UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS – CCE
DEPARTAMENTO DE FÍSICA – DFI
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA – PFI

GIULIANO AGOSTINHO RIDOLFI

CORRELAÇÕES DE LONGO ALCANCE EM TAMANHOS DE FRASES

Orientador: Renio dos Santos Mendes

Coorientador: Sérgio de Picoli Junior

Maringá

2016

GIULIANO AGOSTINHO RIDOLFI

CORRELAÇÕES DE LONGO ALCANCE EM TAMANHOS DE FRASES

Orientador: Renio dos Santos Mendes

Coorientador: Sérgio de Picoli Junior

Dissertação apresentada como requisito parcial para a obtenção do título de mestre em Física do programa de Pós-Graduação em Física, da Universidade Estadual de Maringá.

Maringá

2016

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

Ridolfi, Giuliano Agostinho
R547c Correlações de longo alcance em tamanhos de frases/ Giuliano Agostinho Ridolfi. -- Maringá, 2016.
62 f. : il. color, figs. , tabs.

Orientador: Prof. Dr. Renio dos Santos Mendes.
Coorientador: Prof. Dr. Sergio de Picoli Junior.

Dissertação (mestrado) – Universidade Estadual de Maringá, Centro de Ciências Exatas, Programa de Pós-Graduação em Física, 2016.

1. Sentenças em textos. 2. Análise de corrrelações. 3. Sistemas complexos. 4. Correlação. 5. Autocorrelação. 6. DFA. 7. Hurst. I. Mendes, Renio dos Santos, orient. II. Picoli Junior, Sergio de, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Programa de Pós-Graduação em Física. IV. Título.

CDD 22. ED.530.13
JLM-001648

Agradecimentos

Em primeiro lugar, é necessário dizer que agradecimentos têm necessariamente de ser feitos com uma determinada ordem, a ordem das pessoas a quem são dedicados, de maneira a parecer favorecer alguém em detrimento de outro. No entanto, os envolvidos aqui citados têm todos uma grande porção de importância e, sem essas pessoas, esse trabalho jamais poderia ser concluído.

Dito isso, agradeço primeiramente ao Prof. Renio, que esteve ao meu lado do começo ao fim, empregando seu conhecimento e disposição para me ajudar a produzir um trabalho de qualidade. Posso dizer que esse período todo utilizado para elaborar esta dissertação foi uma escola, não só com respeito às técnicas matemáticas e fluidez de ideias em trabalhos acadêmicos, mas também de empatia na relação professor-aluno, algo extremamente necessário a qualquer um que queira estar, um dia, na posição de mestre, orientando e ensinando.

Em igual grau, devo agradecer aos meus pais, cujo suporte emocional foi vital para este projeto. Aliás, seria muito injusto se eu usasse este espaço para agradecê-los por este período, somente. A toda a formação que tenho hoje, a tudo o que me torna capaz de produzir coisas boas, devo aos meus pais, Paolo e Cris.

Agradeço também à minha irmã Laura, à Suely, bem como ao restante da minha família, e também a todos os meus amigos. Meus amigos são todos valorosos e com características únicas. Sem dúvidas, a eles lhes sou muito grato. Aos colegas de trabalho também devo inúmeros "obrigados", por terem me dado do seu conhecimento para a realização deste estudo. Em especial ao Haroldo e ao Prof. Sérgio.

Por fim, saliento que as pessoas acima tiveram todas sua parcela de importância para a elaboração deste projeto, mas devo à CAPES e ao CNPq meus agradecimentos pelo suporte financeiro que me foi dado neste período de dois anos empenhados no meu crescimento profissional e acadêmico.

Sumário

Lista de Figuras

Lista de Tabelas

"Todos os pensamentos verdadeiramente grandes são concebidos durante a caminhada" (Friedrich Nietzsche)

Resumo

Correlações entre entidades matemáticas têm sido estudadas há bastante tempo em sistemas complexos. O objeto de estudo deste trabalho, o qual se conjectura apresentar algum grau de autocorrelação, é composto pela base de dados relativa aos tamanhos de frases em textos, quantificados em termos do número de palavras em cada uma delas. Estes textos provêm do segundo testamento da bíblia em português e em outras dezenove línguas. Não é apenas diretamente dos tamanhos de frases que as correlações são investigadas, mas também de outras duas séries temporais extraídas da original. Nesse estudo, verifica-se que o expoente de Hurst indica persistência para todos os idiomas nas séries temporais formadas pelos tamanhos das frases e, também, para aquelas formadas pelos módulos das diferenças consecutivas nos tamanhos dessas frases, atestando, possivelmente, a presença de correlações de longo alcance. Para séries temporais formadas a partir dos sinais dessas mesmas diferenças, os expoentes encontrados indicam a antipersistência ou, talvez, ausência de correlações. Quantitativamente, os expoentes de Hurst das séries dos tamanhos são, em geral, próximos, indicando que eles são aproximadamente independentes do idioma considerado. As mesmas conclusões quantitativas foram observadas para a série dos módulos e a dos sinais. A técnica aqui empregada para a extração deste expoente característico é a DFA − análise de flutuação destendenciada.

Palavras-chave: Textos, frases, sentenças, correlação, autocorrelação, DFA, Hurst.

Abstract

Correlations between mathematical entities have been studied for very long in the complex systems area. The study object of this analysis, assumed to display some degree of self-correlation, is composed of a database concerning sentences lengths in texts, quantified in terms of the number of words within them. These texts are extracted from the second testament of the bible, in Portuguese and other nineteen languages. Correlations are investigated not solely directly from sentences lengths, but also from two additional time series, extracted from the original one. In this study, one verifies that the Hurst exponent points to persistence for length time series in all considered languages, and for those built on the absolute values of sentence lengths differences as well, possibly pointing towards a long-range correlation presence. For time series built on the signs of these same differences, the found exponents indicate anti-persistence or, perhaps, absence of correlations. Quantitatively, the Hurst exponents of length series are generally close to each other, indicating that they are approximately independent on the language in consideration. The same quantitative conclusions were observed for the modules and signs series. The technique that is utilized here for the extraction of such an exponent is the DFA − detrended fluctuation analysis.

Keywords: Texts, phrases, sentences, correlation, self-correlation, DFA, Hurst.

Introdução

Já se passaram 150 anos desde que Ludwig Boltzmann publicou seu primeiro trabalho sobre mecânica estatística [1, 2]. Hoje, os estudos sobre a dinâmica dos gases e sistemas interagentes em geral, sob os conceitos de entropia e de probabilidade, já têm um alto grau de consolidação, fornecendo uma grande concordância com os dados experimentais. Tal sucesso permite o seguinte questionamento: é possível tomar como parâmetro os procedimentos da mecânica estatística para o desenvolvimento de análises diversas, que não tenham como objeto um gás de partículas (ou de um sistema mecânico-estatístico usual)? Essa pergunta já tem sido respondida afirmativamente por muitos cientistas que se dedicam ao estudo de *sistemas complexos*.

No contexto de sistemas complexos, analisam-se sistemas que são e, também, que não são usualmente alvo das disciplinas da física. Esse caráter abrangente dá margem à possibilidade de análise dos mais diversos sistemas possíveis, mas todos têm características que os torna passíveis de abordagem semelhante. Uma delas é que tenham entidades que se apresentem em grande quantidade, podendo compor uma vasta base de dados [3, 4]. As ferramentas computacionais das últimas décadas têm permitido que esses dados sejam manipulados em uma velocidade muito grande. Ainda assim, é importante dizer que a análise não é, quase nunca, determinista, mas, sim, essencialmente de caráter probabilístico. Os sistemas complexos podem servir à ecologia, por exemplo, como se vê no tratamento dado à dinâmica de populações de espécies diversas. Se, dentro disso, consideram-se fluxos migratórios, também é possível descrever interações sociais humanas, cujos resultados podem eventualmente servir para reconstruir a história, ratificando ou refutando especulações já estabelecidas [5]. À genética isso tem se mostrado de grande interesse, uma vez que hoje já se está reconstruindo a identidade genética de diversas po-

pulações. Sistemas sociais também têm sido amplamente investigados segundo o enfoque típico da mecânica estatística. Como uma área específica dessa conexão do uso de ferramentas típicas de física estatística, pode-se citar a linguística [7]. As linhas a seguir serão dedicadas a explicar brevemente a integração da *linguística* com o enfoque de sistemas complexos.

E, justamente dessa fusão, nasce o ramo da *linguística quantitativa* [6]. A linguística, por si só, investiga elementos da linguagem como a fonologia e fonética, referentes à formação dos sons sob uma perspectiva cognitiva ou física, e, também, a morfologia e a sintaxe, respectivamente relativas à formação de palavras e frases. A semântica, por outro lado, é um estudo dos significados, e ainda haveria como pontuar outras áreas. O que cabe dizer aqui é que, por meio de técnicas computacionais viabilizadas nos últimos anos, as análises fonéticas, morfológicas, dentre outras, passaram também a ser estudadas a partir de uma outra perspectiva. Talvez coubesse outrora à sensibilidade de um grupo limitado de linguistas trabalhando exaustivamente a tarefa de, por exemplo, comparar inúmeros fonemas em suas inúmeras ocorrências ao longo de textos diversos a fim de se detectarem padrões linguísticos e promoverem boas conclusões. O que as análises estatísticas podem fazer hoje é otimizar a leitura de elementos linguísticos por meio de alguns algoritmos que imitem a forma pela qual os linguistas os interpretam. Trazendo novamente à tona as comparações feitas anteriormente, deve-se tomar como exemplo o alívio trazido pela termodinâmica e pela mecânica estatística às análises de comportamento de gases ideais, geralmente difíceis sob o uso puro de uma física determinista newtoniana.

Dentro deste campo da linguística quantitativa, é possível identificar, no mínimo, três leis, cuja validade tem sido bastante testada como o uso de *corpora* (plural de *corpus*) diversos em relação ao idioma, ao gênero textual, à data de publicação, entre outros. Elas são as leis de Menzerath, de Heaps e de Zipf [8].

A primeira lei estabelece que, quanto maior (menor) é o tamanho médio dos componentes de um dado objeto, o tamanho deste objeto tende a diminuir (aumentar). Essa inversão de proporções pôde ser aplicada às análises morfológicas, uma vez estabelecidas sílabas como componentes de um objeto maior, neste caso, a palavra. Além disso, a validade desta lei pareceu abranger também os estudos do genoma de algumas espécies, nas quais se verificou que o número em que se dispunham os cromossomos e o seu tamanho

médio se apresentavam em proporções inversas [9–11].

Já a lei de Heaps, por outro lado, relaciona o número de palavras totais e diversas em um texto por meio de um expoente característico, encontrado em geral para o limite de palavras totais (ou, alternativamente, tamanho do texto em questão) tendendo a infinito [12–14].

Por fim, a lei de Zipf investiga a relação entre a frequência e o *ranking*[1] de um determinado elemento pertencente a um conjunto de vários outros elementos diferentes. As perguntas que suscitam do uso dessa lei podem ser acerca da relação matemática entre as variáveis ou sobre quais são os objetos a serem analisados por ela. De algumas observações na lei de Zipf aplicada à frequência e ao *ranking* de palavras surgiram algumas dúvidas quanto à validade da lei de potência na relação entre as variáveis [15]. Por exemplo, vê-se a lei de Zipf aplicada à linguística: sobre palavras em língua inglesa [16] [17], mas também em chinês [18]. Há ainda outros trabalhos relacionados ao tema, como os que estendem a validade da lei de Zipf a textos randomicamente gerados [19].

As situações mencionadas anteriormente ilustram assuntos bastante explorados em linguística quantitativa. Conviria, então, empreender um esforço muito grande para promover mudanças adicionais, ainda que pequenas, no repertório dessas recorrências já relativamente bem estabelecidas entre a comunidade acadêmica. Quando isso acontece, detalhes nas expressões que descrevem o funcionamento desses objetos são ajustados, mas a lei, em si, já garante boa previsibilidade. Dessa forma, outros pesquisadores decidem se aventurar em áreas ainda não tão bem exploradas, muitas vezes por meio de inferências inéditas, ora malsucedidas, ora bem-sucedidas. Pode-se, então, apresentar como exemplos dessas outras pesquisas dentro da linguística quantitativa aquela que investiga o decaimento da distância euclidiana (correlação) entre palavras com o tamanho da sentença [20]; outra, que compara a genealogia dos idiomas com a análise taxonômica em espécies de seres vivos [21]; a que usa a entropia como uma medida de predictabilidade de letras em palavras [22]; uma análise alternativa sobre os *corpora* textuais do Google para a melhora da qualidade de conclusões estatísticas [23]; a proposição de a dependência da ocorrência

[1] Consideram-se vários elementos diferentes e um número N de eventos. Em cada um destes eventos, há a ocorrência de um destes elementos, que podem se apresentar n_i ($n_i = 0, 1, 2, 3...$) vezes ao longo do processo todo. O *ranking* é determinado da seguinte forma: ao elemento de maior frequência é atribuído o número um, ao segundo mais frequente, o número dois, e assim por diante.

de palavras recair sobre pares de outras palavras [24]; a correlação entre fluxos migratórios e a regularização de verbos em inglês [25]; ou mesmo a aplicação de caminhadas aleatórias sobre *rankings* de palavras evoluindo no tempo [26].

Deste modo, o trabalho que doravante se desenvolve é, também, uma sucessão de análises exploradas dentro do que abrange a área de sistemas complexos. Será trazida à discussão a presença de correlações nos tamanhos de sentenças ao longo de textos. É possível pontuar alguns trabalhos prévios sobre o estudo de sentenças enquanto objeto matemático sujeito a leis estatísticas.

Um desses estudos [27], utilizava uma base de dados composta de três longos textos em inglês embaralhados de diversas formas, e, em um segundo passo, suas correlações eram calculadas por meio de um expoente característico, o expoente de Hurst. Também outros indicadores eram testados, como cumulantes e espectros de potência. Os embaralhamentos poderiam ocorrer em níveis acima ou abaixo do nível de sentenças. Tendo sido alterações significativas encontradas somente para os casos em que se embaralhavam sentenças ou grupos delas, concluiu-se que as correlações deveriam ser características da relação entre elas, e não a partir de dentro delas.

Já em um outro estudo divulgado em 2012 [28], o princípio de se analisar sentenças à luz de correlações se manteve, porém os métodos empregados foram ligeiramente diferentes. Neste segundo caso, utilizou-se a MDFA (sigla em inglês para análise de flutuação destendenciada fractal), enquanto que o primeiro viabilizou as análises de correlação por meio de uma simples análise de flutuação. É importante ressaltar que a complexidade da disposição de sentenças já foi testada outras vezes [29]. Alguns estudos, especificamente para a língua japonesa, [30] contrariaram a hipótese de um simples padrão multiplicativo, em detrimento de um complexo e hierárquico.

