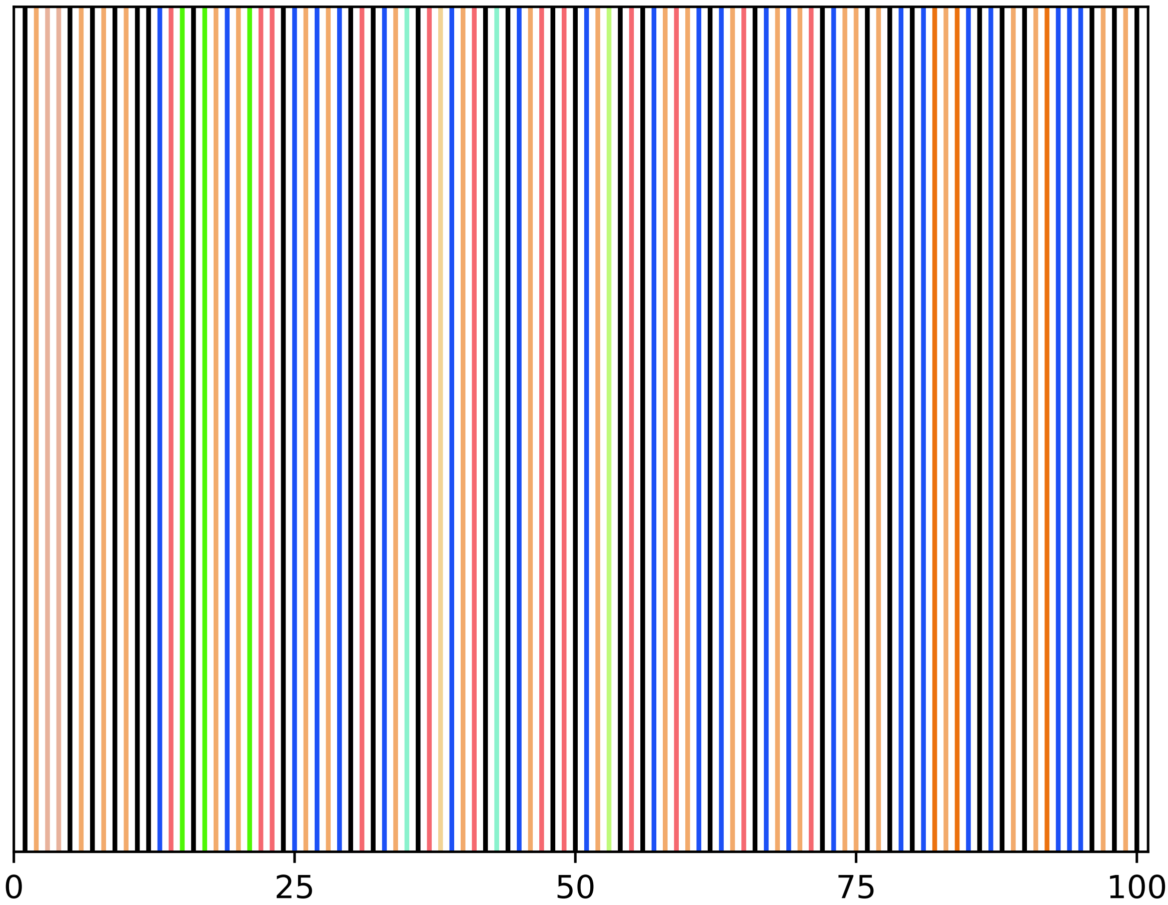
Executando o algoritmo acima para um texto de tamanho $T = 100$, obtém-se o padrão ilustrado na Figura 4.5. Observando os padrões formados, é possível perceber o comportamento esperado para verbetes com $\sigma > 1$ em que há aglomerados distintos de ocorrências separados por ocorrências individuais e outros aglomerados.

Podemos fazer também o gráfico das leis de Zipf e Heaps para os três maiores textos produzidos com esse algoritmo e as ilustramos na Figura 4.6. Vemos que os expoentes de Zipf mostrados são compatíveis com os observados em linguagem natural (como discutido nas subseções 2.1.1 e 2.1.2), enquanto os expoentes de Heaps possuem valores semelhantes aqueles encontrados por Pola [59].

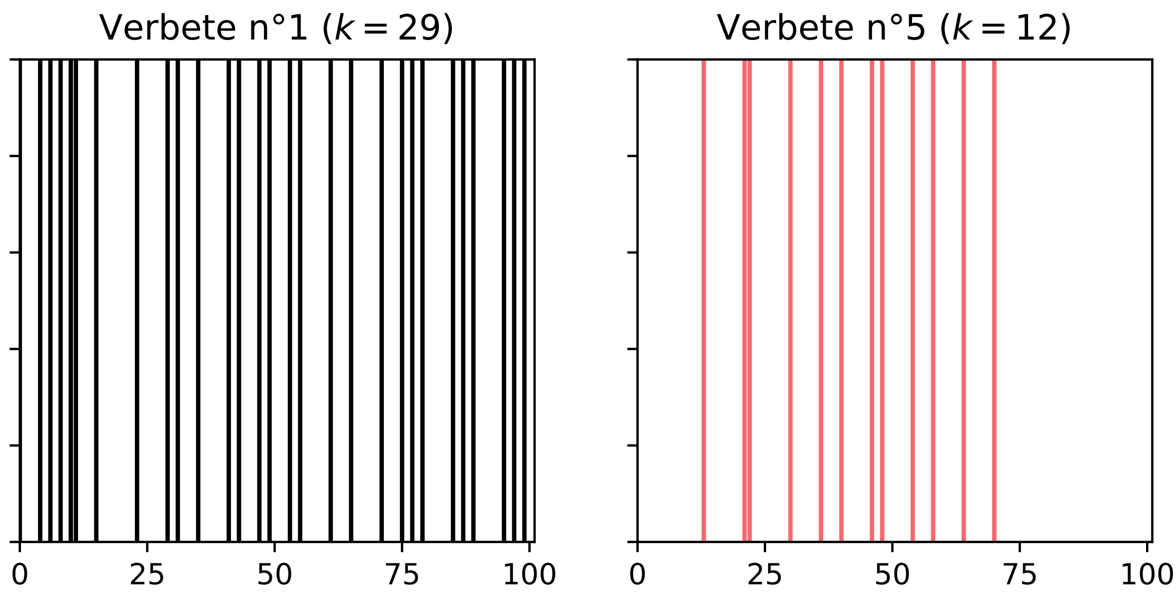
Determinando a intermitência e entropia para cada verbete dos textos, podemos criar os gráficos apresentados na Figura 4.7. Neles, observam-se algumas características importantes da distribuição de distâncias resultante:

i) os verbetes possuem regimes super-poissonianos, poissonianos e sub-poissonianos em todos os casos;

ii) todos eles se encontram entre as regiões limitantes como observado em textos reais (tema mais profundamente discutido na seção 4.1);



(a) Todos os verbetes do texto.



(b) Dois verbetes que possuem distribuição de distâncias típicas.

Figura 4.5: Espectro das posições dos verbetes em um texto genérico de tamanho $T = 100$ com distâncias dadas pela distribuição de distâncias entre números primos sucessivos. Cada verbete foi representado por uma cor distinta e o vocabulário do texto possui tamanho $V = 10$.

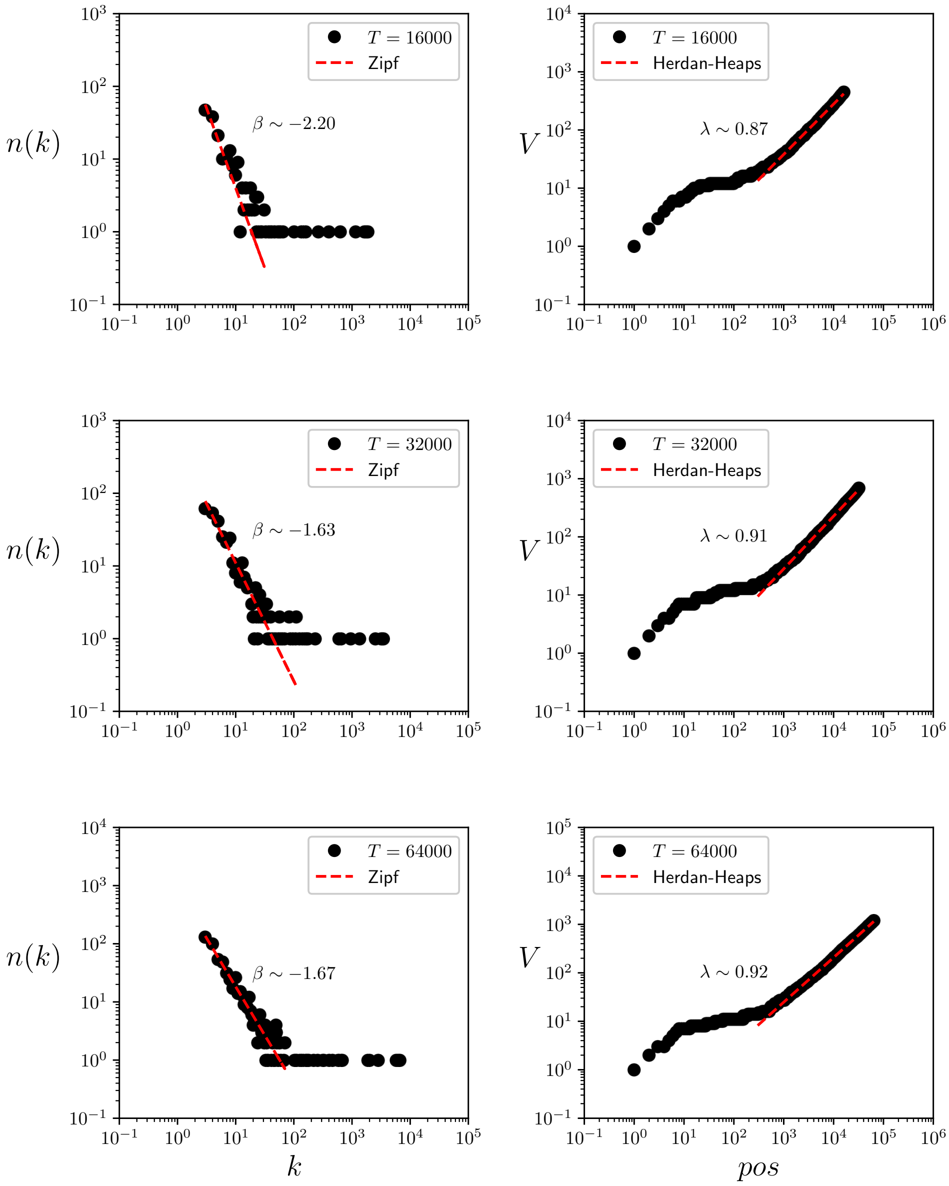


Figura 4.6: Gráfico log-log das leis de Zipf e Herdan Heaps para os três maiores textos genéricos gerados pelo modelo de distâncias entre números primos consecutivos. Os gráficos trazem os tamanhos dos textos e os valores dos expoentes da regressão.

iii) as entropias que surgem desse tipo de modelo possuem um comportamento bastante distinto daquele discutido para textos reais para $k > 10$.

Outra característica importante, é o fato da frequência máxima k_{max} ser maior do que os valores típicos em textos. Para determinar qual deveria ser o valor esperado para k_{max}, utilizamos a lei de potência proposta por Rolim [38], na qual a frequência máxima k_{max} nos textos presentes no *corpora* obedece uma lei de potência com o tamanho do texto T descrita por:

$$k_{max} = (0{,}048 \pm 0{,}015)\, T^{0{,}92}\,. \tag{4.13}$$

Sendo assim, os valores esperados para esses textos seriam respectivamente $k^1_{max} \sim 354$,

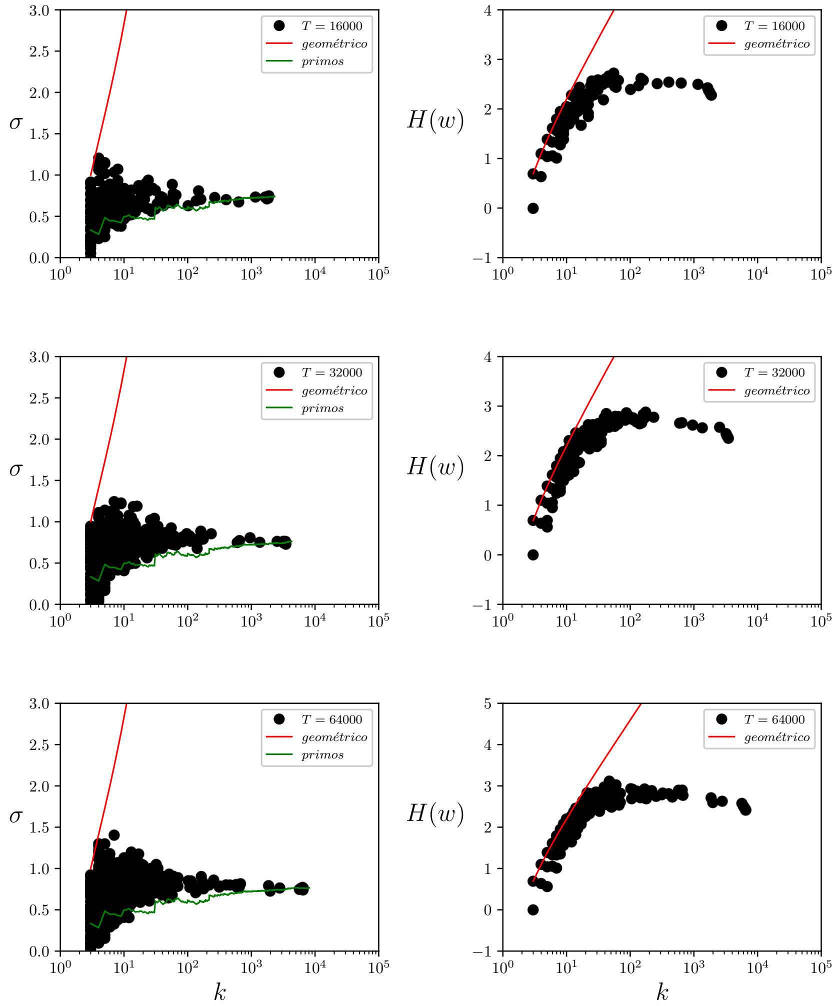


Figura 4.7: Gráficos da intermitência e entropia como função da frequência para os três maiores textos genéricos gerados modelo de distâncias entre números primos consecutivos. Foram adicionadas as curvas para o modelo geométrico e o modelo dos primos consecutivos ilustrando as regiões limítrofes.

$k^2_{max} \sim 670$ e $k^3_{max} \sim 1267$ para os textos de tamanho $T = 16{,}000$, $T = 32{,}000$ e $T = 64{,}000$.

Essa discrepância entre os valores encontra justifcativa nos passos 3) e 4) do algoritmo do modelo. Por utilizar uma distribuição em que distâncias curtas são muito mais frequentes que longas distâncias, o texto vai sendo ocupado rapidamente pelos primeiros verbetes. Isto, por sua vez, permite a existência de verbetes com frequências muito maiores que as dos demais.

Ao determinar as flutuações $F(\tau)$ e $F_{CC}(\tau)$ para o DFA-4 e DCCA-4 das séries temporais das frequências e intermitências[1], podemos as ilustrar como uma função

[1] Os "verbetes" criados em textos genéricos são simplesmente marcações da existência de um elemento distinto aos que foram previamente sorteados, logo eles não possuem tamanho. Isto, por sua vez,

do comprimento τ gerando os gráficos exibidos na Figura 4.8.

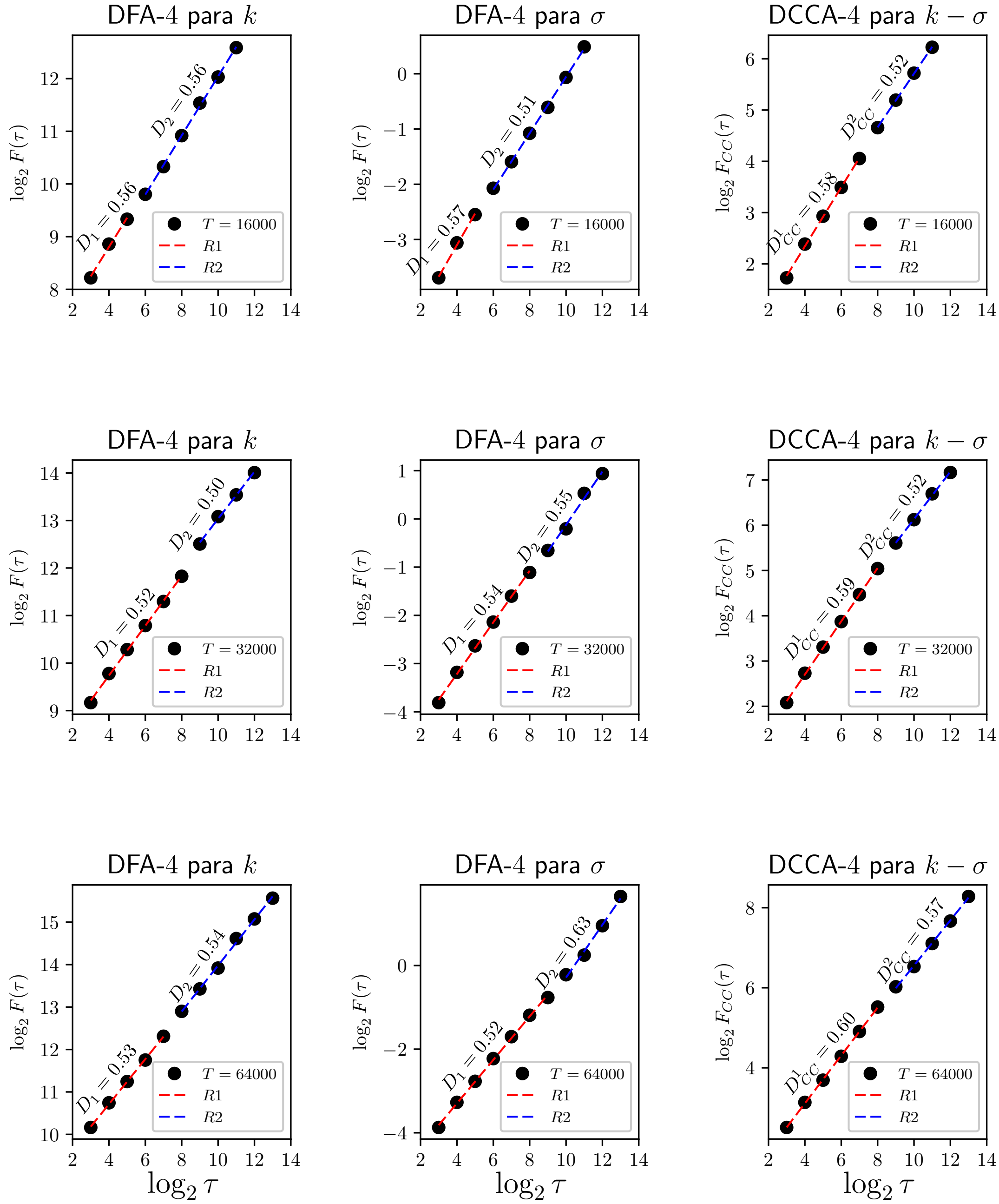


Figura 4.8: Gráficos das flutuações para DFA-4 e DCCA-4 como função do comprimento τ para os três maiores textos gerados pelo modelo de distâncias entre números primos consecutivos. As respectivas regressões e seus expoentes são mostrados nos gráficos.