Dentro, ainda, do contexto da linguística quantitativa, cabe citar outras realizações no campo da exploração de textos por meio da matemática [31]. Algumas delas [32, 33] consistiram em descobrir a presença de correlação de longo alcance entre palavras para além daquela de curto alcance restrita às interações dentro de frases. Outro [34], ainda, com a seleção dos tópicos de maior relevância dentro de textos, propõe que a dinâmica

das correlações se apresente de maneira "explosiva"[2]. É possível, por outro lado, também, analisar a ordem de símbolos gráficos por meio das medidas de entropia [35]. Trazendo à tona o expoente de Hurst, é possível pontuar estudos que utilizaram do expoente uma função derivada, notadamente parabólica para determinados textos literários (ainda que tenham sido feitos estudos estatísticos também sobre textos falados [36]) embaralhados [37], como uma medida, por exemplo, de complexidade [38].

O presente trabalho, além de analisar o grau de correlação das séries temporais de tamanho de sentenças, utiliza também outras duas séries derivadas (como já feito em um trabalho anterior [39], sobre batidas do coração) a partir daquela, a fim de se investigar outros padrões específicos em textos. Uma destas duas séries é tomada a partir dos sinais das diferenças dos tamanhos consecutivos de sentenças na série original, a *série dos sinais*. Um indicativo de correlação negativa, para esta série, em particular, apontaria simplesmente para um comportamento intermitente da taxa de variação dos tamanhos. Neste caso, incrementos positivos têm maior probabilidade de serem seguidos por subtrações. Nesse contexto, a *série dos módulos* das diferenças dos tamanhos também é investigada.

Além da variação no grau de correlação com respeito ao tipo de série, investiga-se também a dependência do expoente característico com a língua. Para isso, usam-se vinte idiomas em que o mesmo texto − o novo testamento da bíblia − foi escrito. Em resumo e pontuando-se uma vez mais as diretrizes deste estudo, serão vistas análises de correlação nos textos bíblicos sob parâmetros linguísticos (diversidade de idiomas), por meio de diversas séries temporais relacionadas ao tamanho de frases, à luz do expoente de Hurst, o indicador escolhido para se atestarem tais correlações.

O trabalho que aqui se apresenta está disposto em quatro capítulos. No primeiro deles, mostram-se os dados a serem investigados, assim como uma breve apresentação do método aqui empregado para investigar correlações, a DFA (sigla em inglês para análise de flutuação destendenciada). No capítulo seguinte, dispõem-se gráficos, tabelas e resultados para o expoentes de Hurst referentes a três séries temporais extraídas do novo testamento da bíblia em português. O terceiro capítulo é uma extensão das análises do capítulo anterior a outras dezenove línguas, com diferentes graus de proximidade entre elas, de

[2]Tradução do termo que, em inglês, é comumente utilizado para designar séries temporais que apresentem caráter "explosivo", *bursty*.

acordo com a sua genealogia.

Nos dois últimos capítulos antecedentes à conclusão, mostram-se também resultados sobre correlações nos textos por meio desse expoente, e o penúltimo, por dispor de vinte idiomas em análise, confronta-os em dois aspectos diferentes, referentes à relação da média (do tamanho de frases) com o desvio padrão e o número de frases no texto. Os principais resultados obtidos, sobretudo a partir das análises apresentadas nos capítulos 2 e 3, estão dispostas na conclusão.

Capítulo 1

Dados e métodos

Neste capítulo, serão apresentados os dados analisados nos capítulos subsequentes, bem como a técnica empregada para a análise desses dados. Assim será feito, porque os próximos capítulos tratarão de análises de correlação entre tamanhos de frases dentro da bíblia em português, primeiramente, e, depois, para vários idiomas. Quanto à técnica, será empregada a DFA na investigação dos dados para que conclusões acerca de correlações sejam obtidas por meio da interpretação dos chamados expoentes de Hurst.

1.1 Dados

A análise que será apresentada no presente estudo está baseada nas frases que compõem um texto. Frases são compreendidas por trechos dentro de um texto delimitados por ponto final, de exclamação ou interrogação e os seus tamanhos são computados em termos do número de palavras. É oportuno ressaltar que a presença de vírgula ou outros sinais de pontuação, que não sejam os mencionados anteriormente, não interfere no tamanho da frase. Assim, por exemplo, a frase imediatamente anterior tem tamanho igual a 25.

O texto escolhido para ser investigado aqui é a bíblia. Sua escolha para as análises de correlação se justifica não só pela grande extensão do corpo textual, mas também porque é dividida em livros escritos por autores diversos e em diferentes épocas, em que se cogita a presença de alguma heterogeneidade (em relação à frequência em que

determinadas palavras são utilizadas ou ao tamanho em que são dispostas as sentenças). Outra motivação para a escolha da bíblia é o fato de que há traduções disponíveis em mais de 4.000 línguas [40], permitindo-se que a análise de correlação seja efetuada ao se considerarem os cálculos sobre textos em diferentes idiomas, com diferentes estruturas sintáticas e morfológicas.

Como em geral é sabido, a bíblia se divide em dois testamentos (o novo e o velho), cada um composto de vários livros, como já dito, escritos via de regra por autores diferentes. Como ilustração, tem-se na tabela 1.1 o número total M de frases de cada livro do novo testamento da bíblia em português (Almeida revista e corrigida) [41], bem como a média (ou valor médio) e o desvio padrão dos seus tamanhos.

Bíblia em português			
Livro	M	μ	σ
Mateus	1.088	18,95	12,38
Marcos	715	18,08	11,61
Lucas	1.136	19,56	13,05
João	1.011	16,61	9,21
Atos dos Apóstolos	979	21,56	13,24
Romanos	451	19,34	14,95
I Coríntios	484	17,51	11,23
II Coríntios	229	23,84	17,15
Gálatas	141	19,96	15,07
Efésios	74	37,76	33,93
Filipenses	82	25,00	16,03
Colossenses	57	33,01	29,98
I Tessalonicenses	66	27,14	20,89
II Tessalonicenses	29	33,90	39,35
I Timóteo	89	24,19	19,11
II Timóteo	71	21,68	19,94
Tito	33	27,03	24,66
Filemom	18	23,39	17,13
Hebreus	241	25,84	17,20
Tiago	125	17,37	10,38
I Pedro	67	33,64	30,30
II Pedro	37	36,54	25,43
I João	124	18,49	11,46
II João	16	17,31	11,29
III João	19	14,37	11,85
Judas	18	31,94	20,69
Apocalipse	438	25,39	16,10

Tabela 1.1: **Dados referentes à bíblia em português**. M é o número de sentenças para cada livro do novo testamento da bíblia em questão, e μ e σ são, respectivamente, a média e o desvio padrão em relação ao número de palavras por sentença. Os livros estão ordenados segundo a sequência em que se dispõem ao longo do segundo testamento da bíblia.

O conjunto dos tamanhos das sentenças na ordem em que aparece no texto é a série temporal básica (*série original*) de análise desta dissertação. A partir desta série básica, três outras séries temporais foram desenvolvidas. A primeira é obtida por meio dos tamanhos originais das sentenças subtraídos do tamanho relativo ao instante de tempo anterior, e é chamada *série das diferenças*. A segunda e a terceira séries são obtidas a partir da primeira: respectivamente, tomam-se os módulos das diferenças e se chega à *série dos módulos* e os sinais das diferenças, encontrando-se a *série dos sinais*. A tabela 1.2 dá a definição (assim como a nomenclatura a ser empregada ao longo deste texto) do i-ésimo termo de cada uma das séries mencionadas. As figuras 1.1 e 1.2 dispõem essas séries em gráficos, fornecendo uma visualização global e parcial delas, respectivamente.

Série	**Definição**	**Expressão**
Original	N_i	*Número de palavras/frase*
Diferenças	W_i	$N_i - N_{i-1}$
Módulos	Z_i	$\|W_i\|$
Sinais	S_i	$W_i/\|W_i\|$, se $W_i \neq 0$; ou 0, se $W_i = 0$

Tabela 1.2: **Definição e nomenclatura das séries temporais**. A nomenclatura e a definição dos objetos básicos sob discussão neste trabalho, as séries dos tamanhos das frases (série original), da diferença dos tamanhos (diferenças), das magnitudes das diferenças (série dos módulos) e a dos sinais das diferenças (série dos sinais), estão expostas nesta tabela.

A versão utilizada para a primeira parte da análise foi a Almeida revista e corrigida (2009) [41]. Para a análise da bíblia em outras línguas, foram consideradas as versões em dinamarquês, norueguês, alemão, holandês, islandês, inglês, sueco, albanês, húngaro, croata, ucraniano, sérvio, russo, búlgaro, francês, crioulo haitiano, italiano, espanhol e romeno. Dado que é conveniente dispor de uma ampla base de dados, a utlização da bíblia em vários idiomas como fonte de dados se mostrou útil: como já mencionado, trata-se de um texto relativamente longo, escrito por vários autores e, ao analisá-lo sob diferentes línguas, espera-se que o estudo seja o menos enviesado possível, além de fornecer uma investigação da relevância do idioma na análise. A tabela 1.3 apresenta a quantidade de frases, bem como a média e o desvio padrão dos seus tamanhos, para bíblias em diversos idiomas.

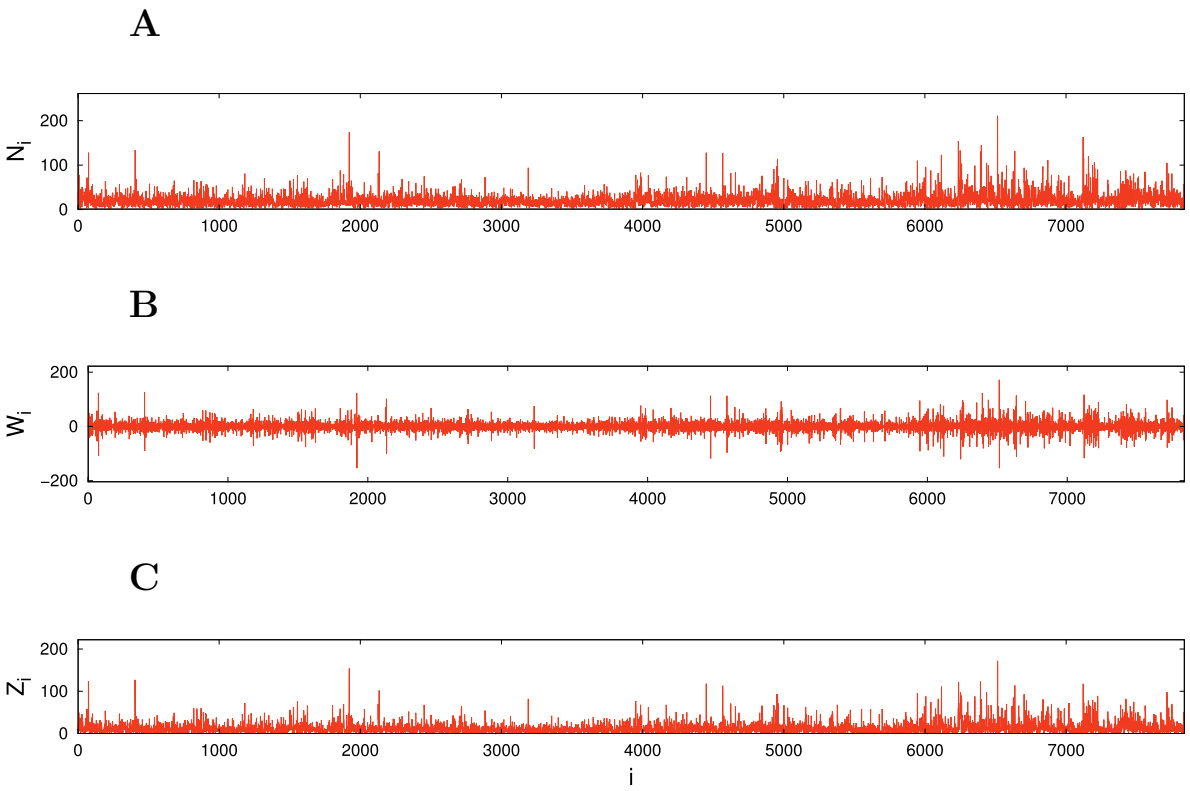

Figura 1.1: **As séries original, das diferenças e dos módulos**. As séries original N_i (A), das diferenças W_i (B) e dos módulos Z_i (C) dispostas em função do número i da sentença, considerando o novo testamento da bíblia em português (Almeida revista e corrigida). Nestes gráficos, os pontos foram unidos por segmentos de retas. A extensão da série original, ou o seu número de sentenças, é igual a 7.838.

Já foi dito que a diversidade oferecida pelos textos da bíblia, tanto em relação àquela de períodos históricos quanto à multiplicidade de autores a eles atribuídas, motivou o seu uso para as análises de correlação. Mostrou-se razoável, ainda, investigar possíveis dependências linguísticas na análise de correlação. Assim, incluiu-se na base de dados, também, textos escritos em outros idiomas, além do português. Dessa forma, os vinte idiomas em questão foram escolhidos de maneira que se abrangesse um vasto grupo de estruturas linguísticas diferentes. Se, porventura, a presença de correlações apresenta dependência sobre uma determinada estrutura linguística particular de um idioma, é esperado que essa não-uniformidade se evidencie nos resultados.

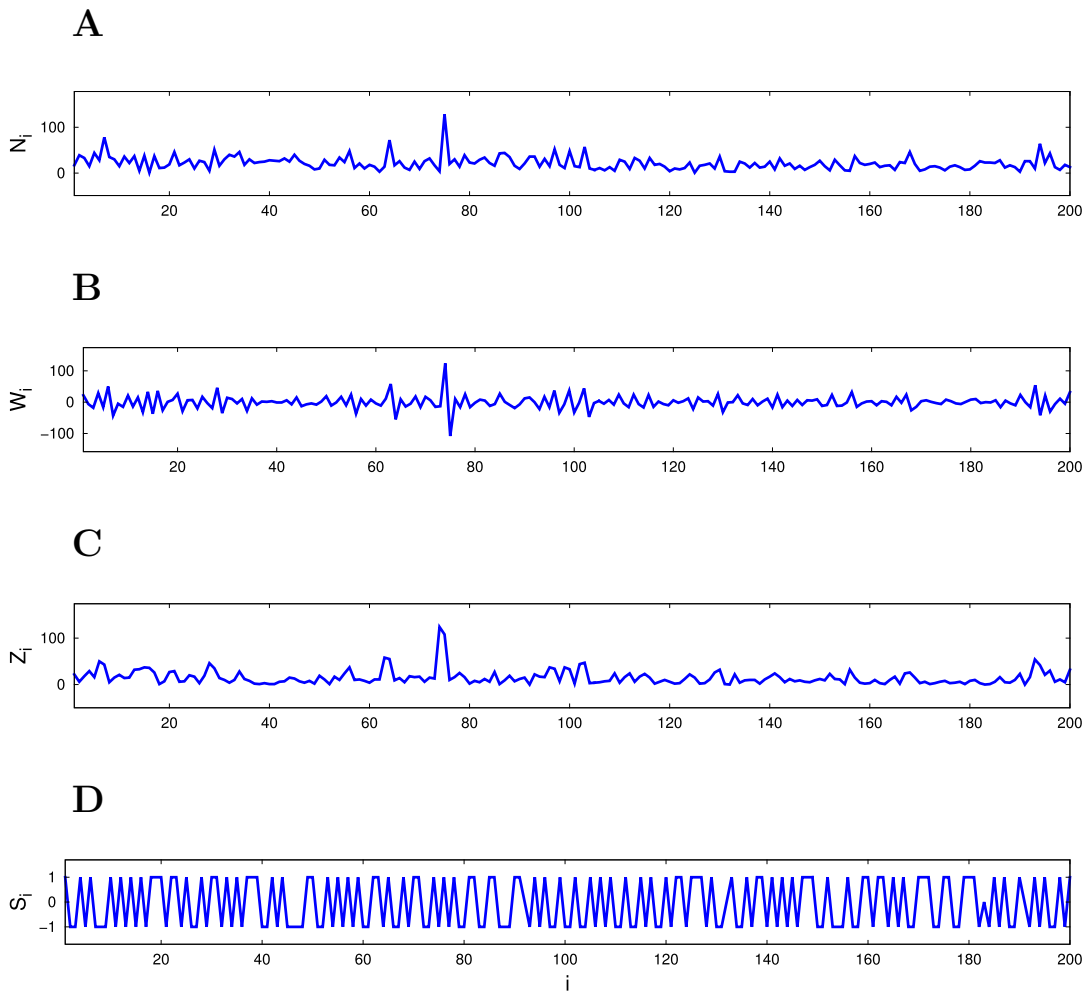

Figura 1.2: **Detalhamento das séries original, das diferenças, dos módulos e dos sinais**. As séries original N_i (A), das diferenças W_i (B), dos módulos Z_i (C) e dos sinais S_i (D) dispostas em função do número i da sentença, para o novo testamento da bíblia em português (Almeida revista e corrigida). Estão dispostas no intervalo $1 \leq i \leq 2 \times 10^2$, fornecendo um detalhamento dos dados apresentados na figura 1.1. Nesses gráficos, os dados foram unidos por segmentos de retas.