Os valores dos expoentes nos regimes $R1$ e $R2$ indicam que:

i) a frequência deveria possuir um único regime levemente correlacionado para o DFA. Ao lembrarmos dos resultados obtidos na seção 3.2 em que se discute a existência de dois regimes distintos para as frequências em textos reais, podemos afirmar que o

impossibilita a determinação de uma série temporal dos comprimentos dos verbetes.

modelo aqui proposto não é capaz de gerar os padrões de frequências observados em linguagem natural;

ii) a intermitência possui dois regimes bem definidos para os textos mostrados no DFA. Apesar do primeiro regime se mostrar levemente correlacionado, o mesmo não pode ser afirmado sobre o segundo regime, uma vez que seus valores são bastante distintos. O primeiro regime encontra paralelo quantitativo com os achados em textos reais;

iii) as flutuações do DCCA entre as séries das frequências e intermitência apontam a existência de dois regimes. No entanto, diferentemente do encontrado em textos reais, os valores do segundo regime indicam que as séries possuem leves correlações cruzadas em longas escalas.

Estes resultados combinados indicam que a distribuição é somente capaz de gerar correlações de curto alcance, isso se deve à propriedade da distribuição de distâncias entre números primos consecutivos explicada nos parágrafos anteriores.

Portanto, apesar de gerar valores esperados para os expoentes de Zipf e Herdan-Heaps, apresentar padrões semelhantes para a intermitência como função da frequência e expoentes em $R1$ que são semelhantes aos observados em textos reais, esse modelo não produz uma das características mais comuns dos textos em linguagem escrita: as estruturas de longo alcance correlacionadas. Uma das abordagens possíveis é adotar uma distribuição de distância que possua em sua definição a possibilidade de criação de estruturas de longo alcance. Para tal, adotamos o uso da distribuição de Weibull.

Distribuição de Weibull

Altmann e colaboradores [34] sugeriram que as distâncias entre sucessivas ocorrências d de um verbete de frequência k poderia ser apropriadamente modelada se utilizada uma distribuição de Weibull $f_\gamma(d)$ com γ do tipo:

$$f_\gamma(d) = a\gamma d^{\gamma-1} e^{-ad^\gamma} . \tag{4.14}$$

Para obter a, faz-se a normalização da distribuição acima e temos:

$$a = \left[\nu\Gamma\left(\frac{\gamma+1}{\gamma}\right)\right]^\gamma \tag{4.15}$$

em que Γ é a função gama, $0 < \gamma \leq 1$ e ν é o inverso da distância média:

$$\nu = \frac{k-1}{T}.$$

Se imaginarmos um processo em que d é o tempo em que ocorre um fracasso da variável desejada, a distribuição de Weibull nos fornece uma distribuição na qual a taxa de fracassos é proporcional a uma potência do tempo d. Portanto, a escolha dos valores do expoente se justifica pelo fato de $\gamma < 1$ representar a região em que a taxa de fracasso diminui com o tempo e $\gamma = 1$, o expoente em que a taxa de fracasso se mantém constante com d.

Dada as devidas justificativas acima, Altmann e colaboradores utilizaram como texto a discussão do grupo *talk.origins* na plataforma USENET e consideraram somente as 2.128 palavras que possuíam frequência $k \geq 10.000$. Ao tomar a regressão para cada verbete utilizando a equação (4.14), eles obtiveram o gráfico da Figura 4.9 que ilustra o expoente de cada verbete como função da distância média $\frac{T}{k-1}$.

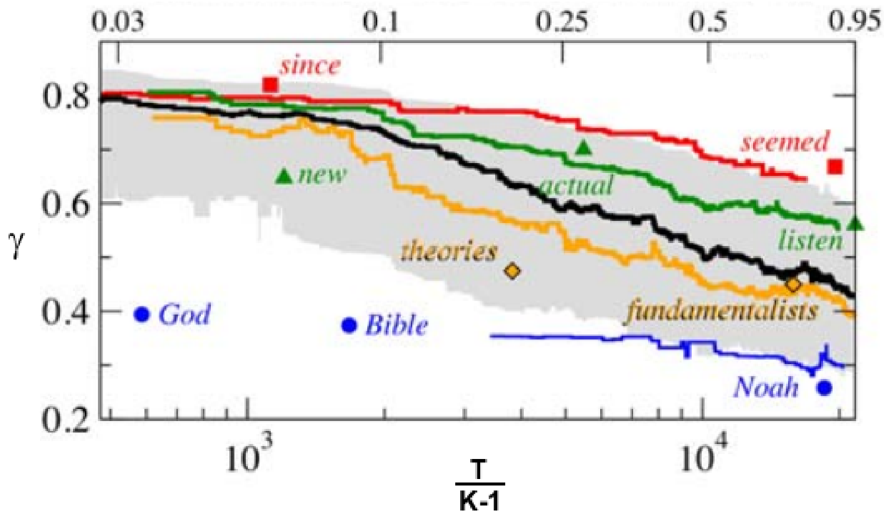


Figura 4.9: Gráfico log-linear entre γ e $\frac{T}{k-1}$ para o grupo de discussão *talk.origins* na USENET. Foi feita uma selação de verbetes com frequências distintas para ilustrar as propriedades dos valores do expoente γ. Figura adaptada de [34].

O gráfico acima ilustra os valores possíveis de γ para os verbetes que preenchem os requisitos fixados por Altmann e colaboradores. Vemos que existe uma região bem definida para os valores do expoente $0{,}25 < \gamma < 0{,}85$, na qual os verbetes com menores valores tendem a ser palavras-chave no tópico discutido e os com maiores expoentes são palavras comuns.

Partindo desses resultados, propomos a criação de um texto artificial utilizando a distribuição de Weibull para determinar a distância entre as sucessivas ocorrências. Para a criação do texto artificial, adotamos os seguintes passos:

1) determinamos o comprimento T do texto;

2) sorteia-se um valor para o expoente γ de um verbete no intervalo $0{,}1 < \gamma < 1$;

3) definido γ, escolhe-se através de um sorteio a sua frequência que pode estar no intervalo $1 < k < T_{disp}$, em que T_{disp} é o número de posições disponíveis no texto;

4) a primeira ocorrência é definida alternadamente entre a primeira e última posição livre do texto, quando mudam-se os verbetes. As demais ocorrências estarão distribuídas aleatoriamente por distâncias que obedecem uma distribuição de Weibull com expoente γ e parâmetros T e k para o valor de a;

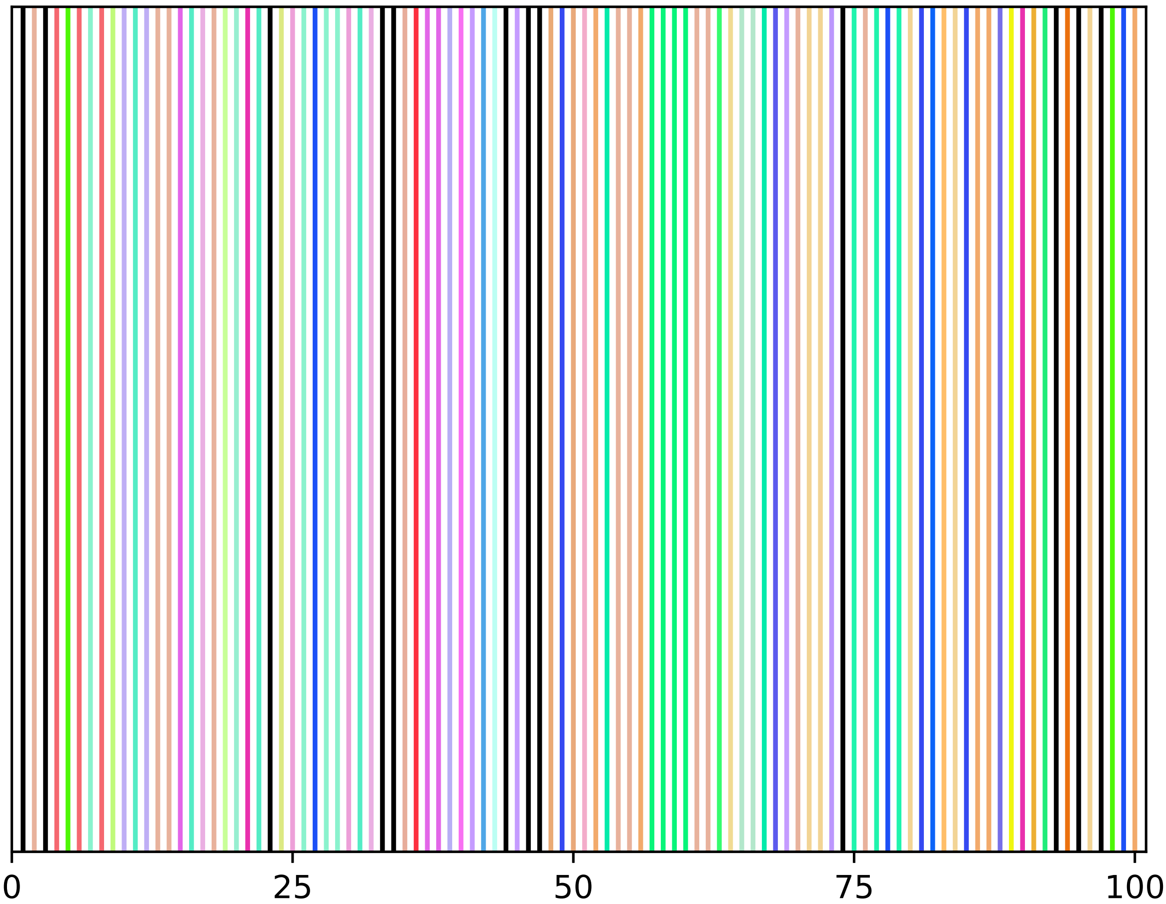
5) se alguma das posições exceder o comprimento do texto ou se a distância sorteada for igual a 0, retorna-se ao passo 3);