No entanto, alguns limites foram impostos para a análise: utilizaram-se somente idiomas essencialmente de alguns povos europeus e, também, apenas o novo testamento da bíblia. A justificativa para o primeiro caso é a de que já se encontra relativa diferença estrutural entre os idiomas abordados (capaz de motivar uma busca por diferenças nos resultados dependentes da língua), além de serem de conhecimento mais abrangente pelo mundo. Já reduzir a análise àquela segunda parte da bíblia é justificado pelo fato de ela apresentar maior uniformidade (em termos do número de livros disponíveis) em relação ao velho testamento, para essas versões.

Grupo	**Bíblia (idioma)**	M	μ	σ
Germânicas	Bibelen pa hverdagsdansk (dinamarquês)	12.433	16,08	8,88
	Det Norsk Bibelselskap 1930 (norueguês)	7.472	22,80	17,03
	Het Boek (holandês)	14.770	13,14	7,94
	Hoffnung fur Alle (alemão)	13.925	13,80	7,11
	Icelandic Bible (islandês)	10.129	15,70	8,81
	21st Century King James Version (inglês)	8.781	20,87	14,95
	Nya Levande Bibeln (sueco)	12.435	15,92	8,63
NI e albanês	Albanian Bible (albanês)	8.358	21,10	15,71
	Hungarian Karoli (húngaro)	7.846	18,15	12,43
Eslavas	Hrvatski Novi Zavjet Rijeka 2001 (croata)	9.136	15,48	10,32
	Ukrainian Bible (ucraniano)	9.556	14,39	10,53
	Serbian New Testament Easy-to-Read Version (sérvio)	8.289	15,39	9,49
	Russian New Testament Easy-to-Read Version (russo)	10.625	14,73	7,94
	1940 Bulgarian Bible (búlgaro)	7.666	19,79	14,06
Latinas	Almeida Revista e Corrigida 2009 (português)	7.838	20.44	14,78
	Haitian Creole Version (crioulo haitiano)	14.881	15,25	8,53
	La Bibbia della Gioia (italiano)	11.282	16,45	10,77
	La Bible du Semeur (francês)	10.981	17,60	10,44
	La Biblia de las Americas (espanhol)	7.696	22,99	15,91
	Noua Traducere In Limba Romana (romeno)	9.505	18,11	12,38

Tabela 1.3: **Dados referentes às bíblias em vários idiomas**. M é o número de sentenças para o novo testamento de uma dada bíblia, e μ e σ são, respectivamente, a média e o desvio padrão em relação ao número de palavras por sentença. A sigla NI significa não indo-europeu.

Que todas as línguas aqui em análise apresentam alguma distância estrutural uma em relação a outra, isso já é sabido. Mas é necessário explicar o seu agrupamento segundo similaridades. A figura 1.3 esquematiza a similaridade de línguas segundo o seu parentesco. De fato, como em uma árvore genealógica, agrupam-se segundo famílias ou sub-famílias. A princípio, dois grandes grupos foram tomados: o das línguas indo-europeias e aquelas não indo-europeiras.

A grande parte das línguas europeias derivam de um idioma comum, de uso extinto há mais ou menos 5.000 anos [42], o proto-indo-europeu. As línguas que se desenvolveram a partir deste idioma comum são chamadas de línguas indo-europeias, sendo as demais classificadas como não indo-europeias. As primeiras são classificadas por várias famílias, dentre as quais três delas foram tomadas aqui como exemplo: germânica, eslava e latina. Ainda que haja mais classificações dentro de uma mesma família, foram apenas mostrados os seus representantes, sem distinções. Para exemplificar isso, tomam-se como exemplo o português e o espanhol. Ambas são línguas latinas, mas, além disso, línguas ibéricas. O italiano e o francês, por outro lado, por não serem idiomas ibéricos, apresentam uma

distância maior em relação ao português do que apresenta o espanhol, deste modo. Porém, como já dito, aqui se colocam o português, o espanhol, o italiano e o francês dentro do mesmo grupo, sem que sejam evidenciadas as distâncias que apresentam um com o outro.

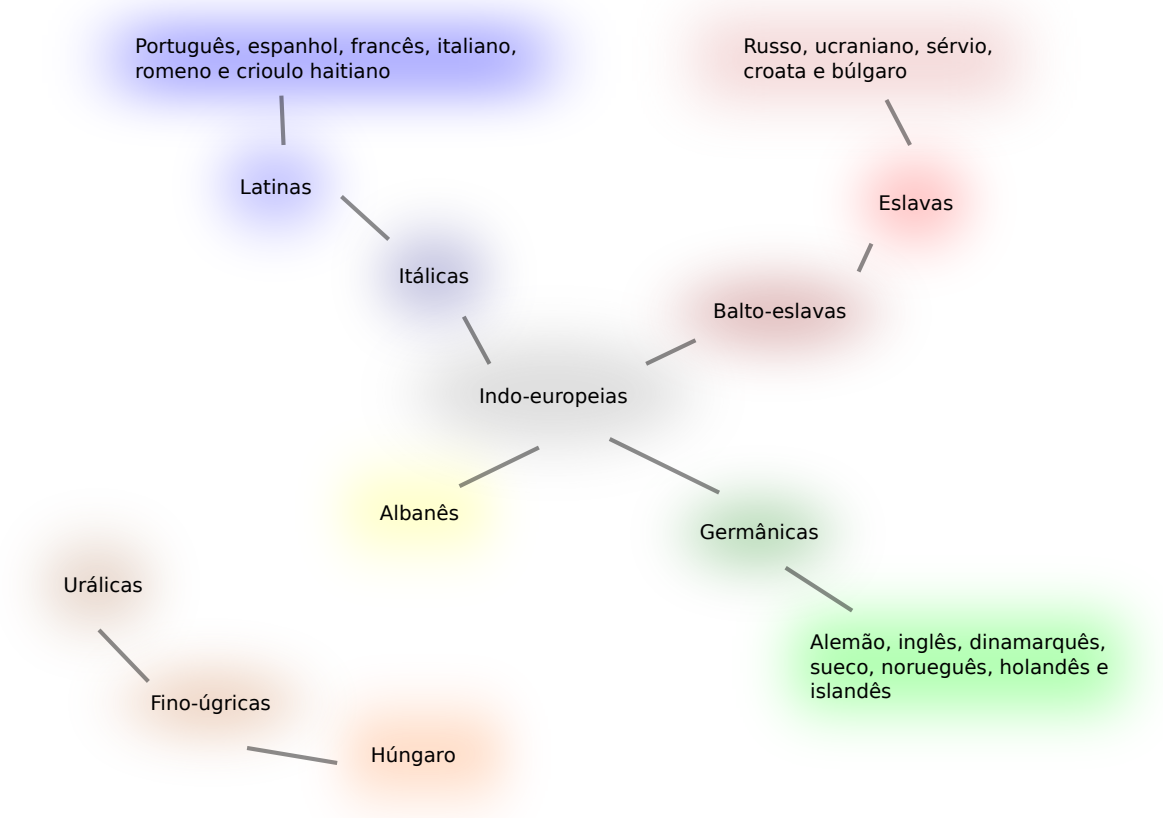

Figura 1.3: **Genealogia dos idiomas**. Nesta figura, dispõem-se todos os idiomas considerados neste trabalho e suas conexões genealógicas. Eventualmente, dentro dos balões mais à superfície, aqueles que contêm as línguas em si, omite-se relações genealógicas que possam haver de uma língua com outra. Por exemplo, dentro do conjunto de línguas latinas, há as ibéricas, das quais só o português e o espanhol, nesta amostragem, fazem parte; informações como essas são omitidas, no entanto. Vê-se que o húngaro é um caso distinto, não pertencendo ao grupo das famílias indo-europeias.

1.2 Extração dos dados

Para que se dispusse, finalmente, da base de dados correspondente aos tamanhos de sentenças, foi necessário inicialmente acessar os textos bíblicos já disponíveis na internet [40], copiá-los e fazer sobre eles o uso de alguns algoritmos, que os transformasse nas séries temporais dos tamanhos. A figura 1.4 dispõe algumas linhas de código utilizadas

para a extração da bíblia em diversos idiomas e sua conversão em arquivos de texto. A linguagem de programação empregada desde a extração até a conversão em arquivos de extensão ".txt" foi o Python [43].

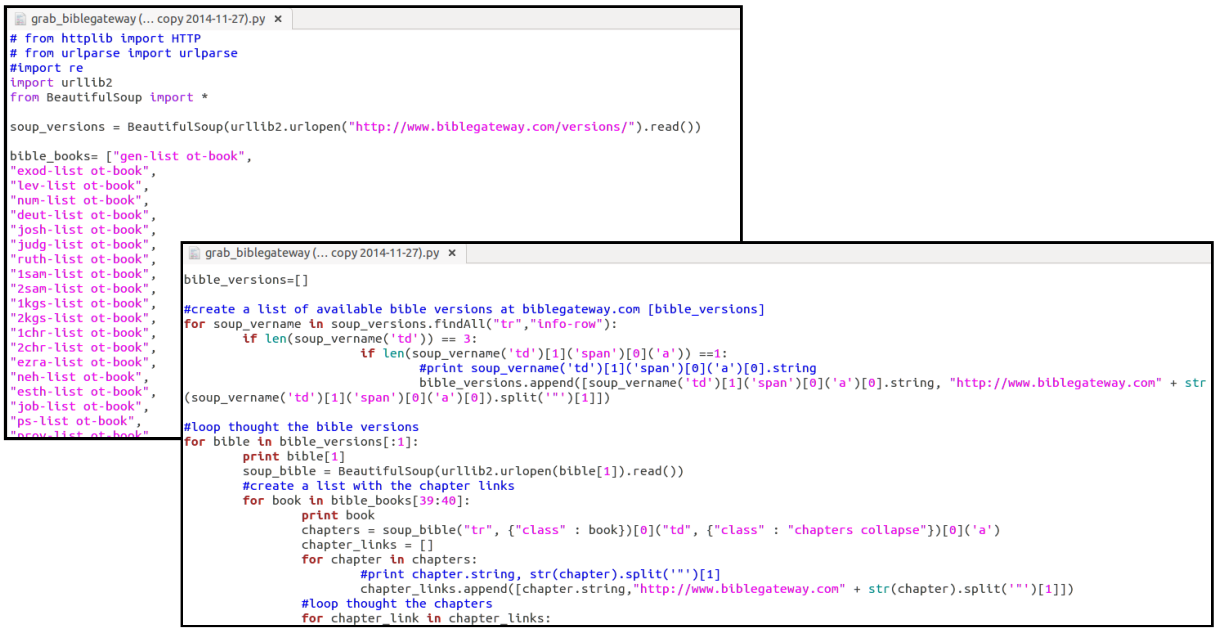

Figura 1.4: **Algumas linhas de código para a extração de dados**. Como se vê no quadro mais atrás, o algoritmo interpreta o código em HTML da página da qual se extraem os dados, mudando o endereço conforme os livros da bíblia (cada uma das linhas corresponde a um livro diferente). Mais à frente, há um trecho da continuação do programa, que está escrito em linguagem Python.

Dispondo-se dos arquivos de texto separados por livros, o passo seguinte consistiu na contagem das frases e de seus tamanhos. Esses passos foram viabilizados pelo programa Mathematica [44], que oferece uma interface amigável a análises de texto. Ao longo da dissertação, como dito no início do capítulo, as séries temporais calculadas são essencialmente baseadas na definição de frase com base nos trechos de palavras delimitados por ".", "?" e "!" (ponto final, de interrogação e de exclamação). Outras definições são postas à prova, com a inclusão de outros sinais gráficos, a saber, ":" e ";" (dois pontos e ponto-e-vírgula). A tabela 1.4 mostra, para o novo testamento da bíblia em português, quantas sentenças são formadas a partir de cada uma das definições de pontuação, juntamente às médias e desvios padrão em relação ao tamanho delas. Dessa tabela, cabe notar que os desvios padrão variam aproximadamente de 72% a 87% em relação a suas médias.

	.!?	.!?:	.!?;	.!?:;
M	7.838	10.134	10.057	12.358
μ	20,44	15,83	15,95	12,98
σ	14,78	13,84	12,03	10,54

Tabela 1.4: **Número de sentenças para diferentes definições**. Quatro diferentes definições de pontuação geram um número de sentenças M diferente para o mesmo texto, neste caso, o novo testamento da bíblia em português (Almeida revista e corrigida). Também são dispostas as médias (μ) e os desvios padrão (σ) referentes ao tamanho dessas frases.

1.3 Análise de autocorrelações

Visando motivar o método básico de análise nesta dissertação, a DFA[1] (análise de flutuações destendenciadas), esta parte do presente capítulo se inicia com uma breve discussão sobre caminhadas aleatórias, focando-se no seu desvio padrão [45]. Como será visto, o objeto central da DFA é a função de flutuação, que está diretamente relacionada com o cálculo do desvio padrão de uma caminhada aleatória. A seguir, apresenta-se uma conexão entre caminhadas aleatórias e análise de flutuações (FA[2]). Por fim, esta apresentação culmina com a exposição da DFA.