6) caso todas as ocorrências estejam em posições válidas, computa-se as ocorrências do verbete no texto e repete-se os passos 1-5 até que o texto esteja completamente preenchido.

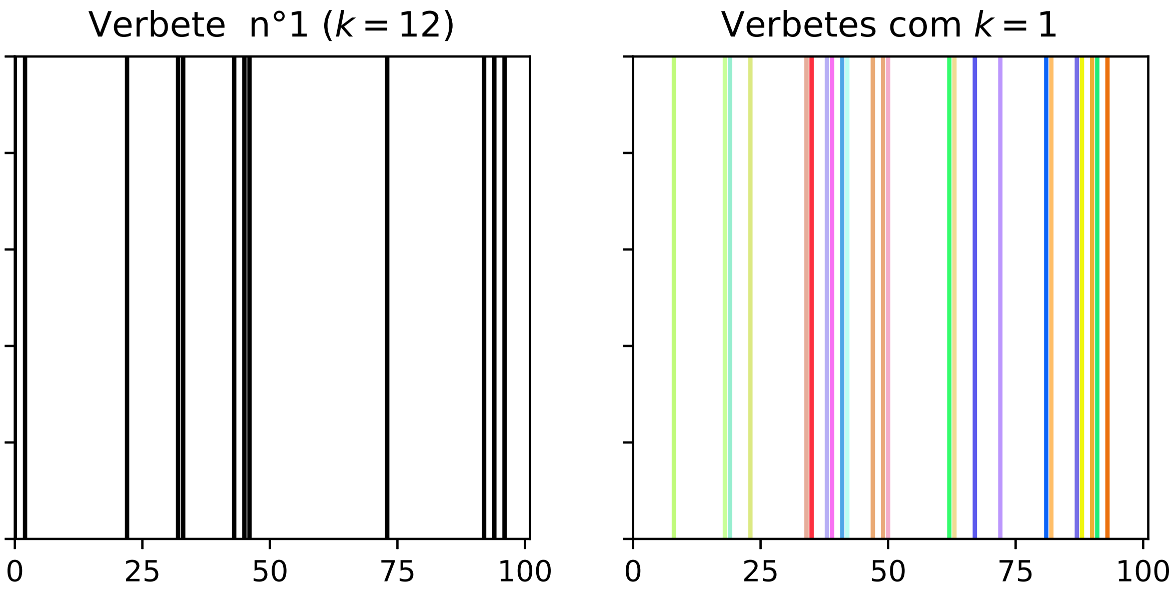
Ao executar o algoritmo descrito acima para um texto de tamanho $T = 100$ obtemos um padrão como mostrado na Figura 4.10. Ao observar atentamente os padrões formados no espectro de posições, vemos que esse modelo produz um comportamento semelhante aos presentes em textos, tais como a formação de aglomerados que são separados por ocorrências individuais ou outros aglomerados.

Porém, ele mostra outro comportamento, este com características não correlatas ao que é observado em linguagem natural. Vemos que certas regiões são majoritariamente preenchidas por verbetes com poucas ocorrências. Esse comportamento se justifica pelos valores de γ, uma vez que eles determinam quais valores de k são possíveis devido ao passo 5) do algoritmo. Sendo assim, os textos genéricos criados por esta abordagem apresentam distorções nas propriedades comuns a textos em linguagem natural como veremos a seguir.

A partir de textos genéricos podemos determinar as leis de Zipf e Herdan-Heaps como mostradas na Figura 4.11 para os três maiores textos gerados através desse



(a) Todos os verbetes do texto.



(b) Verbete mais frequente e todos os verbetens com frequência $k = 1$.

Figura 4.10: Espectro das posições dos verbetes em um texto genérico de tamanho $T = 100$ com distâncias dadas pela distribuição de Weibull. Cada verbete foi representado por uma cor distinta e o vocabulário do texto possui tamanho $V = 44$.

algoritmo. Vemos que os valores para o expoente de Zipf são aproximadamente $\beta \sim 2{,}50$ e para o expoente de Herdan-Heaps $\lambda \sim 0{,}90$. Esses valores possuem duas interpretações:

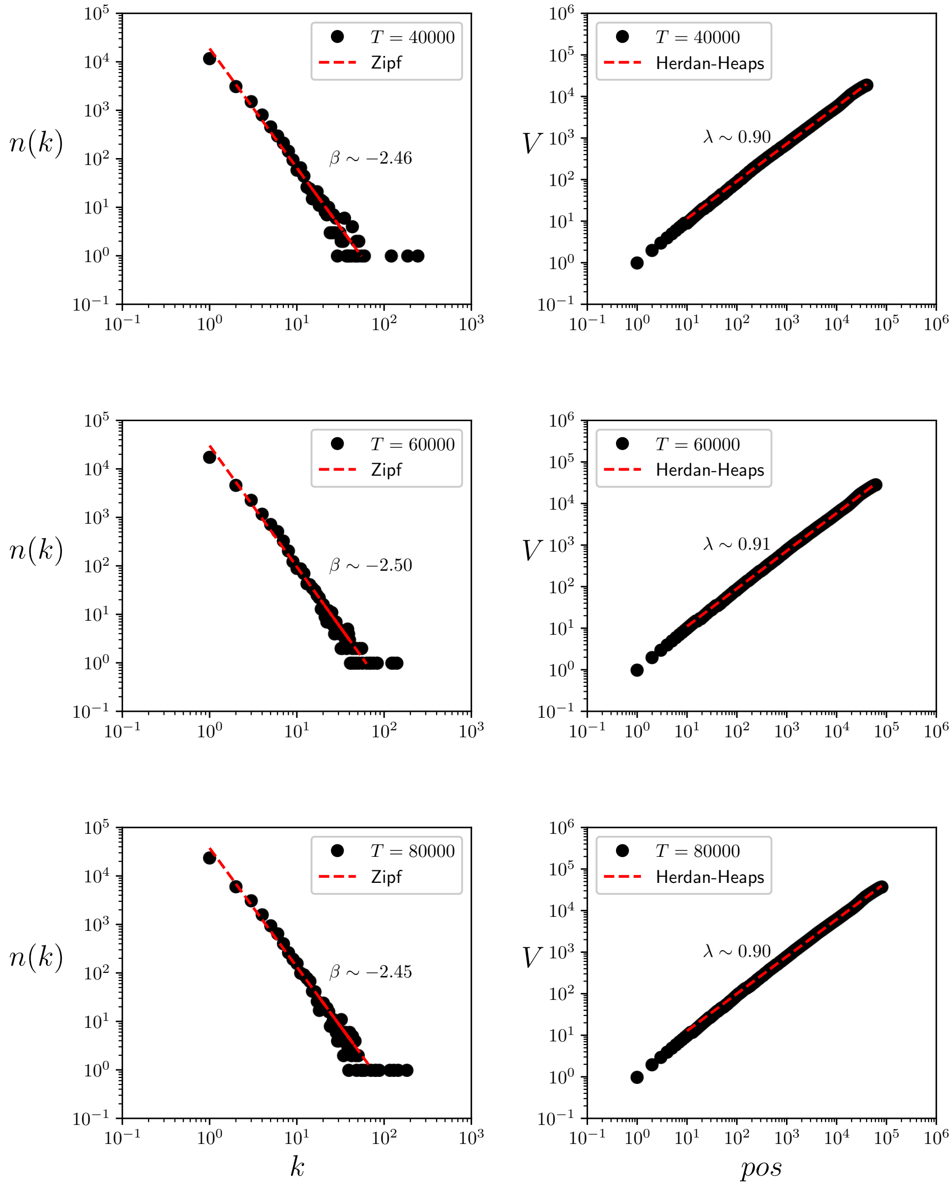


Figura 4.11: Gráfico log-log das leis de Zipf e Herdan Heaps para os três maiores textos genéricos gerados pelo modelo de Weibull. Os gráficos trazem os tamanhos dos textos e os valores dos expoentes da regressão.

i) o valor de $\beta \sim 2{,}50$ indica que os verbetes com frequências menores são muito mais numerosos do que em um texto real. A presença dessa quantidade maior de verbetes foi justificada anteriormente;

ii) o valor de $\lambda \sim 0{,}90$ apesar de distinto do valor médio por língua, ele é compatível com os valores do expoente em textos como evidenciado por Pola [59].

De maneira similar, podemos determinar as intermitências e entropias para cada um dos verbetes nesses textos e gerar o gráfico da Figura 4.12, ilustradas em conjunto com suas curvas limitantes. Vemos que os valores para a intermitência e entropia possuem comportamento semelhante aos discutidos em textos reais (ver a discussão completa na seção 4.1).

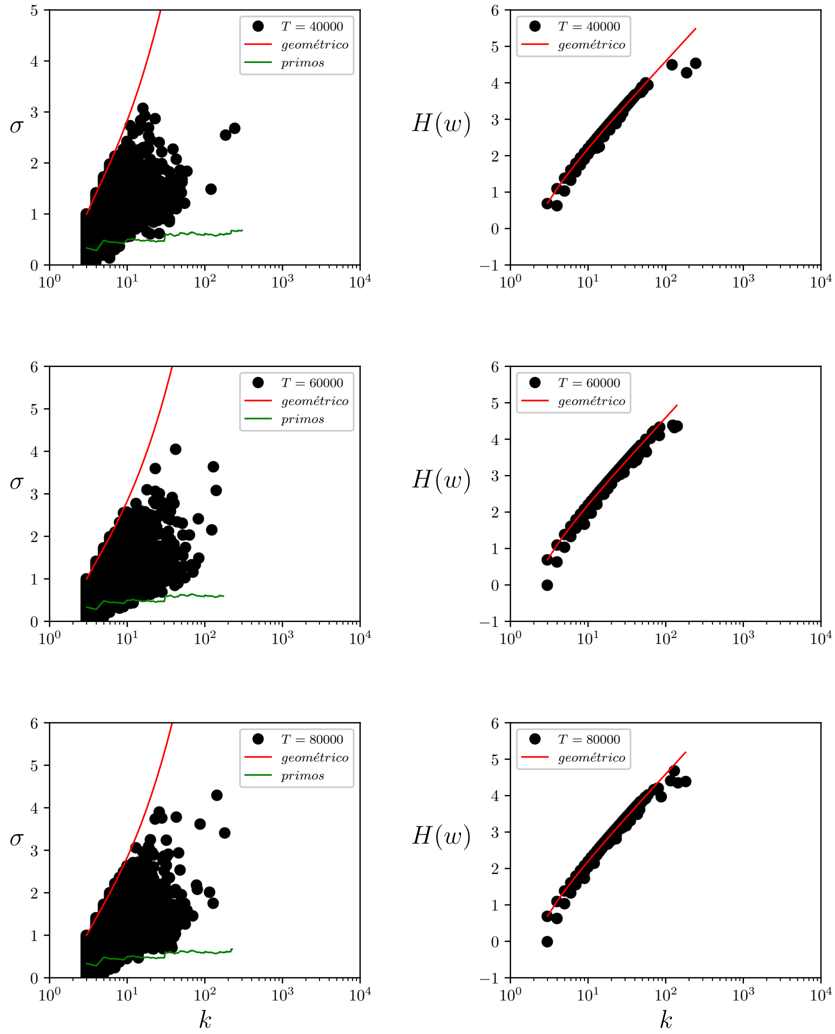


Figura 4.12: Gráficos da intermitência e entropia como função da frequência para os três maiores textos genéricos gerados pelo modelo de Weibull. Foram adicionadas as curvas para o modelo geométrico e o modelo dos primos consecutivos ilustrando as regiões limítrofes.

No entanto, os valores das frequências máximas k_{max} obtidas podem diferir em uma ordem de grandeza em relação aos encontrados em textos em linguagem natural que possuem tamanhos semelhantes. Utilizando a equação (4.13), obtemos que $k^1_{max} \sim 822$, $k^2_{max} \sim 1194$ e $k^3_{max} \sim 1556$ para os textos de tamanho $T = 40{,}000$, $T = 60{,}000$ e $T = 80{,}000$, respectivamente. Este fato é novamente justificado pelo mecanismo contido entre os passos 2) e 5) do algoritmo do modelo.

Calculamos também as flutuações $F(\tau)$ e $F_{CC}(\tau)$ do DFA-4 e DCCA-4 para as séries temporais da frequência e intermitência. Como é mostrado na Figura 4.13, temos novamente dois regimes com expoentes distintos para as flutuações e seus valores são semelhantes aos encontrados para textos reais (ver as seções 3.2 e 3.3).

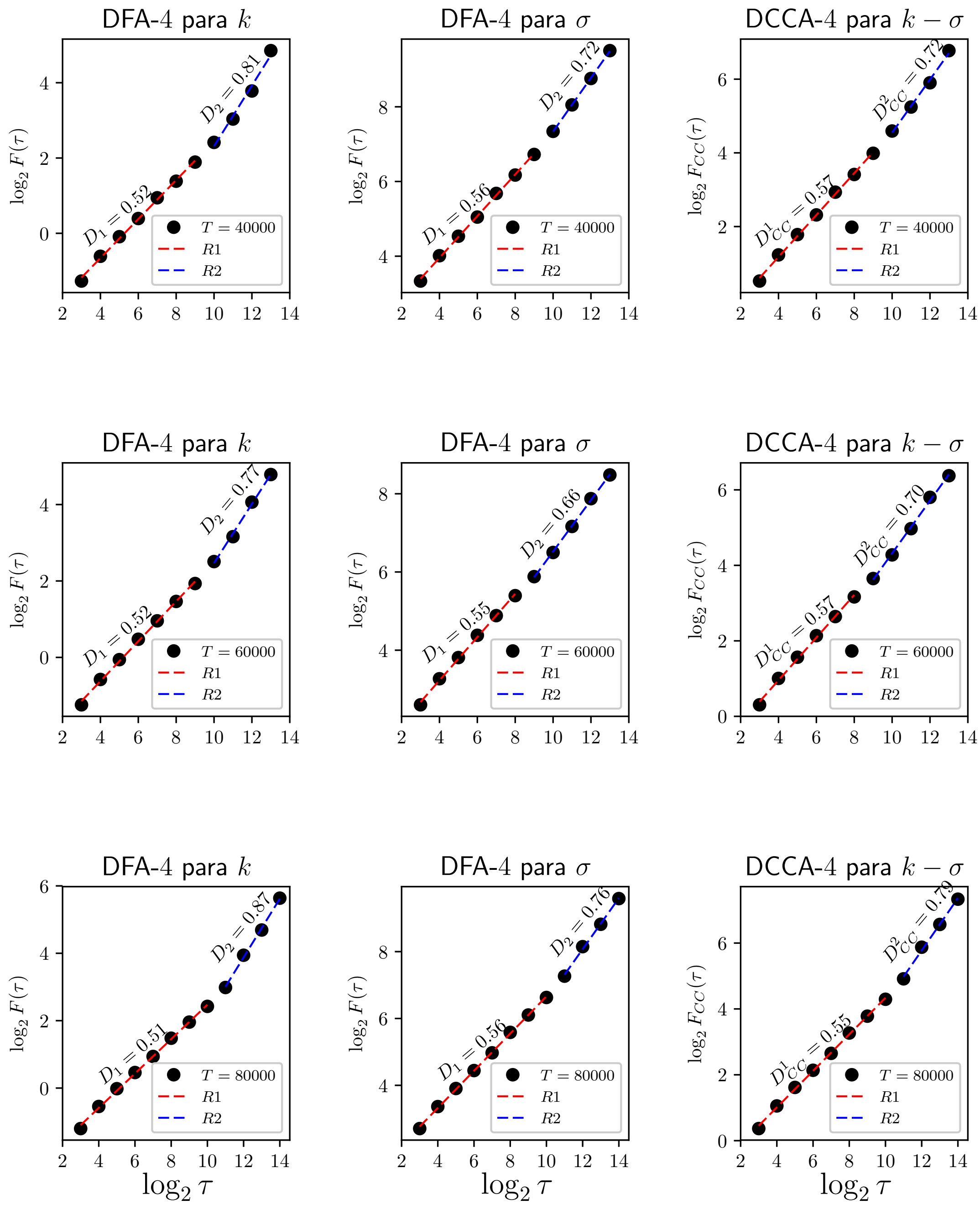


Figura 4.13: Gráficos das flutuações para DFA-4 e DCCA-4 como função do comprimento τ para os três maiores textos gerados pelo modelo de Weibull. As respectivas regressões e seus expoentes são mostrados nos gráficos.

A presença dessas propriedades semelhantes às exibidas em textos em linguagem natural indica que esse processo de criação de textos genéricos possui paralelos qualitativos com aquele apresentado pela escrita. Em especial, os resultados apontam que a distribuição de distâncias entre as sucessivas ocorrências em textos reais tem um comportamento bem aproximado pela distribuição de Weibull.