1.3.1 Desvio padrão em caminhadas aleatórias

Considera-se, a princípio, uma caminhada aleatória de n passos. A posição final $X^{(n)}$ é, portanto, a soma dos deslocamentos discretos x_i até $i = n$:

$$X^{(n)} = \sum_{i=1}^{n} x_i \,, \tag{1.1}$$

em que x_i é o tamanho do i-ésimo passo. Por sua vez, a consideração de que todos passos são estatisticamente idênticos, tomada incialmente por simplicidade, conduz a:

$$\left\langle X^{(n)} \right\rangle = \left\langle \sum_{i=1}^{n} x_i \right\rangle = \sum_{i=1}^{n} \left\langle x_i \right\rangle = n \left\langle x \right\rangle , \tag{1.2}$$

em que $\left\langle X^{(n)} \right\rangle$ e $\left\langle x_i \right\rangle = \left\langle x \right\rangle$ são os valores médios de $X^{(n)}$ e de x_i, respectivamente.

[1] Do inglês, *detrended fluctuation analysis*.
[2] Do inglês, *fluctuation analysis*.

O desvio padrão de $X^{(n)}$ é, por sua vez:

$$\begin{aligned}\sigma^{(n)} &= \left[\left\langle \left(X^{(n)} - \left\langle X^{(n)}\right\rangle\right)^2 \right\rangle\right]^{1/2} \\ &= \left[\left\langle \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \langle x_i\rangle\right)^2 \right\rangle\right]^{1/2} \\ &= \left[\left\langle \left(\sum_{i=1}^{n} \left(x_i - \langle x_i\rangle\right)\right)^2 \right\rangle\right]^{1/2} .\end{aligned} \tag{1.3}$$

Com o objetivo de reescrever $\sigma^{(n)}$ em uma forma mais conveniente para a discussão que se segue, usa-se, para uma série genérica y_i, a seguinte relação:

$$\left\langle \left(\sum_{i=1}^{n} y_i\right)^2 \right\rangle = \sum_{i=1}^{n} \left\langle y_i^2 \right\rangle + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} \langle y_i y_j \rangle \,. \tag{1.4}$$

Além disso, considera-se inicialmente que a série y_i é de caráter aleatório, não-correlacionada e usa-se a já sabida igualdade $\langle y_i \rangle = 0$. Então:

$$\langle y_i y_j \rangle = a^2 \delta_{ij} \,, \tag{1.5}$$

em que $a = \left[\langle y_i^2 \rangle\right]^{1/2}$ é uma constante positiva. Isso porque, para $i \neq j$, tem-se que $\langle y_i y_j \rangle = \langle y_i \rangle \langle y_j \rangle = 0$. Assim, se for adotado $x_i - \langle x_i \rangle \equiv y_i$, a equação 1.3 pode ser reescrita como:

$$\sigma^{(n)} = \left[\sum_{i=1}^{n} \left\langle y_i^2 \right\rangle + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} \langle y_i y_j \rangle\right]^{1/2} . \tag{1.6}$$

Portanto, ao se empregar o resultado 1.5 na equação 1.6, verifica-se que:

$$\sigma^{(n)} = a\, n^{1/2}. \tag{1.7}$$

No entanto, relacionados a esses últimos resultados, pode-se tomar um caso mais geral em que as correlações entre os y_i não sejam desprezadas. Assim, ter-se-ia:

$$\langle y_i y_j \rangle \neq a^2 \delta_{ij}. \tag{1.8}$$

É necessário dizer aqui que, como $y_i = x_i - \langle x_i \rangle$, a correlação entre eles é dita positiva se a probabilidade dos termos da caminhada aleatória x_i e x_j serem ambos simultaneamente maiores ou menores que a média $\langle x \rangle$ fizer com que $\langle y_i y_j \rangle > 0$. Da mesma forma, se a probabilidade da diferença em relação à média para sinais contrários conduzir a $\langle y_i y_j \rangle < 0$, tem-se correlação negativa. Se, por um lado, na série predominam correlações positivas, então:

$$\sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} \langle y_i y_j \rangle > 0. \tag{1.9}$$

Por outro lado, se a magnitude das correlações negativas se sobrepuser às positivas, o resultado será:

$$\sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} \langle y_i y_j \rangle < 0. \tag{1.10}$$

Portanto, sabendo-se que a primeira soma na equação 1.6, por si só, já é responsável por $\sigma^{(n)} \propto n^{1/2}$, a inclusão de 1.9 ou 1.10 deverá conduzir a uma versão mais completa do desvio padrão, que usualmente é da forma:

$$\sigma^{(n)} = \tilde{a}\, n^{\alpha}, \tag{1.11}$$

em que $\tilde{a}$ e α são constantes.

É importante concluir que o fato de se ter $\sigma^{(n)} \propto n^{1/2}$ é uma consequência da relação 1.5, que indica a total ausência de correlação entre os termos da série, ou uma correlação insuficiente para mudar essa proporcionalidade. No entanto, é direta a verificação de que, a partir de 1.6 e 1.9, com $y_i = x_i - \langle x_i \rangle$, $\sigma^{(n)}$ é sistematicamente maior que $a\, n^{1/2}$ quando há predominância de correlações positivas (persistentes). Portanto, a possibilidade de $\sigma^{(n)} = \tilde{a}\, n^{\alpha}$, com $\alpha > 1/2$, é consistente com correlações positivas, ou seja, correlações persistentes podem apontar para $\sigma^{(n)} = \tilde{a}\, n^{\alpha}$. A partir de 1.10, um raciocínio similar mostra que $\sigma^{(n)} = \tilde{a}\, n^{\alpha}$, com $\alpha < 1/2$, é consistente com correlações negativas, ou seja, correlações antipersistentes podem favorecer $\sigma^{(n)} = \tilde{a}\, n^{\alpha}$, com $\alpha < 1/2$. Em geral, $\alpha > 1/2$ está conectado com correlações persistentes de longo alcance, e $\alpha < 1/2$ diz respeito a correlações antipersistentes de longo alcance [46]. Além disso, é comum conectar caminhadas aleatórias com processos difusivos. Em tal cenário, se $\alpha > 1/2$, o processo é dito

superdifusivo, enquanto que $\alpha < 1/2$ se refere a um processo subdifusivo [47]. A difusão usual (normal), por sua vez, corresponde a $\alpha = 1/2$.

Em uma típica situação experimental, há apenas um conjunto finito de valores para $X^{(n)}$, ao invés de sua distribuição de probabilidade, e não é possível calcular exatamente o $\langle X^{(n)} \rangle$, mas, sim, estimá-lo. A melhor estimativa de $\langle X^{(n)} \rangle$ será denotada por $\langle X^{(n)} \rangle_e$ e é dada por:

$$\langle X^{(n)} \rangle_e = \frac{1}{s} \sum_{i=1}^{s} X_i^{(n)} \,, \tag{1.12}$$

em que s é a quantidade de valores que se têm para $X^{(n)}$ e os $X_i^{(n)}$'s são estes valores. Nesse contexto:

$$\sigma_e^{(n)} = \left[\frac{1}{s-1} \sum_{i=1}^{s} \left(X_i^{(n)} - \langle X^{(n)} \rangle_e \right) \right]^{1/2} \tag{1.13}$$

é a melhor estimativa para o desvio padrão de $X^{(n)}$ [48, 49].

Com o objetivo de simplificar as notações, aqui serão suprimidos os subíndices e em $\sigma_e^{(n)}$ e em $\langle X^{(n)} \rangle_e$. Para $s \gg 1$, $1/(s-1) \approx 1/s$, cabe fazer a seguinte substituição:

$$\left\langle \left(X^{(n)} - \left\langle X^{(n)} \right\rangle \right)^2 \right\rangle \to \frac{1}{s} \sum_{i=1}^{s} \left(X_i^{(n)} - \left\langle X_i^{(n)} \right\rangle \right)^2 = \frac{1}{s} \sum_{i=1}^{s} \left(Y_i^{(n)} \right)^2 , \tag{1.14}$$

com

$$Y_i^{(n)} \equiv X_i^{(n)} - \left\langle X^{(n)} \right\rangle \tag{1.15}$$

e

$$\left\langle X^{(n)} \right\rangle = \frac{1}{s} \sum_{i=1}^{s} X_i^{(n)} . \tag{1.16}$$

A expressão para o desvio padrão sobre todas as realizações será, portanto:

$$\sigma^{(n)} = \left[\frac{1}{s} \sum_{i=1}^{s} \left(Y_i^{(n)} \right)^2 \right]^{1/2} , \tag{1.17}$$

que poderia ser comparada à relação de proporcionalidade $\sigma^{(n)} = \tilde{a}\, n^{\alpha}$. É importante deixar claro, aqui, que a igualdade acima é válida somente para o caso em que s é muito grande. Caso s seja finito, o lado direito da igualdade se torna apenas uma estimativa para $\sigma^{(n)}$.

1.3.2 Função de flutuação

Será feita, agora, uma conexão entre o desvio padrão que se acabou de discutir e a análise de correlação de uma série temporal. Para tal, considera-se uma série temporal de M termos. Se for, então, dividida em s janelas, todas de tamanho n, dispor-se-á de $s = M/n$ vetores n-dimensionais. Eventualmente a divisão M/n possui uma parte não-inteira, o que leva a alguns termos da série ficarem de fora das janelas. Em uma situação prática de manipulação de dados, é comum a realização de um procedimento duplo: a contagem a partir do início, deixando-se os últimos termos fora dela e outra contagem a partir do último termo da série, eliminando-se os primeiros que correspondam à parte não inteira da divisão M/n. Desse modo, o número total de janelas a serem analisadas nesse caso hipotético seria de $2s$, em que s é a parte inteira de M/n.

Assim, se forem tomadas estas sub-séries, pode-se pensar que cada uma dessas janelas de n termos como uma realização de uma caminhada aleatória $X_i^{(n)}$, e, consequentemente, escrever o desvio padrão correspondente à equação 1.17. A partir de agora se dará ao desvio padrão $\sigma^{(n)}$ o nome de *função de flutuação*, $\tilde{F}(n)$, isto é, $\tilde{F}\,(n) = \sigma^{(n)}$. Desta maneira, a função de flutuação $\tilde{F}(n)$ é definida por:

$$\tilde{F}\,(n) = \left[\frac{1}{2s}\sum_{i=1}^{2s}\left(Y_i^{(n)}\right)^2\right]^{1/2}, \tag{1.18}$$

de forma que a $Y_i^{(n)}$ se apresente tal qual em 1.15, sendo conveniente substituir $1/2s$ por $1/(2s-1)$, em particular, quando s é pequeno.

Como já visto, em uma caminhada aleatória, é comum se considerar $\sigma^{(n)} \propto n^{\alpha}$. No estudo em questão, a mesma relação é usualmente escrita como

$$\tilde{F}\,(n) \propto n^{h}, \tag{1.19}$$

e se define h como sendo o *expoente de Hurst* [50]. Do mesmo modo como se interpreta o valor do expoente α, conclusões similares podem ser feitas sobre as séries analisadas via função de flutuação $\tilde{F}\,(n)$, empregando-se h nas séries temporais. Disto se infere que séries positivamente (negativamente) correlacionadas devem apresentar $h > (<)\, 1/2$.

Para a ausência de correlações entre os termos da série, deve-se chegar a $h = 1/2$.

Esse procedimento para se obter o expoente de Hurst é comumente chamado de análise de flutuação, FA (abreviação do inglês *fluctuation analysis*). Com isso, tem-se um procedimento para a investigação de correlações em uma série temporal. Como tais correlações se referem àquelas entre elementos de uma mesma série, a análise passa a ser especificamente sobre autocorrelações.

Um fato recorrente nos gráficos de $\tilde{F}(n)$ em escala logarítmica é a presença mais acentuada de flutuações quanto maiores são os valores de n. A explicação para esse fenômeno se justifica no número decrescente de janelas em análise para valores crescentes do tamanho n das janelas, e, desse modo, a lei de potência esperada é mais suscetível a oscilações estatísticas do que aquelas tomadas sobre uma quantidade de réplicas mais extensa. Portanto, é conveniente o descarte de alguns pontos do gráfico correspondente a um número muito pequeno de janelas, ou seja, n deve ser limitado a um valor que não descaracterize a possível lei de potência. Por exemplo, nesta dissertação, $n_{max} = M/4$. Adicionalmente, há ainda outros fatores que podem comprometer conclusões confiáveis acerca de correlações. Em particular, o expoente de Hurst pode conter um resultado residual devido à não-estacionariedade de uma série[3]. Isso quer dizer que, se uma série não oscila em torno de um único ponto central, um valor interpretado como devido puramente a correlações pode se dever a esse tipo de não uniformidade.

1.3.3 DFA

Em geral, as bases de dados que se costumam utilizar podem não satisfazer a condição de estacionariedade. A seguir, expõe-se um método já proposto [51–53] para otimizar a leitura dos resultados para o expoente de Hurst. A esse método se dá o nome de DFA.

Supõe-se que uma série qualquer seja, porventura, não-estacionária. Isso significa que, ao invés de flutuar em torno de um único ponto central, ela transita sobre vários pontos de flutuação. Em geral, a dinâmica que descreve esses pontos pode seguir uma tendência ajustável por uma função (a exemplo da polinomial). O objetivo do procedimento é,

[3]Uma série é dita não-estacionária se sua distribuição de probabilidade não é constante em relação ao tempo, tendo, como consequência, parâmetros como a média e a variância variáveis [54].

agora, remover tendências da série. Em geral, considera-se para isso a série acumulada:

$$X_i^{(n)} = \sum_{j=1}^{n} x_{ij}\,, \tag{1.20}$$

cujas tendências são removidas por meio da subtração desta série por polinômios de ajuste, o que é equivalente a tomar:

$$\tilde{X}_i^{(j)} = X_i^{(j)} - P_i^{(l)}\,(j) = X_i^{(j)} - \sum_{k=0}^{l} a_{ik} j^k, \tag{1.21}$$

em que os subíndices i e j se referem, respectivamente, à janela e ao termo dentro desta janela. Desta maneira, a subsérie acumulada *destendenciada* $\tilde{X}_i^{(j)}$ é obtida a partir da subtração da subsérie original $X_i^{(j)}$ por um polinômio $P_i^{(l)}\,(j)$ de grau l relativo à janela i. Este polinômio tem suas constantes a_{ik} encontradas de maneira que melhor se ajustem à série dentro da janela i.