Conclusões e perspectivas

Essa dissertação foi voltada ao estudo de características frequencistas e estruturais da distribuição de verbetes em textos e suas respectivas contribuições para compreender as propriedades presentes na escrita em linguagem natural. Os resultados encontrados foram comparados entre as 485 obras, divididas entre textos literários e verbetes da wikipedia, em 10 idiomas distintos pertencentes a três famílias linguísticas.

Ao analisar as séries temporais de frequência, comprimento e intermitência extraídas dos textos, mostramos a existência de dois regimes distintos de autocorrelação em todas as séries. Tais resultados sugerem a existência de, ao menos, 2 escalas distintas de correlação: uma de curto alcance que possui tamanhos variando entre comprimentos de frases até capítulos, a outra possuindo um comportamento de longo alcance, relacionada com própria estrutura do texto.

As análises utilizadas mostraram a presença de correlações cruzadas com comportamento semelhante às autocorrelações. A presença desse tipo de correlação indica que o processo de escrita produz esses padrões para frequência, tamanho e intermitência dos verbetes de maneira bastante direta e fortemente correlacionada em regimes de longo alcance.

Em seu segundo momento, este trabalho propõe a criação de textos genéricos a partir de distribuições de distâncias distintas. Para tal, foram escolhidas as distribuições de distâncias entre números primos sucessivos e a distribuição de Weibull. Ambas foram escolhidas por serem capazes de reproduzir regimes super-poissonianos, poissonianos e sub-poissonianos.

O modelo para distância entre sucessivos números primos mostrou comportamentos semelhantes aos textos reais para as leis de Zipf e Herdan-Heaps, para a distribuição de intermitências com as respectivas frequências dos verbetes e nas correlações em curtas escalas. No entanto, por gerar textos a partir de uma distribuição que não contém, por definição, mecanismos que permitam a criação de estrutura de longas escalas, o modelo falha em descrever propriedades fundamentais comuns aos textos reais.

O segundo modelo, no qual utilizamos uma distribuição de Weibull para as distâncias entre sucessivas ocorrências, apresentou propriedades como as leis de Zipf e Herdan-Heaps, distribuição semelhantes das intermitências em função da frequência e regimes de correlação em curtas e longas escalas com expoentes próximos aqueles observados em linguagem natural.

Tal resultado indica que o processo de escrita deve possuir um mecanismo semelhante ao apresentado no modelo. No entanto, as análises feitas indicam que ele não é capaz de gerar os valores esperados para a frequência máxima previstos. Essa deficiência do modelo é justificada em sua definição ao limitar os valores de frequências possíveis quando determinado o valor do expoente da distribuição de Weibull.

Esta deficiência sugere a criação de um novo algoritmo utilizando distribuições do tipo exponencial esticada que seja menos restringente aos valores das frequências possíveis. Outro possível desdobramento desse trabalho está no estudo do valor do expoente da distribuição utilizada e seus intervalos de valores típicos nas línguas escritas, criando a possibilidade de gerar textos genéricos representativos de uma certa língua.

Apêndice A

Os textos utilizados ao longo dessa dissertação foram os mesmos utilizados por Rolim [38], todos extraídos do *Project Gutenberg* [108] e do Portal Domínio Público [109] para as obras em português. Foi feita a escolha de 25 livros, dos quais romances totalizam a maioria, em dez idiomas: alemão, dinamarquês, espanho, finlandês, francês, húngaro, inglês, italiano, português e sueco. A seleção foi feita tal que todas as línguas possuíssem livros entre o mesmo intervalo de comprimento $10^4 < T < 10^5$. As próximas páginas trarão a referência da obra (código utilizado ao longo da dissertação para as obras), livro, autor(es) e o *release* na fonte utilizada, com PT indicando *Project Gutenberg* e DP, o Domínio Público.

O Apêndice B trará os artigos retirados da wikipedia com suas respectivas referências, verbete na língua original, data de acesso e 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	Herman Löns	PG October 2, 2007 Ebook #22824
DE-05	Die Luftschiffahrt der Gegenwart	Herman 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 Ebook #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 Freudin	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	Kaptjnen 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	Bjørnaæt	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 Nytaarrstid 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	PG February 22, 2011 Ebook #16484
ES-03	Fiebre de amor (Dominique)	Euène 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 15, 2004 Ebook #12848
ES-13	La letra escarlatta	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 bejo	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 allhaiset	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 Helminkangas	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 Boysleve	PG February 20, 2009 EBook #28124
FR-05	Biribi	Georges Darien	PG August 8, 2005 EBook #16492
FR-06	La tombe de fer	Hendrike 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	Jehann Bagnyon	PG November 27, 2013 Ebook #44301
FR-10	Le créspucule 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 Boysleve	PG July 23, 2006 Ebook #18899
FR-18	Le blé que 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	Dezsö 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 széspségei; A legvitézebb huszár	Mór Jókai	PG December 10, 2012 Ebook #41601
HU-17	A przemysli repülö 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 e 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	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 Ebook #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 e 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 Ebook #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 9, 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 Mysteria da Estrada de Cintra	Eça de Queirós e 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 Garret	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 Brazilio	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	Förvillelser	Hjalmar Söderberg	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 kosuln	Mathilda Malling	PG December 18, 2007 Ebook #23891
SE-16	Modern	Ernst Ahlgren e 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 Lagerlöf	PG March 14, 2012 Ebook #39147
SE-22	Vi Bookar, Krokar och Rothar	Hjalmar Bergman	PG April 28, 2005 Ebook #15724
SE-23	Hemsöborna	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

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	Bergman	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.wikipedia.org/wiki/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	
Referência	Url para download
DKW-01	https://mega.nz/#!o0R0naTT!m06T2gFJ20g5vgPbjbTscHzdji3ZYJwksUbblEQocXU
DKW-02	https://mega.nz/#!J8RCxKrB!TwVBxYiuAvGCp5bd5jHZCE5ZZzGhWe69nvepXt4amZM
DKW-03	https://mega.nz/#!sswlEAhI!jLVsotMMTCl26jDLfuIm-dW2QXLmzzE-NiqTLygD8m8
DKW-04	https://mega.nz/#!01wjDChD!HXcuJVjOaMQiP16CG1_hqTEPsIUGJOHnp2pJ3srHu0k
DKW-05	https://mega.nz/#!1o5lyQBQ!01tY8SwI3P-lT4RyDS1OZ7BuqCfvF0ngVzEG5D5EYug
DKW-06	https://mega.nz/#!dxI1jCyK!USgPhmDNMqHgUuCYjFP3i0UPJj5Iefks9fgztSz_5wc
DKW-07	https://mega.nz/#!wp52lQRL!Dr2Nbhy62SEP1oWBEYkkvEqJU01AbuTovD1BRSN80kQ
DKW-08	https://mega.nz/#!JppC2KbZ!LlOWN4fqQwEWK41mAvggKh-x04aF-Oq1p48esDXuqPI
DKW-09	https://mega.nz/#!Q1p12b6I!Djyb9DMrNXrgE1mNmwOwEJy9EpYnUeM1IBksraNHh5E
DKW-10	https://mega.nz/#!N9oWhQKb!cpvUlXP5AGSOEfZwDfwuOijr0C5qKw4CV-KzcPyBDsA

Espanhol			
Referência	Verbete	Data de acesso	Endereço
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	Deida	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.wikipedia.org/wiki/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 (geografia)	19/07/2014	http://es.wikipedia.org/wiki/Litoral_(geografia)
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	Mitologia	19/07/2014	http://es.wikipedia.org/wiki/Mitologia
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			
Referência	Verbete	Data de acesso	Endereço
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 maailmasota	19/07/2014	http://fi.wikipedia.org/wiki/Ensimmäinen_maailmasota
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.wikipedia.org/wiki/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/Liberalismi
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	Metafysiika	19/07/2014	http://fi.wikipedia.org/wiki/Metafysiika
FIW-16	Modernismi	19/07/2014	http://fi.wikipedia.org/wiki/Modernismi
FIW-17	Ohjelmistotuotanto	19/07/2014	http://fi.wikipedia.org/wiki/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.wikipedia.org/wiki/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			
Referência	Verbete	Data de acesso	Endereço
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/Publicité
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.wikipedia.org/wiki/Réseautage_social
FRW-18	Rivière	19/07/2014	http://fr.wikipedia.org/wiki/Rivière
FRW-19	Serveur informatique	19/07/2014	http://fr.wikipedia.org/wiki/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.wikipedia.org/wiki/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			
Referência	Verbete	Data de acesso	Endereço
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	Idö	19/07/2014	http://hu.wikipedia.org/wiki/Idö
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	Müvészet	19/07/2014	http://hu.wikipedia.org/wiki/Müvészet
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 (színmü)	19/07/2014	http://hu.wikipedia.org/wiki/Opera_(színmü)
HUW-19	Ösrobbanás	19/07/2014	http://hu.wikipedia.org/wiki/Ösrobbanás
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			
Referência	Verbete	Data de acesso	Endereço
INW-01	Aeneid	19/07/2014	http://en.wikipedia.org/wiki/Aeneid
INW-02	Atacama Desert	19/07/2014	http://en.wikipedia.org/wiki/Atamaca_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			
Referência	Verbete	Data de acesso	Endereço
ITW-01	Adolescenza	21/07/2014	http://en.wikipedia.org/wiki/Adolescenza
ITW-02	Balleto	21/07/2014	http://en.wikipedia.org/wiki/Balleto
ITW-03	Biosfera	21/07/2014	http://en.wikipedia.org/wiki/Biosfera
ITW-04	Bullismo	21/07/2014	http://en.wikipedia.org/wiki/Bullismo
ITW-05	Colpa (diritto)	21/07/2014	http://en.wikipedia.org/wiki/Colpa_(diritto)
ITW-06	Conoscenza	21/07/2014	http://en.wikipedia.org/wiki/Conoscenza
ITW-07	Cultura	21/07/2014	http://en.wikipedia.org/wiki/Cultura
ITW-08	Empatia	21/07/2014	http://en.wikipedia.org/wiki/Empatia
ITW-09	Eroe	21/07/2014	http://en.wikipedia.org/wiki/Eroe
ITW-10	Fiume	21/07/2014	http://en.wikipedia.org/wiki/Fiume
ITW-11	Isaac Asimov	21/07/2014	http://en.wikipedia.org/wiki/Isaac_Asimov
ITW-12	Lettura	21/07/2014	http://en.wikipedia.org/wiki/Lettura
ITW-13	Lingua (linguistica)	21/07/2014	http://en.wikipedia.org/wiki/Lingua_(linguistica)
ITW-14	Milizia	21/07/2014	http://en.wikipedia.org/wiki/Milizia
ITW-15	Montagna	21/07/2014	http://en.wikipedia.org/wiki/Montagna
ITW-16	Morte	21/07/2014	http://en.wikipedia.org/wiki/Morte
ITW-17	Prosa	21/07/2014	http://en.wikipedia.org/wiki/Prosa
ITW-18	Rete Sociale	21/07/2014	http://en.wikipedia.org/wiki/Rete_Sociale
ITW-19	Scienza	21/07/2014	http://en.wikipedia.org/wiki/Scienza
ITW-20	Scuola	21/07/2014	http://en.wikipedia.org/wiki/Scuola
ITW-21	Senilitá	21/07/2014	http://en.wikipedia.org/wiki/Senilitá
ITW-22	Sociologia	21/07/2014	http://en.wikipedia.org/wiki/Sociologia
ITW-23	Teatro	21/07/2014	http://en.wikipedia.org/wiki/Teatro
ITW-24	Uomo	21/07/2014	http://en.wikipedia.org/wiki/Uomo
ITW-25	Vita	21/07/2014	http://en.wikipedia.org/wiki/Vita

Português			
Referência	Verbete	Data de acesso	Endereço
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	Logística	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			
Referência	Verbete	Data de acesso	Endereço
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	Jordbäavning	19/07/2014	http://sv.wikipedia.org/wiki/Jordbäavning
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

Referências Bibliográficas

[1] Yaneer Bar-Yam, *Dynamics of complex systems.* Maryland, USA: Addison-Wesley, 1997. 13

[2] J. Ladyman and J. Lambert, "What is a complex system?." `http://philsci-archive.pitt.edu/9044/4/LLWultimate.pdf`. Acessado em 15 de Janeiro de 2017. 13

[3] James Gleick, *The information: a history, a theory, a flood.* London, UK: Reaktion Books, 2011. 13

[4] `https://www.ethnologue.com/`. Acessado em 16 de Janeiro de 2017. 13

[5] Steven Roger Fischer, *A history of writing.* London, UK: Reaktion Books, 2003. 13, 16

[6] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: What is it, who has it, and how did it evolve?," *Science's Compass*, vol. 298, pp. 1569–1579, 2002. 14

[7] Thomas L. Griffiths, "Rethinking language: How probabilities shape the words we use," *PNAS*, vol. 108, pp. 3825–3826, 2011. 14

[8] Gabriel Altmann, "On the symbiosis of physicists and linguists," *Romanian Reports in Physics*, vol. 60, pp. 417–422, 2008. 14

[9] George Kingsley Zipf, *The psycho-biology of language: an introduction to dynamic philology.* Boston, USA: Houghton Mifflin, 1935. 14, 18, 21, 23, 27

[10] George Kingsley Zipf, *Human behavior and the principle of least effort.* Massachusetts, USA: Addison-Wesley, 1949. 14, 18, 20, 21, 23, 27

[11] Claude Elwood Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 623–656, 1948. 14, 37

[12] Claude Elwood Shannon, "Prediction and entropy in printed English," *The Bell System Technical Journal*, pp. 50–64, 1951. 14, 18

[13] Benoit Mandelbrot, "An informational theory of the statistical structure of language," *Communication Theory*, pp. 486–502, 1953. 14, 23, 27

[14] R. Ferrer i Cancho and R. V. Solé, "Least effort and the origins of scaling in human language," *PNAS*, pp. 788–791, 2003. 14, 23

[15] P. Carpena, P. Bernaola-Galván, and P. C. Ivanov, "New class of level statistics in correlated disordered chains," *Physical Review Letters*, vol. 93, 2004. 14

[16] J. P. Herrera and P. A. Pury, "Statistical keyword detection in literary corpora," *The Europen Physical Journal B*, vol. 63, pp. 135–146, 2008. 14, 15, 41

[17] Gustav Herdan, *Type-token mathematics*. The Hague, NL: Mouton Publishers, 1960. 14

[18] Harold S. Heaps, *Information retrieval: computational and theoretical aspects*. Academic Press, 1978. 14, 23, 24

[19] Leo Egghe, "Untangling Herdan's law and Heaps' law: mathematical and informetric arguments," *PNAS*, pp. 788–791, 2003. 14

[20] S. T. Piantadosi, H. Tily, and E. Gibson, "Word lengths are optimized for efficient communication," *PNAS*, 2010. 14

[21] M. A. Montemurro and D. H. Zanette, "Towards the quantification of the semantic information encoded in written language," *Advances in Complex Systems*, vol. 13, pp. 135–153, 2010. 14, 40, 41

[22] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," *Physical Review X*, vol. 3, 2013. 14, 15, 24, 26, 43, 61

[23] M. Cassandro, P. Collet, A. Galves, and C. Galves, "A statistical-physics approach to language acquisition and language change," *Physica A*, vol. 263, pp. 427–437, 1999. 14

[24] M. Ausloos, "Measuring complexity with multifractals in texts. Translation effects," *Chaos, Solitons and Fractals*, vol. 45, pp. 1349–1357, 2012. 14

[25] U. Strauss, P. Gryzbek, and G. Altmann, *Contributions to the science of text and language*. Springer, 2007. 14

[26] M. Ausloos, "Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series," *Physical Review E*, vol. 23, 2012. 14, 43, 44, 47, 48

[27] L. Vargas-Guzmán, B. Obregón-Quintana, D. Aguilar-Velázquez, R. Hernández-Pérez, and L. S. Liebovitch, "Word-length correlations and memory in large texts: a visibility network analysis," *Entropy*, vol. 17, pp. 7798–7810, 2015. 14, 16, 48, 50, 54

[28] M. W. Berry and J. Kogan, *Text mining: applications and theory.* West Sussex, UK: Wiley, 2010. 15

[29] Hans P. Luhn, "The automatic creation of literature abstracts," *IBM Journal*, pp. 159–165, 1958. 15, 18, 27

[30] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and M. Somoza, "Keyword detection in natural languages and DNA," *Europhysics*, vol. 57, pp. 759–764, 2002. 15, 29, 30, 31, 34, 35

[31] H. Zhou and G. W. Slater, "A metric to search for relevant words," *Physica A*, vol. 329, pp. 309–327, 2003. 15, 34

[32] 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*, vol. 79, 2009. 15, 34, 35, 66

[33] C. Carretera-Campos, P. Bernaola-Galván, A. V. Coronado, and P. Carpena, "Improving statistical keyword detection in short texts: entropic and clustering approaches," *Physica A*, 2012. 15

[34] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, "Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words," *Plos One*, vol. 4, 2009. 15, 16, 42, 68, 74, 75

[35] K. Tanaka-Ishii and A. Bunde, "Long-Range memory in literary texts: on the universal clustering of the rare words," *Plos One*, 2016. 15

[36] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, "Languages cool as they expand: Allometric scaling and the decreasing need for new words," *Scientific Reports*, vol. 2, 2012. 15

[37] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiber, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, "Quantitative analysis of Culture using millions of digitezed books," *Science*, vol. 331, pp. 176–182, 2011. 15, 26

[38] Maelyson Rolim, *Aspectos estatísticos da distribuição espacial de palavras em linguagem escrita.* Dissertação de mestrado, Departamento de Física - UFRPE, 2014. 15, 24, 25, 28, 30, 33, 34, 41, 61, 67, 71, 84

[39] S. Bird, E. Klein, and E. Loper, *Natural language processing with python.* O'Reilly, 2009. 16

[40] E. G. Altmann, G. Cristadoro, and M. D. Esposti, "On the origin of long-range correlations in texts," *PNAS*, vol. 109, pp. 11582–11587, 2012. 16

[41] K. Kosmidis, A. Kalampokis, and P. Argyrakis, "Language time series analysis," *Disponível em aRXiv*, 2006. 16, 48, 50, 54

[42] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *Society for Industrial and Applied Mathematics*, vol. 51, 2009. 18

[43] Per Bak, *How nature works: the science of self-organized criticality.* Springer Science, 1996. 19

[44] Sílvio R. A. Salinas, *Introdução à Física Estatística.* EDUSP, 2005. 19, 20

[45] Harry Eugene Stanley, *Introduction to phase transitions and critical phenomena.* Clarendon Press, 1971. 20

[46] Michael Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Mathematics*, vol. 1, 2004. 20, 23, 61