Para ilustrar esse procedimento para reduzir efeitos de tendências, considera-se um exemplo. Se uma série de $n = 1000$ termos for dividida em subséries de tamanho $n = 5$, cada uma das 200 janelas comportará cinco pontos. Supõe-se, agora, que se queira eliminar uma possível tendência na quarta janela. Logo, deve-se encontrar um polinômio que se ajuste suficientemente bem à subsérie $X_4^{(j)}$ $(1 \leq j < 5)$. A escolha do grau do polinômio é, de certa maneira, arbitrária e, neste exemplo, toma-se $l = 1$, o que implica em $P_4^{(1)}\,(z) = a_{40} + a_{41} z$. Via um método de ajuste conveniente, essas duas constantes podem ser encontradas e a nova subsérie *destendenciada* é obtida:

$$\begin{aligned} \tilde{X}_4^{(j)} &= X_4^{(j)} - P_4^{(1)}\,(j) \\ &= X_4^{(j)} - (a_{40} + a_{41} j) \qquad\qquad \forall \quad 1 < j < 5. \end{aligned} \tag{1.22}$$

De uma maneira geral, a partir da subsérie $\tilde{X}_i^{(j)}$, a função de flutuação *destendenciada* pode ser encontrada:

$$\tilde{Y}_i^{(n)} = \tilde{X}_i^{(n)} - \left\langle \tilde{X}^{(n)} \right\rangle \tag{1.23}$$

e

$$F(n) = \left[\frac{1}{2s} \sum_{i=1}^{2s} \left(\tilde{Y}_i^{(n)} \right)^2 \right]^{1/2} . \tag{1.24}$$

Dá-se, pois, o nome deste método de *análise de flutuação destendenciada*, de onde vem a sigla em inglês DFA (*detrended fluctuation analysis).* Neste ponto, deve estar claro que, ao longo desta dissertação, será usado $F(n)$ para estimar o expoente de Hurst ($F(n) \propto n^h$), via gráficos log-log, investigando-se, assim, autocorrelações presentes nas séries (relacionadas a tamanhos de frases) que foram expostas na seção 1.1. A escolha mais comum para o grau l do polinômio para se diminuir a tendência da série é $l = 1$. Isso porque verifica-se que usar $l = 2$ (ou maior) não conduz a uma mudança significativa no expoente de Hurst obtido. Nessa dissertação, $l = 1$ é o valor usado. Se a série apresentar $h < 0,5$, considera-se em geral a série integrada em vez de sua versão original, para que haja maior precisão no cálculo de h. Assim, se $l = 1$ foi a escolha, este valor deve ser reconsiderado como $l = 2$ em consistência com a integração da série.

Apesar de não ser o foco dessa dissertação, uma função de flutuação generalizada, que depende de um parâmetro q, pode ser concebida a partir da expressão anterior:

$$F_q(n) = \left[\frac{1}{2s} \sum_{i=1}^{2s} \left| \tilde{Y}_i^{(n)} \right|^q \right]^{1/q} . \tag{1.25}$$

Neste caso, tem-se a MDFA, em que a inclusão do M à sigla se refere a *multifractal.* No caso particular dessa dissertação, tem-se $q = 2$ e, portanto, $F_q(n) = F_2(n) = F(n)$, para as análises das séries dos tamanhos, dos módulos e dos sinais.

1.4 Correlações: outros métodos de análise

Antes de se concluir este capítulo é oportuno comentar acerca de outros métodos de investigação de correlações (autocorrelações) em uma série temporal.

O procedimento direto para se investigar a autocorrelação em uma série temporal é via função de autocorrelação. Quando se tem um conjunto de dados x_i compondo uma

série estacionária, a função de autocorrelação é definida como:

$$C(n) = \frac{\langle (x_{i+n} - \mu)(x_i - \mu) \rangle}{\sigma^2}, \tag{1.26}$$

em que

$$\langle (x_{i+n} - \mu)(x_i - \mu) \rangle = \frac{1}{M-n} \sum_{1}^{M-n} (x_{i+n} - \mu)(x_i - \mu) \tag{1.27}$$

é o valor médio do produto $(x_{i+n} - \mu)(x_i - \mu)$, e μ e σ são, respectivamente, a média e o desvio padrão de x_i.

Essa função informa diretamente o quanto o produto do desvio da média de x_i se correlaciona com o desvio da média de x_{i+n}. Usualmente, $C(n)$ decai com o aumento de n. Por exemplo, no caso não-correlacionado, $C(0) = 1$ e $C(n) = 0$ se $n \neq 0$. Se a correlação é de curto alcance, como em $C(n) \propto e^{-\beta n}$, β é um valor característico de decaimento da correlação. Por sua vez, correlações de longo alcance são frequentemente caracterizadas por uma relação do tipo $C(n) \propto n^{-\gamma}$. Neste último caso, limita-se a informar que a conexão entre o expoente de autocorrelação γ e o expoente de Hurst h é $\gamma = 2 - 2h$ [55]. No entanto, quando o conjunto de dados não é grande, o cálculo de $C(n)$ falha em fornecer um resultado preciso que permita o acesso ao expoente de Hurst por meio dessa relação de conexão. Desta forma, a DFA se apresenta mais proveitosa para a investigação do expoente de Hurst do que o uso direto da função de autocorrelação $C(n)$.

Por fim, cabe ressaltar que os métodos FA e DFA não são os únicos que fornecem, a partir de uma base de dados, o expoente de Hurst. Por exemplo, tem-se a análise via *wavelet* [56] e a reescala do alcance da série temporal [57], sendo esses procedimentos alternativos à DFA. De uma maneira geral, todos eles forncecem boas estimativas para o expoente de Hurst, mas aqui, sobre a presente base de dados, será empregada apenas a DFA para a investigação de comportamento autocorrelacionado.

Capítulo 2

Correlações na bíblia em português

Neste capítulo, serão estudadas correlações de longo alcance, analisadas via DFA, de séries temporais baseadas no número de palavras por sentença, extraídas do novo testamento da bíblia em português (Almeida revista e corrigida) [41] e reportadas no capítulo anterior. Primeiramente, serão investigadas correlações na série do número de palavras por sentença. A seguir, as análises serão feitas sobre as duas séries extraídas da série das diferenças: a dos sinais e a dos módulos.

2.1 Linguagem e DFA

No contexto da linguagem, os símbolos (pertencentes a um repertório limitado, como um conjunto de letras ou ideogramas) se combinam de maneira a formar repertórios de níveis superiores (como letras formam palavras, como palavras se agrupam em sentenças, e assim por diante) e, ao longo de um texto, podem ser tratadas probabilisticamente [58]. A lei de Zipf aplicada à linguística relaciona, por exemplo, apenas a frequência de aparecimento de um dado símbolo com o seu *ranking*[1] e, portanto, não leva em conta a ordem em que esses símbolos são dispostos [59]. No entanto, o presente estudo visa levar em consideração vínculos gramaticais e sintáticos presentes nos textos, de forma que uma análise de correlação seja pertinente.

[1] O *ranking* de um símbolo é definido em termos da sua frequência: ao mais frequente se atribui o número 1, ao segundo mais frequente, o 2 etc.

Mais precisamente, nesta dissertação, será investigada a existência de correlações de longo alcance em séries relacionadas aos tamanhos de frases. A técnica aqui empregada para esse fim será a DFA, que já tem sido usada em algumas séries temporais extraídas de textos e em vários outros contextos. Apesar de haver uma breve revisão de DFA no capítulo anterior, alguns aspectos serão ressaltados agora. Esta seção fornece uma conexão direta entre linguagem e DFA. Para aqueles que optaram por não ler a última sessão do capítulo anterior, esta seção serve também para fixar a notação empregada na presente dissertação.

DFA é a sigla em inglês para *detrended fluctuation analysis*, ou seja, uma análise das flutuações de uma série temporal cujas tendências foram minimizadas. O emprego deste método tem se mostrado útil na análise de séries temporais não-estacionárias e ruidosas, a fim de se detectarem correlações que não sejam apenas resultado do seu próprio caráter não-estacionário [60]. Além do uso desse método no estudo de textos, a DFA se mostrou produtiva em vários outros contextos, a exemplo de batidas do coração [61] e de sequências de DNA [62] [63]. É possível, ainda, identificar outras situações semelhantes citadas em [53], como no caso das variações dos índices econômicos [64] e temperatura atmosféria [65], de caminhada humana [66], de espalhamento de raios-X [67], ou mesmo de receptores neurais [68]. Em particular, vários estudos desenvolvidos na Universidade Estadual de Maringá empregaram essa técnica, por exemplo [53], na distribuição de velocidades relacionadas à postura [69] e em atividades psicomotoras [70]. Essas séries têm todas em comum o fato de apresentarem correlações que decaem por lei de potência, e, consequentemente, não há uma escala característica em cada uma delas.

A análise de correlação via DFA tem como ingrediente central a função de flutuação $F(n)$ (equação 1.24), em termos do tamanho n das janelas em que as séries são divididas e onde suas tendências locais são minimizadas. Tipicamente, quando há correlações de longo alcance, $F(n) \propto n^h$. Em um gráfico log-log de $F(n)$, espera-se, portanto, obter uma reta, cuja inclinação fornece o expoente h, conhecido comumente como o expoente de Hurst. Esse expoente é o indicador utilizado ao longo deste trabalho para a detecção de correlações nas séries temporais já apresentadas. Essas séries, por sua vez, são aquelas constituídas pelos tamanhos das frases ao longo de um texto e daquelas obtidas a partir delas. As principais análises foram feitas com as frases sendo delimitadas pelos sinais ".",

“!” e “?”.

Aqui se faz importante um detalhamento do expoente h. Quando esse valor é igual a 0,5, a série em questão não apresenta correlações de longo alcance. Sendo assim, se fosse tomada uma série completamente aleatória (obtida, por exemplo, via embaralhamento de uma outra série), seria esperado obter $h = 0,5$. Vale ressaltar, apesar disso, que, se uma série fornece esse mesmo valor de h, não se pode concluir que é de todo não-correlacionada, pois séries com correlação de curto alcance também exibem o mesmo expoente. Por outro lado, quando $h \neq 0,5$, a série apresenta correlação de longo alcance. Valores crescentes a partir de 0,5 ($h > 0,5$) indicam que há uma correlação de longo alcance positiva (persistente), ou seja, há a tendência de valores similares se seguirem. Em contraposição, valores decrescentes do expoente de Hurst ($h < 0,5$) indicam que a correlação é negativa (antipersistente), de forma que valores díspares se sucedam mais frequentemente. Daqui em diante, será usado simplesmente o termo *correlacionado* para representar séries em que $h > 0,5$, e *anti-correlacionado* nos casos em que $h < 0,5$.

Portanto, foi parte essencial da análise calcular a flutuação $F(n)$ para as séries descritas no capítulo anterior. Como esperado, leis de potência foram obtidas; sendo que, em geral, tem-se $h \neq 0,5$. Para se garantir que os expoentes das séries pertinentes (N_i, Z_i e S_i, conforme a tabela 1.2) fossem um indicativo puro das correlações de longo alcance, tomou-se a versão embaralhada para cada umas delas (N_i^*, Z_i^* e S_i^*). Como se espera que todas as correlações de longo alcance de uma série sejam destruídas no embaralhamento, os seus respectivos coeficientes de Hurst foram comparados a $h = 0,5$. Como usualmente se faz ao estudar leis de potência, os gráficos da flutuação foram dispostos em escala logarítmica[2], os quais iniciam por um tamanho de janela $n_{min} = 4$ ($\log[n_{min}] = 0,6$). As janelas máximas aproximam-se de 1/4 do tamanho da série. Adicionalmente, outras análises foram incluídas para um tamanho fixo (e notavelmente reduzido) máximo de janela.

[2]Salvo menção contrária, todos os logaritmos empregados aqui têm base 10.

2.2 Série dos tamanhos

A série dos tamanhos das sentenças extraídas da bíblia em português (Almeida Revista Corrigida) [41], definida pelo número de palavras em cada sentença i do texto, é o objeto a ser investigado via DFA nesta seção. Para que as funções de flutuação não sejam confundidas umas com as outras, seus índices foram marcados com a série a que se referem. Neste caso, a função de flutuação para a série dos tamanhos (considerando todo o novo testamento da bíblia em português) é definida por $F_N(n)$. O gráfico de $F_N(n)$ (log-log) é apresentado na figura 2.1.

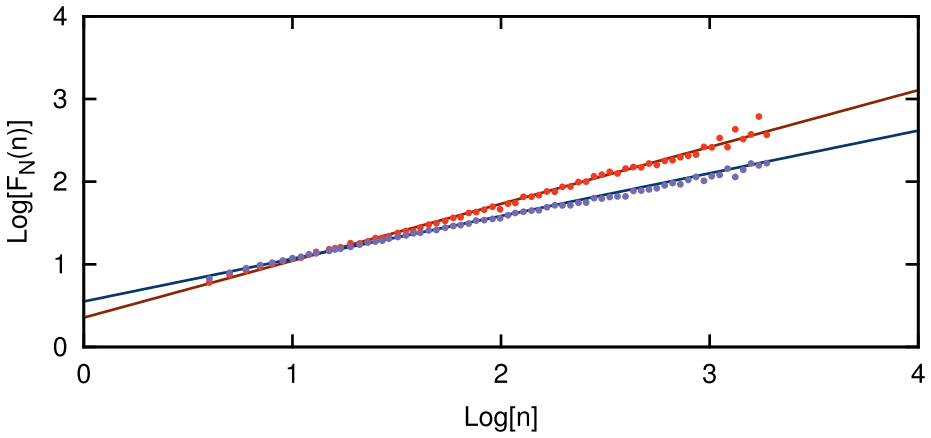

Figura 2.1: **Funções de flutuação para a série dos tamanhos**. Em vermelho e em função do tamanho das janelas, são dispostos os pontos correspondentes à flutuação da série original, $F_N(n)$, bem como um ajuste linear de inclinação $h_N = 0,69$ $(\pm 0,01)$. Em azul, o mesmo foi feito para a DFA da série dos tamanhos embaralhada, $F_{N^*}(n)$, com inclinação igual a $h_{N^*} \approx 0,5$.

A partir dessa figura, é imediata a identificação de comportamento correlacionado para a série dos tamanhos, pois o uso de $F(n) \propto n^h$ (eq. 1.24) no ajuste dos dados da função de flutuação conduziu a $h_N = 0,69$, com boa aproximação. Nota-se que os pontos da figura 2.1 estão em escala logarítmica, igualmente espaçados, isto é, estão log-espaçados. Vale dizer, ainda, que o método de ajuste linear aqui empregado é o da minimização dos quadrados das distâncias no eixo vertical. Aqui foi dito "com boa aproximação" pois, após uma apreciação da figura 2.1, pode-se perceber que há algum resíduo de curvatura no gráfico de $F_N(n)$. Usando-se essa aproximação, bem como as informações citadas na seção anterior acerca do significado dos valores do expoente de Hurst, conclui-se que a série dos tamanhos é correlacionada. Tal fato induz à conclusão

de que sentenças longas têm maior probabilidade de serem seguidas por outras também longas, da mesma forma em que sentenças curtas são frequentemente sucedidas por outras igualmente curtas. Por fim, esse resultado se contrapõe a uma possível predisposição a se acreditar na aleatoriedade completa na disposição dos tamanhos das frases em um texto, o que, neste caso, corresponderia a $h_{N^*} = 0,5$.

Na figura 2.1 há uma ilustração de embaralhamento da série dos tamanhos, que conduziu a $h_{N^*} = 0,46$. Outros embaralhamentos sobre a mesma série (a dos tamanhos) foram realizados diversas vezes e resultados similares foram obtidos[3] ($h_{N^*} \approx 0,5$). Se os expoentes de Hurst das séries embaralhadas também fossem da ordem de 0,69, concluir-se-ia que a correlação presente nas séries não advém da disposição relativa em que se encontram as sentenças, mas, provavelmente, de algo característico dessas sequências. Assim, $h_N = 0,69$ para a série, com $h_{N^*} \approx 0,5$ para sua versão embaralhada, endossa a existência de correlação de longo alcance nos tamanhos de sentenças proveniente da ordem em que são dispostas no texto.