[47] P. Bak and C. Tang, "Earthquakes as a self-organized critical phenomenon," *Journal of Geophysical Research*, vol. 94, 1989. 20

[48] A. Blank and S. Solomon, "Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components)," *Physica A*, vol. 287, 2000. 20

[49] `https://www.gutenberg.org/files/4300/4300-0.txt`. Acessado em 17 de Janeiro de 2017. 21

[50] Steven T. Piantadosi, "Zipf's word frequency law in natural language: a critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, 2015. 22, 23

[51] Eugene Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," *Communications in Pure and Applied Mathematics*, vol. 13, 1960. 22

[52] Herbert A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, 1955. 23, 26, 43, 61

[53] M. P. H. Stumpf and M. A. Porter, "Critical truths about power laws," *Science*, vol. 335, 2012. 23

[54] Wetian Li, "Random texts exhibit Zipf's-law-like word frequency distribution," *Santa Fe Institute*, 1991. 23, 43

[55] B. Conrad and M. Mitzenmacher, "Power laws for monkeys typing randomly: the case of unequal probabilities," *IEEE Transactions on Information Theory*, vol. 50, 2004. 23

[56] R. Ferrer i Cancho and B. Elvevåg, "Random texts do not exhibit the real Zipf's law-like rank distribution," *Plos One*, vol. 5, 2010. 23

[57] Gustav Herdan, *The advanced theory of language as choice and chance.* Springer, 1966. 23, 24

[58] F. Font-Clos, G. Boleda, and A. Corral, "A scaling law beyond Zipf's law and its relation to Heaps' law," *New Journal of Physics*, vol. 15, 2013. 24

[59] Tommaso Pola, *Statistical analysis of written languages.* Tesi di laurea in sistemi dinamici e applicazioni, Università de Bologna, 2013. 24, 26, 43, 69, 78

[60] A. Gelbukh and G. Sidorov, "Zipf and Heaps laws' coefficients depend on language," *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, 2001. 24, 25

[61] D. van Leijenhorst and T. P. van der Weide, "A formal derivation of Heaps' law," *Information Sciences*, vol. 70, 2005. 26

[62] L. Lü, Z.-k. Zhang, and T. Zhou, "Zipf's law leads to Heaps' law: analyzing their relation in finite-size systems," *Plos One*, 2010. 26

[63] L. Lü, Z.-k. Zhang, and T. Zhou, "Deviation of Zipf's and Heaps' laws in human languages with limited dictionary size," *Scientific Reports*, 2013. 26

[64] C. K. Peng, S. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, and H. E. Stanley, "Long-range correlations in nucleotide sequences," *Nature*, vol. 356, 1992. 28

[65] A. Schenkel, J. Zhang, and Y.-C. Zhang, "Long-range correlation in human writings," *Fractals*, vol. 01, 1993. 28

[66] W. Ebeling and A. Neiman, "Long-range correlations between letters and sentences in texts," *Physica A*, vol. 215, 1995. 28, 29

[67] G. J. Stephens and W. Bialek, "Statistical mechanics of letters in words," *Physical Review E*, vol. 81, 2010. 29

[68] M. A. Montemurro and P. A. Pury, "Long-range fractal correlations in literary corpora," *Fractals*, vol. 10, 2002. 29, 48

[69] 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," *Reviews of Modern Physics*, vol. 53, 1981. 29

[70] Marek Wolf, "Nearest-neighbor-spacing distribution of prime numbers and quantum chaos," *Physical Review E*, vol. 89, 2014. 30, 63, 69

[71] M. J. Berryman, A. Allison, and D. Abbott, "Statistical techniques for text classification based on word recurrence intervals," *Fluctuation and noise letters*, vol. 03, 2003. 34

[72] D. R. Amancio, E. G. Altmann, O. N. Oliveira Jr, and L. F. Costa, "Comparing intermittency and network measurements of words and their dependence on authorship," *New Journal of Physics*, vol. 13, 2011. 34

[73] M. Hackenberg, A. Rueda, P. Carpena, P. Bernaola-Galván, G. Barturen, and J. L. Oliver, "Clustering of DNA words and biological function: a proof of principle," *Journal of Theoretical Biology*, vol. 297, 2012. 34

[74] David S. Landes, *The unbound Prometheus*. Cambridge University Press, 2003. 36

[75] José Fernando Rocha, *Origens e evolução das ideias da física.* EDUFBA, 2011. 36

[76] Rudolf J. Clausius, *The mechanical theory of heat.* Macmillan and Corporation, 1879. 36

[77] E. T. Jaynes, "Gibbs vs Boltzmann entropies," *American Journal of Physics*, vol. 33, 1965. 37

[78] Josiah W. Gibbs, *Elementary principles in statistical mechanics.* Scribner's sons, 1902. 37

[79] Pedro Rogério S. Gomes, "Modelos de urna." `http://fma.if.usp.br/~pedrorsg/ModeloUrna.pdf`. Acessado em 8 de Fevereiro de 2017. 37

[80] R. P. Feynman, R. B. Leighton, and M. Sands, *Lições de física volume 1.* Bookman, 2009. 37

[81] T. M. Cover and J. A. Thomas, *Elements of information theory.* Wiley-Interscience, 2006. 38, 39

[82] M. A. Montemurro and D. H. Zanette, "Entropic analysis of the role of words in literary texts," *Advances in Complex Systems*, vol. 05, 2002. 39, 40

[83] Marcelo A. Montemurro, "Quantifying the information in the long-range order of words: semantic structures and universal linguistic constraints," *Cortex*, vol. 55, 2014. 40, 41

[84] Damián G. Hernández, "Information approach to co-occurrence of words in written language," *Complex Systems*, vol. 24, 2015. 41

[85] A. Mehri and A. H. Darooneh, "The role of entropy in word ranking," *Physica A*, vol. 390, 2011. 41

[86] Jens Feder, *Fractals.* Springer Science, 1988. 43, 44, 47

[87] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications.* Springer Science, 2000. 43, 48

[88] Torsten Kleinow, *Testing continuous time models in financial markets.* Tese de doutorado, 2002. 44, 47

[89] B. Qian and K. Rasheed, "Hurst exponent and financial market predictability," *Conference on Financial Engineering and Applications*, 2004. 44, 47

[90] X.-Y. Qian, G.-F. Gu, and W.-X. Zhou, "Modified detrended fluctuation analysis based on empirical mode decomposition for the characterization of anti-persistent processes," *Physica A*, vol. 390, 2011. 45, 49

[91] H. E. Hurst, R. P. Black, and Y. M. Simaika, *Long-term storage, an experimental study*. London Publisher, 1965. 46, 47

[92] B. B. Mandelbrot and J. R. Wallis, "Noah, Joseph, and operational hydrology," *Water Resources Research*, vol. 4, 1968. 47

[93] A. Carbone, G. Castelli, and H. E. Stanley, "Time-dependent Hurst exponent in financial time series," *Physica A*, 2004. 47

[94] X. Liu, B. Wang, and L. Xu, "Statistical analysis of hurst exponent of essential/-nonessential genes in 33 bacterial genomas," *Plos One*, 2015. 47

[95] J. L. McCauley, K. E. Bassler, and G. H. Gunaratne, "Martingales, the efficient market hypothesis and spurious stylized facts," *Arxiv preprint*, 2008. 48

[96] J. Bhan, S. Kim, J. Kim, Y. Kwon, S.-i. Yang, and K. Lee, "Long-range correlations in Korean literary corpora," *Chaos, Solitons and Fractals*, vol. 29, 2006. 48

[97] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Physica A*, vol. 295, 2001. 48

[98] C. K. Peng, S. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotide," *Physical Review E*, vol. 49, 1994. 48, 49

[99] C. Matsoukas, S. Islam, and I. Rodriguez-Iturbe, "Detrended fluctuation analysis of rainfall and streamflow time series," *Journal of Geophysical Research*, vol. 105, 2000. 48

[100] J. Alvarez-Ramirez, J. Alvarez, and E. Rodriguez, "Short-term predictability of crude oil markets: a detrended fluctuation analysis approach," *Energy Economics*, vol. 30, 2008. 48

[101] G. Şahin, M. Erentürk, and A. Hacinliyan, "Detrended fluctuation analysis in natural languages using non-corpus parametrization," *Chaos, Solitons and Fractals*, vol. 41, 2009. 48

[102] E. A. F. Ihlen, "Introduction to multifractal detrended fluctuation analysis in Matlab," *Frontiers in Physiology*, 2012. 52

[103] B. Podobnik and H. E. Stanley, "Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series," *Physical Review Letters*, 2008. 55, 56

[104] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley, "Cross-correlations between volume change and price change," *PNAS*, vol. 106, 2009. 56

[105] L. Sandoval Junior and I. d. P. Franca, "Correlation of financial markets in times of crisis," *Physica A*, vol. 391, 2012. 56

[106] N. Argaman, F.-M. Dittes, E. Doron, J. P. Keating, A. Y. Kitaev, M. Sieber, and U. Smilansky, "Correlations in the actions of periodic orbits derived from quantum chaos," *Physical Review Letters*, vol. 71, 1993. 63

[107] E. Goles, O. Schulz, and M. Markus, "Prime number selection of cycles in a predator-prey model," *Complexity*, vol. 6, 2001. 63

[108] `https://www.gutenberg.org/`. 84

[109] `http://www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.jsp`. 84