Vale dizer que, apesar de a série dos tamanhos ser composta por todo o novo testamento, e, portanto, compreender de livros escritos por diversos autores, de diferentes épocas e com estilos diferentes de escrita, correlações de longo alcance ainda assim puderam ser detectadas. Dessa maneira, esse resultado é um indicativo de que os mesmos padrões podem ser encontrados em corpos textuais diversos em língua portuguesa, além da bíblia. De forma igualmente relevante, uma outra vertente seria saber se as correlações encontradas aqui têm alguma dependência em relação ao idioma. O próximo capítulo é dedicado à investigação das dependências linguísticas em um texto no que se refere à possibilidade de existência de comportamento (anti-)correlacionado, baseando-se em traduções da bíblia em dezenove outros idiomas.

[3]Mais especificamente, esse processo consistiu na realização de 500 vezes o embaralhamento e a sucessiva extração do expoente de Hurst da série original do novo testamento da bíblia em português. Quando isso foi realizado, submeteram-se os expoentes a uma distribuição de probabilidade aproximadamente gaussiana, pois a curtose foi aferida para $3,14$ e a assimetria, $0,10$. Os parâmetros dessa distribuição consistiram na média $\mu = 0,50$ e desvio padrão $\sigma = 0,02$. Assim, no intervalo $[0,44\ ,\ 0,56]$ estão cerca de $99,7\%$ dos valores do expoente de Hurst h_{N^*} para a série embaralhada.

2.3 Série dos módulos e dos sinais

Identificar um comportamento positivamente correlacionado para a série dos tamanhos proporcionou que se tirassem conclusões apenas com respeito à correlação no número de palavras em sentenças. Por outro lado, quando, a partir dessa série, é obtida a série das diferenças, duas outras informações são passíveis de ser encontradas: (i) se, de uma sentença para a seguinte, o número de palavras aumenta ou diminui; ou (ii) quanto varia o tamanho de uma sentença a outra, em termos do número absoluto de palavras. É importante ressaltar que esta seção emprega as definições de S_i (série dos sinais) e Z_i (série dos módulos) já apresentadas na tabela 1.2, de maneira que um segundo olhar sobre ela possa ajudar a compreender melhor as condições (i) e (ii). Mais objetivamente, a primeira condição se refere à série dos sinais, enquanto que a segunda, à série dos módulos. O gráfico das funções de flutuação relativas às duas séries (a saber, $F_S(n)$ e $F_Z(n)$) são dispostos nas figuras 2.2 e 2.3, juntamente com os pontos obtidos para as séries embaralhadas e os seus ajustes lineares.

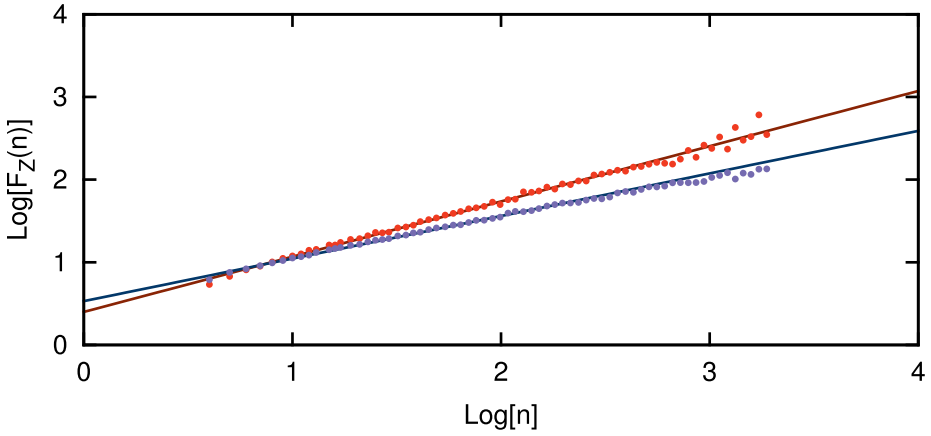

Figura 2.2: **Função de flutuação para a série dos módulos**. Em vermelho e em função do tamanho das janelas, são dispostos os pontos correspondentes à função de flutuação da série dos módulos, $F_Z(n)$, para todo o novo testamento da bíblia em português. A inclinação dessa função em escala logarítmica forneceu $h_Z = 0,67$ $(\pm 0,01)$. Em azul, o mesmo foi feito para a DFA da série embaralhada, $F_{Z^*}(n)$, com $h_{Z^*} \approx 0,5$.

Repetindo-se os mesmos processos feitos para a série dos tamanhos, chega-se também aos valores do expoente de Hurst para as duas séries derivadas da série das diferenças, a saber, h_Z (módulos) e h_S (sinais), juntamente aos seus correspondentes às séries embaralhadas, h_{Z^*} e h_{S^*}. O valor obtido para a série dos módulos indica que a variação do

número de palavras no texto, independentemente do sinal, é positivamente autocorrelacionada ($h_Z = 0,67$).

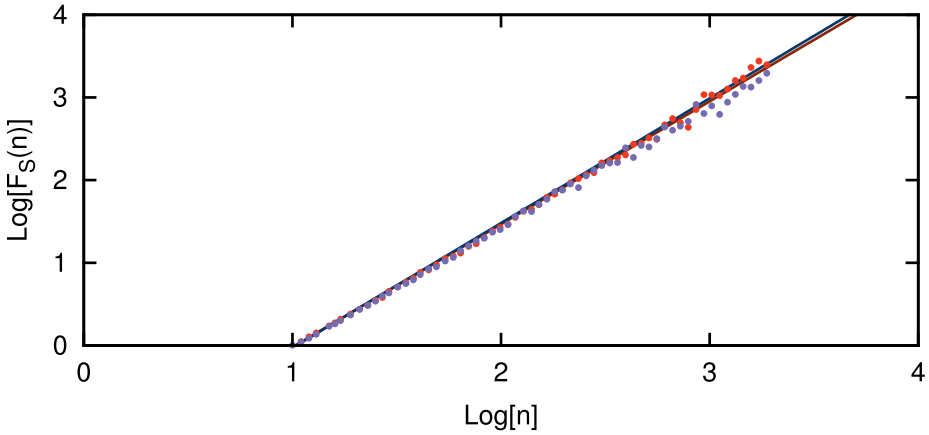

Figura 2.3: **Função de flutuação para a série dos sinais**. Em vermelho e em função do tamanho das janelas, são dispostos os pontos correspondentes à função de flutuação da série dos sinais, $F_S(n)$, para todo o novo testamento da bíblia em português. A inclinação dessa função em escala logarítmica forneceu $h_S = 1,52 - 1 = 0,52$ ($\pm 0,01$). Em azul, o mesmo foi feito para a DFA da série embaralhada, $F_{S^*}(n)$, com $h_{S^*} \approx 0,5$.

Partindo-se da suposição prévia de que a série dos sinais é negativamente autocorrelacionada, um procedimento para tornar os resultados mais confiáveis foi empregado, consistindo na integração[4] de uma dada série [53]. Em outras palavras, cada termo da série integrada é obtido a partir da soma dele com os seus prévios, e, portanto, o expoente de Hurst deve se comportar da seguinte forma $h_{int} \approx h_{orig} + 1$, em que h_{int} se refere ao expoente da série integrada, enquanto h_{orig}, à série não-integrada. Isso é razoável de se pensar, uma vez que o expoente indica o valor proporcional à derivada primeira da função (do tipo $f(x) \propto x^h$) que se aproxima do conjunto de dados experimentais, de forma que a integração da mesma função leva ao aumento de uma unidade ao expoente ($\int f(x)dx \propto x^{h+1}$).

Assim, ao se adotar esse procedimento para a série dos sinais, o expoente obtido foi de $h_S = 1,52 - 1 = 0,52$, atestando ausência de correlação de longo alcance. Os expoentes correspondentes às séries dos módulos e dos sinais embaralhadas permitiram a verificação de que as correlações das séries não-embaralhadas foram perdidas no processo

[4]Quando foi tomado o expoente de Hurst para a série dos sinais não-integrada, obteve-se um valor igual a 0,44. A consideração de que este valor atestava um comportamento anticorrelacionado para a série levou à suposição de que o método da integração deveria fornecer um resultado mais livre de flutuações residuais.

de randomização, uma vez que $h_{Z^*} \approx 0,5$ e, também, que $h_{S^*} \approx 0,5$. Por fim, cabe dizer que a ordem da DFA da série integrada é uma a mais que a série não integrada. Como a presente análise usa DFA de ordem um, a DFA para as séries integradas foi de ordem dois.

2.4 Correlações em livros bíblicos individuais

A fim de se investigar se a coesão dentro dos livros da bíblia em português é razoável dentro do que a técnica de DFA permite visualizar, escolheram-se os sete primeiros livros de todo o novo testamento (que têm maior número de frases) e se calcularam os expoentes de Hurst para cada um deles, referentes às séries dos tamanhos, N_i, dos módulos, Z_i, e dos sinais das diferenças, S_i. A tabela 2.1 expõe esses dados, bem como a média e o desvio padrão a eles relacionados. Vale dizer aqui que se padronizaram os valores dos tamanhos máximos das janelas sobre as quais se aplicou o DFA. Este valor, denotado por n_{max}, será melhor explicado mais adiante.

	h_N	h_Z	h_S
Mateus	0,57	0,62	0,31
Marcos	0,58	0,65	0,38
Lucas	0,58	0,67	0,41
João	0,58	0,57	0,45
Atos dos Apóstolos	0,57	0,66	0,37
Romanos	0,73	0,79	0,37
I Coríntios	0,61	0,58	0,33
Média	0,60	0,65	0,37
Desvio Padrão	0,06	0,07	0,05

Tabela 2.1: **Expoentes de Hurst para livros da bíblia em português**. Os expoentes foram tomados a partir das séries temporais (original, dos módulos e dos sinais) derivadas dos livros mencionados, e a média e o desvio padrão referente a eles também foram expostos. O tamanho máximo das janelas para cada uma das séries tomado foi de $n_{max} = 110$, que é consistente com o valor de n_{max} relativo ao menor destes livros, Romanos. O tamanho mínimo foi mantido como nas análises anteriores, de forma que $n_{min} = 4$.

Nessa apresentação de dados, percebeu-se que os expoentes correspondem a um comportamento correlacionado para as séries original e dos módulos, enquanto que a série dos sinais de cada um dos livros apresentam uma disposição antipersistente. Destaca-se,

também, que o livro Romanos apresentou perceptivelmente nas séries N_i e Z_i um expoente maior em relação aos demais livros. Curiosamente, isso culmina com o fato de ele ser o menor dentre eles, em termos do número de frases: 451 no total. Os tamanhos de todos eles, incuídos os demais livros do novo testamento, podem ser novamente vistos na tabela 1.1.

2.4.1 Múltiplos expoentes de Hurst e o tamanho das janelas

Quando se tem um gráfico da função de flutuação, como mostrado na figura 2.1, são comuns alguns desvios do comportamento $F(n) \propto n^h$. Um deles consiste nas oscilações de caráter aleatório em torno de uma tendência central, tal como se pode ver na mesma figura, em que é constatado um comportamento diferente para os pontos correspondentes à função $F_N(n)$ para $\log[n] > 3$. Essas oscilações são flutuações típicas quando há poucos dados no cálculo de $F(n)$, isto é, para s pequeno na equação 1.24. Outro tipo de desvio ocorre quando mesmo ao serem desconsideradas essas oscilações resta um desvio sistemático da relação do tipo lei de potência entre a função de flutuação e o tamanho das janelas. Uma possível situação em que isso ocorre é o caso em que dois expoentes de Hurst são detectados, cada um deles pertencente a um intervalo do gráfico para o qual uma lei de potência característica é encontrada. Em uma situação mais geral, múltiplos expoentes de Hurst podem ser identificados, ou mesmo situações em que o ajuste da função de flutuação exija a inclusão de outros parâmetros, tornando-a mais complexa. No entanto, é comum se limitar à busca de alguns poucos expoentes, ignorando-se a complexidade de ajustes por funções que, sendo mais complexas, desviam-se do comportamento do tipo lei de potência.

No caso em estudo, para a série dos tamanhos de sentenças, já foi dito que há um pequeno desvio de uma lei de potência pura. Em particular, como será discutido a seguir, o expoente de Hurst para pequenos valores de n é um pouco diferente que no caso em que grandes valores de n são também considerados.

Ainda dentro do contexto da busca do melhor intervalo para o ajuste da reta sobre a função de correlação relativas às séries temporais, dúvidas podem suscitar quanto àquele de maior relevância para que o expoente de Hurst seja confiável. Quando se olha para

os valores individuais do expoente de Hurst para cada um dos livros da bíblia do novo testamento, e sobre eles se toma uma média, percebe-se alguma discrepância do valor aí encontrado em relação ao valor do expoente tomado a partir da DFA sobre o novo testamento como um todo. Possivelmente, este decréscimo no grau de correlação atestados pelos expoentes individuais têm a ver com o intervalo das janelas da DFA consideradas. Assim que se determina um n_{max} (intervalo máximo sobre o qual se medem correlações via DFA, quantificado em termos do número de elementos de determinada série) variável, é esperado que possa variar também o expoente de Hurst. A figura 2.4 apresenta o gráfico motivado por essas dúvidas.

Assim que se analisam os sete primeiros livros individuais dentro do novo testamento da bíblia em português, vê-se que o n_{max} para cada um deles relativos às DFAs realizadas nesta seção é aproximadamente da ordem de 1/4 de toda a série. Em particular, se é tomado como exemplo o livro Romanos − de 451 frases e, portanto, o menor destes sete − tem-se um valor máximo para a janela não muito acima 100. A partir daí, foram calculados (e exibidos na figura 2.4), valores para o expoente de Hurst para a série dos tamanhos cujo intervalo de janelas partia de um mínimo igual a $n_{max} = 100$ ($\log[n_{max}] = 2$), referente ao menor dos n_{max}, considerando-se todos os sete livros, até $n_{max} = 3000$ ($\log[n_{max}] = 3,48$). Ressalta-se que o tamanho de todo o novo testamento desta bíblia é de 7.838 frases.

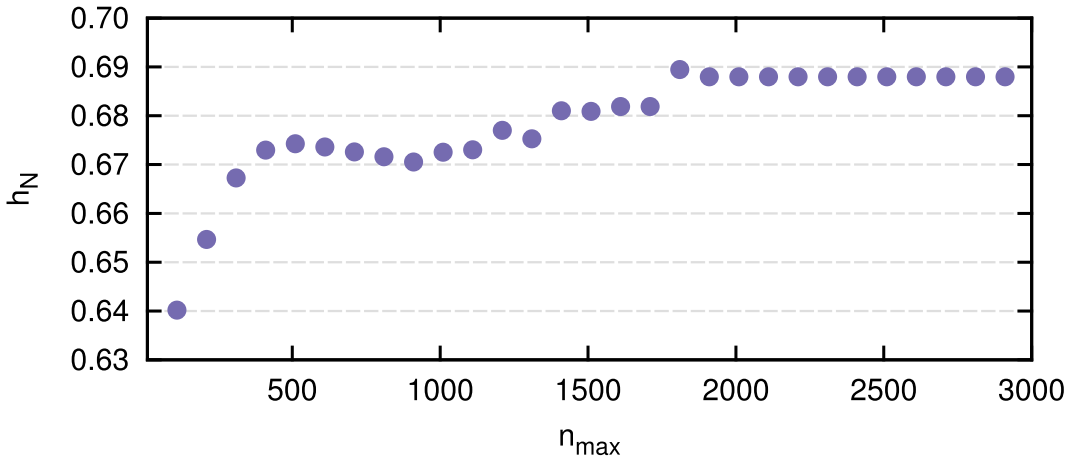

Figura 2.4: **Expoente de Hurst em função do tamanho máximo das janelas**. Os valores para o expoente de Hurst h_N nos casos em que $4 \leq n \leq n_{max}$. Aqui foi empregado o novo testamento da bíblia em português.

É possível observar que, aproximadamente até o valor de $n_{max} = 500$, há uma tendência

marcadamente crescente para o expoente de Hurst em função do número máximo de janelas. Também é razoável dizer que a partir $n_{max} = 100$, o valor do expoente de Hurst apresenta sensível mudança. Parte das análises serão feitas para um valor próximo a este de tamanho máximo de janela, $n_{max} = 110$, correspondente a aproximadamente $M/4$ do menor dos livros do novo testamento dentre os sete primeiros, Romanos ($M = 451$).

2.5 Outros critérios de pontuação

É importante retomar aqui que o foco do presente trabalho é a investigação da existência de correlações dentro de textos. Supôs-se, então, que essas correlações pudessem advir da conexão de diferentes ideias ao longo de um texto, promovidas pelo autor. A partir dessa motivação, surgia uma dúvida: quais são os objetos textuais que podem ser convertidos em objetos matemáticos passíveis de serem analisados sob a ótica das correlações?

Dentre os muitos que se poderiam escolher, como, por exemplo, tamanho (em termos do número de letras) ou tipo de palavras, elegeu-se o *tamanho de sentença* como variável constituinte de uma série temporal dentro da qual se investigou a existência de correlações. Ainda assim, é possível que venha à tona outra dúvida: qual é o critério de definição de uma sentença em um texto? A princípio, as séries temporais aqui analisadas foram construídas a partir da contagem de novas sentenças a cada ponto final, de exclamação ou interrogação (“.”, “!” e “?”), tendo sido dada continuidade à pesquisa com o teste de outras definições.

O que leva a se pensar que eleger outros sinais de pontuação pode alterar a força de correlações presentes em textos é justamente o caráter não-objetivo de separação de ideias que têm esses símbolos. Dizendo de outra forma, é muito tênue a borda que separa o caráter conclusivo do inconclusivo de alguns sinais de pontuação em um texto, ainda que isso se expresse com mais evidência em alguns símbolos do que em outros. Por exemplo, é comum a constatação de que pontos finais quase sempre deem início e concluam ideias sintaticamente autossuficientes, o mesmo não ocorrendo com a vírgula, ponto-e-vírgula ou dois pontos.

Sendo assim, dois outros símbolos menos óbvios − que o ponto final (“.”), de ex-

clamação (“!”) e de interrogação (“?”) − foram incluídos na análise, no que se refere ao seu poder de separar ideias, os dois pontos “:” e o ponto-e-vírgula “;”. A partir disso, três outras séries puderam ser computadas para cada uma já existente: duas que incluíssem cada um dos novos símbolos e outra que contivesse os dois. Os expoentes de Hurst, obtidos de maneira semelhante à dos processos anteriores, para cada uma das séries, podem ser analisados a partir das tabelas 2.2 e 2.3.

	.!?	**.!?:**	**.!?;**	**.!?:;**
Original	0,69	0,63	0,63	0,65
Módulos	0,67	0,59	0,58	0,59
Sinais	0,52	0,40	0,41	0,43

Tabela 2.2: **Expoentes de Hurst para diferentes critérios de pontuação I**. Os expoentes de Hurst foram obtidos considerando-se as séries dos tamanhos de sentenças, identificados a partir de diferentes critérios de divisão de sentença do novo testamento da bíblia em português. A primeira linha mostra os sinais usados como critério de separação. Aqui, $n_{max} = M/4$, em que M é o número de frases do livro.

	.!?	**.!?:**	**.!?;**	**.!?:;**
Original	0,64	0,64	0,60	0,62
Módulos	0,68	0,66	0,65	0,65
Sinais	0,30	0,29	0,33	0,30

Tabela 2.3: **Expoentes de Hurst para diferentes critérios de pontuação II**. Os expoentes de Hurst foram obtidos considerando-se as séries dos tamanhos de sentenças, identificados a partir de diferentes critérios de divisão de sentença do novo testamento da bíblia em português. Aqui, limitou-se o tamanho máximo das janelas da DFA a $n_{max} = 110$.

Previamente, partiu-se da premissa de que, se alguns sinais de pontuação não eram tão bons separadores de ideias, uma divisão de sentença que os incluísse como critério as poderia deixar mais desconexas. Por consequência, uma sucessão de frases desconexas sujeitas a uma análise de correlação (dos tamanhos) poderia estar mais próxima de um caráter aleatório que uma série de frases mais coesas estaria.

O que se viu, no entanto, é que, apesar da inclusão de novos símbolos de pontuação, os expoentes de Hurst se mantiveram muito próximos uns dos outros (em relação às pontuações), comportamento observado na tabela 2.2 e, ainda mais pronunciadamente, na tabela 2.3, quando aplicou-se[5] $n_{max} = 110$. Isso dá margem a duas interpretações diferentes: ou uma série na qual se incluísse novos símbolos como critério de divisão de

[5]Também é interessante notar que o comportamento antipersistente atribuído à série dos sinais é melhor visualizado após a uniformização dos tamanhos máximos das janelas.

frases mantivesse suas ideias, ainda assim, igualmente conexas; ou o impacto da inclusão de novos símbolos fosse significativamente baixo para que alguma diferença pudesse ser notada. Dessa forma, pensou-se: *qual é a magnitude* da inclusão de novos sinais de pontuação?

A tabela 2.4 mostra o número de sinais gráficos ao longo de todo o texto e sua proporção em relação aos outros, de forma a se tentar investigar a resposta a essa pergunta. Nesse sentido, recomenda-se checar a tabela 1.4.

	.	**?**	**!**	**:**	**;**
Núm. absoluto	6.584	992	257	2.301	2.224
Proporção (%)	53,28	8,03	2,08	18,62	18,00

Tabela 2.4: **Número de ocorrências para os sinais gráficos no texto**. Foram contadas todas as vezes em que cada um dos símbolos da tabela ocorriam no novo testamento da bíblia em português, sendo disposta uma linha com o seu número absoluto e uma com a proporção em que ocorrem em relação aos demais.

Um olhar sobre a tabela 2.4 permite visualizar que a contribuição do ponto final (“.”) constitui quase a metade do total dos sinais gráficos considerados durante todo o texto. Adicionalmente, mesmo se intuindo que os sinais de exclamação (“!”) e de interrogação (“?”) têm um poder similar de concluir ou iniciar ideias, estão dispostos em número muito inferior ao ponto final. A partir dessas informações, então, fez-se a análise do expoente de Hurst para o texto em questão se considerado apenas um desses três sinais gráficos. Tomando-se como exemplo a interrogação, os expoentes de Hurst para as séries dos tamanhos, dos módulos e dos sinais forneceram, respectivamente, $h_N = 0,53$, $h_Z = 0,62$ e $h_S = 0,37$ (usando-se $n_{max} = 110$). Ou seja, apenas a série dos tamanhos apresentou sensível mudança no expoente (menos correlacionada) em relação às séries da tabela 2.3.

A partir desse resultado, pode-se imaginar que haja dois fatores em competição que afetem o resultado do expoente de Hurst: o tamanho das frases e a sua coesão. Neste caso, ao se diminuir a quantidade de sinais gráficos, a coesão de ideias em uma sentença pode aumentar. Por outro lado, essa mesma diminuição leva, em geral, ao aumento das sentenças, o que consequentemente pode levar também ao aumento da disparidade dos seus tamanhos (influenciando negativamente, assim, na análise de autocorrelação).

Estendendo-se esse raciocínio ao caso mais geral (em que todos os tipos de pontuação

apresentados na tabela 2.2 são analisados), conclui-se que essa competição pode levar a resultados similares para o expoente de Hurst. Portanto, ao se incorporar os dois pontos e o ponto-e-vírgula à análise, ainda que o tamanho médio das frases diminua (podendo levar ao aumento do expoente de Hurst), a conectividade de ideias também pode diminuir (o que leva à diminuição do mesmo expoente). Como já dito, esses dois comportamentos simultâneos podem levar a uma pouca mudança no expoente da função de correlação, o que foi de fato verificado.

Por fim, haveria ainda a possibilidade de se fazer uma investigação sobre a existência de dois expoentes de Hurst para os casos em que são consideradas frases também aqueles trechos de textos delimitados por ponto-e-vírgula, dois pontos, ou os dois critérios simultaneamente. No entanto, como se pôde ver ao longo dos vários usos da multiplicidade de expoentes de Hurst, as variações se apresentam muito sutilmente, e a implementação de outros critérios para as frases parece não mudar os padrões já visualizados.

Capítulo 3

Correlações na bíblia em vários idiomas

Este capítulo pretende apresentar estudos similares àqueles mostrados no capítulo anterior, a fim de se investigar o poder das constatações obtidas nas análises de correlação para a bíblia em língua portuguesa. Faz parte, então, dos objetivos aqui propostos, a extensão dessas análises a outros idiomas, de forma a se tentar verificar o quanto o expoente de Hurst varia conforme os idiomas em análise. Para que se faça uma conexão entre os valores a serem encontrados para o mesmo expoente nesses textos diversos e as línguas em que estão escritos, faz-se necessário algum conhecimento acerca da estrutura dos idiomas em análise e, portanto, da dinâmica de variação de tamanhos de frases conforme as suas regras sintáticas.

3.1 Apresentação dos idiomas em estudo

Este capítulo compreende, pois, da extensão das análises feitas no capítulo anterior a outros dezenove idiomas. Como já dito no capítulo 1, a base de dados a partir da qual foram obtidos os textos nesses idiomas foi a mesma que serviu à coleta da bíblia em português [71]. Esse *website* contém a tradução da bíblia em número muito superior a vinte idiomas, mas o que explica a restrição da análise a esse número é a necessidade de os

dados se apresentarem de forma clara e objetiva. Para priorizar a completude do trabalho, escolheram-se idiomas suficientemente díspares no que se refere à sua estrutura sintática. Portanto, utilizaram-se tanto representantes de idiomas não indo-europeus como de indo-europeus, como o húngaro e alguns outros pertencentes às sub-famílias germânica, eslava e latina (dentro da qual o português está incluído). Na tabela 3.1, apresentam-se as vinte traduções da bíblia, considerando-se apenas o novo testamento.

Grupo	**Bíblia (idioma)**	M	μ	σ
Germânicas	Bibelen pa hverdagsdansk (dinamarquês)	12.433	16,08	8,88
	Det Norsk Bibelselskap 1930 (norueguês)	7.472	22,80	17,03
	Het Boek (holandês)	14.770	13,14	7,94
	Hoffnung fur Alle (alemão)	13.925	13,80	7,11
	Icelandic Bible (islandês)	10.129	15,70	8,81
	21st Century King James Version (inglês)	8.781	20,87	14,95
	Nya Levande Bibeln (sueco)	12.435	15,92	8,63
NI e albanês	Albanian Bible (albanês)	8.358	21,10	15,71
	Hungarian Karoli (húngaro)	7.846	18,15	12,43
Eslavas	Hrvatski Novi Zavjet Rijeka 2001 (croata)	9.136	15,48	10,32
	Ukrainian Bible (ucraniano)	9.556	14,39	10,53
	Serbian New Testament Easy-to-Read Version (sérvio)	8.289	15,39	9,49
	Russian New Testament Easy-to-Read Version (russo)	10.625	14,73	7,94
	1940 Bulgarian Bible (búlgaro)	7.666	19,79	14,06
Latinas	Almeida Revista e Corrigida 2009 (português)	7.838	20.44	14,78
	Haitian Creole Version (crioulo haitiano)	14.881	15,25	8,53
	La Bibbia della Gioia (italiano)	11.282	16,45	10,77
	La Bible du Semeur(francês)	10.981	17,60	10,44
	La Biblia de las Americas (espanhol)	7.696	22,99	15,91
	Noua Traducere In Limba Romana (romeno)	9.505	18,11	12,38

Tabela 3.1: **Dados referentes às bíblias em vários idiomas**. M é o número de sentenças para o novo testamento de uma dada bíblia, e μ e σ são, respectivamente, a média e o desvio padrão em relação ao número de palavras por sentença. A sigla NI significa não indo-europeu.

3.2 Série dos tamanhos em vários idiomas

A primeira parte da análise da função de flutuação para múltiplos idiomas consistiu no cálculo do expoente de Hurst para a série dos tamanhos, que, seguindo a definição já apresentada no capítulo 1, compreende basicamente da contagem do número de palavras por sentenças em uma bíblia, sendo cada uma destas sentenças consideradas um instante de tempo dentro da série temporal. Mantêm-se, aqui, a mesma definição para esta série: N_i. Mais uma vez, apenas o novo testamento para cada uma das bíblias foi considerado,

conforme detalhado na tabela 1.1.

A abordagem imediata para se analisar uma grande quantidade de expoentes é, de forma usual, a partir dos gráficos das funções de flutuação que geram tais expoentes. Para uma primeira visualização dos resultados, optou-se pela disposição de quatro gráficos da função de flutuação, escolhidos de forma que cada um correspondesse a uma família ou sub-família linguística diferente. Sendo assim, elegeu-se um representante dentro do grupo de idiomas pertencentes a famílias não indo-europeias e outros três dentro da família indo-europeia, mas pertencentes a três diferentes famílias: latina, germânica e eslava. Cada um destes gráficos está disposto nas figuras 3.1 e 3.2.

Para uma análise dos expoentes referentes a todos os idiomas, optou-se pela disposição das tabelas 3.2, 3.3 e 3.4, respectivamente correspondentes aos expoentes da série dos tamanhos, dos módulos e dos sinais.

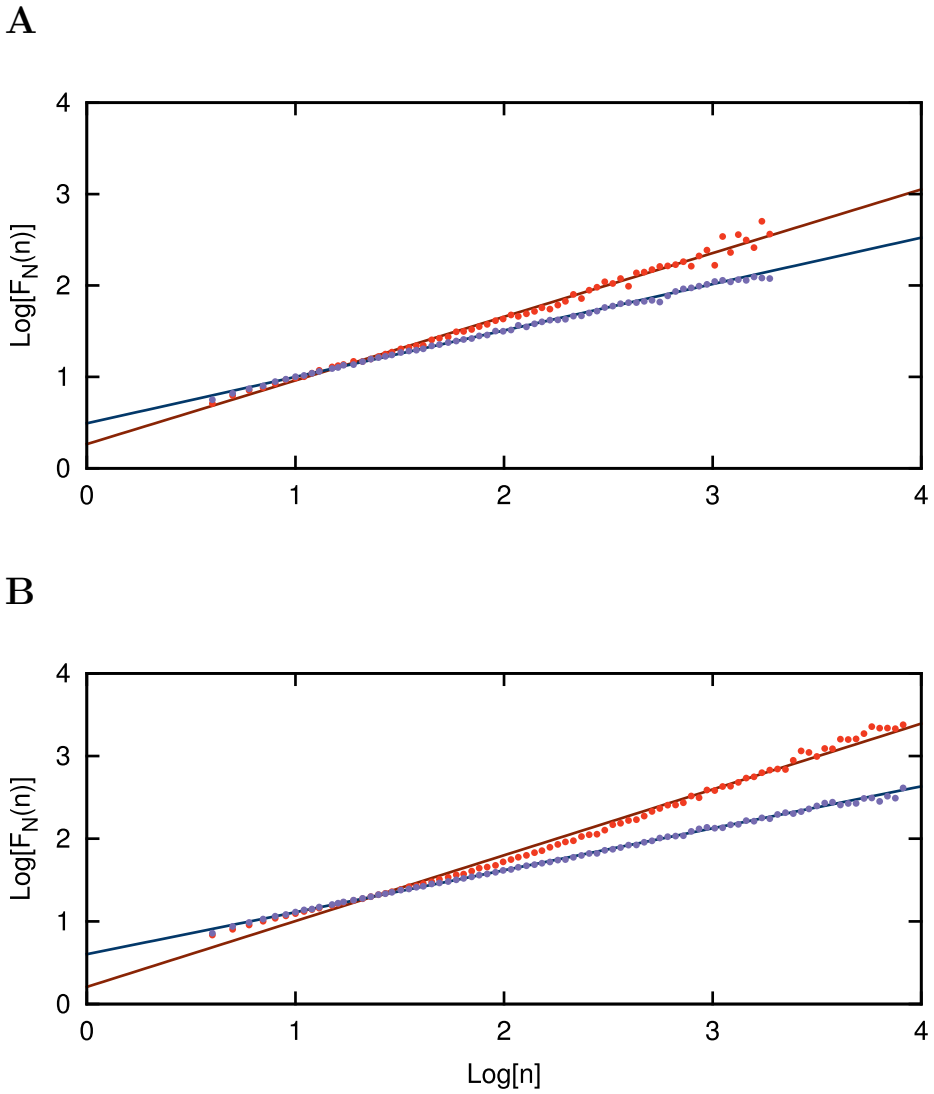

Figura 3.1: **Função de flutuação: série dos tamanhos em húngaro e inglês**. Aqui, as séries dos tamanhos foram obtidas a partir dos textos em húngaro (A) e inglês (B). Em vermelho, os pontos e o ajuste correspondem às séries originais, enquanto que, em azul, às séries embaralhadas. Os ajustes lineares das funções de flutuação para as séries não embaralhadas forneceram expoentes de Hurst iguais a 0,70 ($\pm 0,01$) (A) e a 0,68 ($\pm 0,01$) (B). Para as séries embaralhadas, obteve-se $h \approx 0,5$.

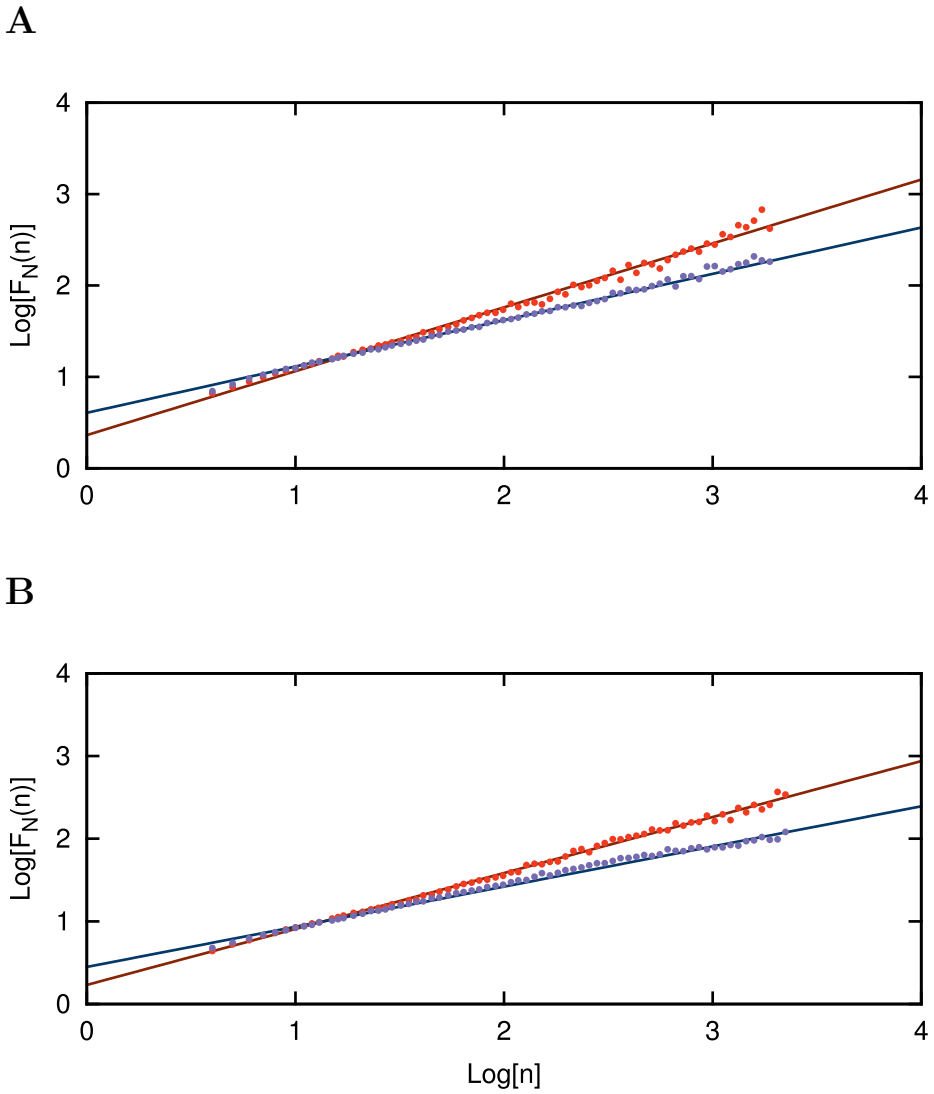

Figura 3.2: **Função de flutuação: série dos tamanhos em espanhol e ucraniano**. Aqui, as séries dos tamanhos foram obtidas a partir dos textos em espanhol (A) e ucraniano (B). Em vermelho, os pontos e o ajuste correspondem às séries originais, enquanto que, em azul, às séries embaralhadas. Os ajustes lineares das funções de flutuação para as séries não embaralhadas forneceram expoentes de Hurst iguais a 0,70 ($\pm 0,01$) (A) e a 0,68 ($\pm 0,01$) (B). Para as séries embaralhadas, obteve-se $h \approx 0,5$.

Faz-se, pois, importante a discussão dos resultados apresentados na tabela e nos gráficos deste capítulo. A tabela 3.1 mostra que, apesar de se lidar com essencialmente o mesmo texto (embora disposto em vários idiomas), há uma relativa discrepância entre os valores de M para diferentes traduções, ou seja, o tamanho do texto em termos do número de sentenças. Por exemplo, a versão dinamarquesa do novo testamento da bíblia, na versão em estudo no presente trabalho, é mais que uma vez e meia o tamanho do mesmo texto em húngaro. Isso poderia, eventualmente, conduzir ao questionamento

sobre se as propriedades estatísticas se mantêm ao longo das diferentes versões.

Neste sentido, um outro estudo realizado se referia às médias e variâncias relativas aos tamanhos das sentenças para cada um dos textos analisados. Uma primeira visualização da figura 3.3 indica que é razoável uma relação linear entre μ e σ. Na sequência, também se dispôs a figura 3.4 para esclarecer algo mais acerca da tabela 3.1: a razoável discrepância nos valores de M entre os idiomas analisados está de acordo com as médias sobre o tamanho de suas frases, nas quais se verifica uma queda sistemática com o aumento do tamanho M ($M = c + d\mu$, com $c = 20.367,60$ e $d = -585,19$). Disso, por sua vez, intuiu-se uma conformidade com a lei de Menzerath [8], exposta na introdução deste trabalho, segundo a qual, se são dispostos corpos constituídos de partes, existe uma tendência rumo ao decréscimo (aumento) do todo, quando suas partes tendem a crescer (diminuir) em tamanho. Esse comportamento é verificado para o caso em estudo, se ao *todo* for comparado o tamanho de todo o novo testamento da bíblia em determinada língua, cujas partes individuais são as sentenças que o compõem. Ou, mais objetivamente, quando se aumenta μ, M decresce.

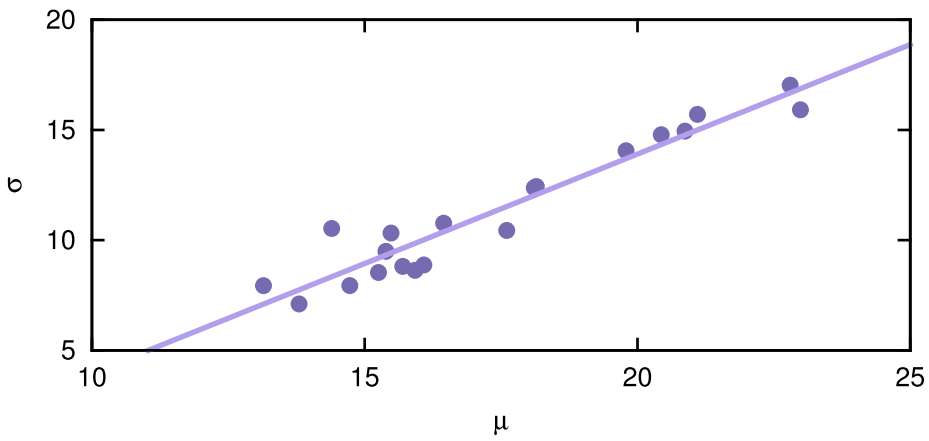

Figura 3.3: **Relação entre a média e o desvio padrão dos tamanhos de frases**. Foram dispostos no gráfico os pares ordenados compostos pela média μ do número de palavras por frase e pelo respectivo desvio padrão σ para os idiomas apresentados na tabela 1.1. Um ajuste linear foi tomado sobre os pontos, cuja inclinação forneceu um valor igual a 0,99 ($\pm 0,07$).

Dessa forma, os passos posteriores foram motivados pela constatação de que a relação entre a média e o desvio padrão no tamanho das sentenças ao longo das traduções disponíveis é aproximadamente linear. Um ajuste sobre as variáveis, $\sigma = a + b\mu$, forneceu o

um valor para a constante b muito próximo de 1 ($a = -6,00$ e $b = 0,99$).

Por fim, é notável a constatação que, apesar de existirem grandes flutuações nos tamanhos das amostras textuais analisadas, algumas propriedades estatísticas ainda se mantêm aproximadamente invariáveis, como essa análise sobre as médias e os desvios padrões permite verificar.

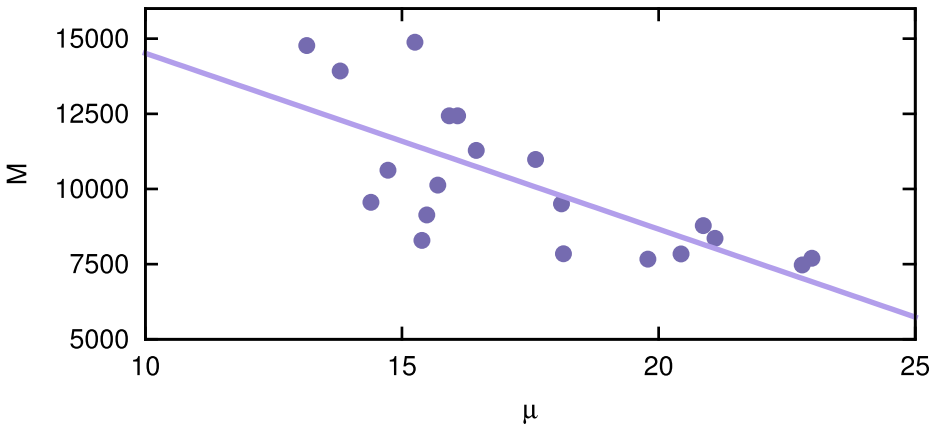

Figura 3.4: **Relação entre número total de frases e média sobre seus tamanhos**. Neste gráfico, constata-se alguma linearidade na relação entre o número de frases M do novo testamento da bíblia em determinado idioma e a média sobre os tamanhos de suas frases. Cada ponto corresponde a um idioma e a inclinação do ajuste linear forneceu um valor igual a -585,19 ($\pm 131,73$).

É importante seguir a análise com os expoentes de Hurst relativos às séries temporais extraídas do novo testamento em vários idiomas. Uma disposição bem completa dos dados permite saber para as três séries N_i, Z_i e S_i (respectivamente, tamanhos, módulos e sinais) os seus expoentes de Hurst sobre todo o gráfico de suas DFAs.

Como intuído, o comportamento do expoente de Hurst variando-se as línguas é muito uniforme. Para a série dos tamanhos, denotada por N_i (tabela 3.2), os expoentes se encontram no intervalo $0,60 \leq h_N \leq 0,70$, com as séries em norueguês, húngaro e espanhol apresentando maior grau de autocorrelação, enquanto aquelas em alemão e sueco se dispondo de maneira menos correlacionada (com o expoente mais próximo a 0,5). Para tal série, obteve-se $\langle h_N \rangle = 0,66$ e $\sigma_{h_N} = 0,03$ quando não se limitou o valor de n_{max}. A fim de se evitar possíveis inconsistências ligadas à não uniformidade nos tamanhos máximos de janelas, numa segunda análise, limitando-se tal tamanho a $n_{max} = 110$, $\langle h_N \rangle = 0,62$ e $\sigma_{h_N} = 0,02$, atestanto um comportamento ainda positivamente correlacionado, mesmo

apresentando uma média menor sobre h_N. Para a análise feita com a condição $n_{max} = 110$, os valores em geral se mostraram menores. Correlações maiores foram atribuídas, neste caso, ao albanês, búlgaro, espanhol e romeno, enquanto que as menores, ao italiano, ao alemão e ao holandês.

Quanto à série dos módulos das diferenças, definida por Z_i (tabela 3.3), pode-se dizer que, de uma maneira geral, considerando-se todos os idiomas, há menor grau de auto-correlação e um translado do intervalo − a saber, $0,57 \leq h_Z \leq 0,69$ − do expoente em direção ao valor de 0,5, tendo a série em ucraniano se apresentado de maneira mais correlacionada e a série em italiano, menos correlacionada. Neste caso, $\langle h_Z \rangle = 0,62$ e $\sigma_{h_Z} = 0,03$, com n_{max} variável conforme o idioma. Consistentemente com o caso anterior (série dos tamanhos), aplicada a condição $n_{max} = 110$ os valores da média e do desvio padrão aferiram $\langle h_Z \rangle = 0,63$ e $\sigma_{h_Z} = 0,03$, em conformidade com o caso em que n_{max} é variável. No caso em que $n_{max} = 110$, os valores recaíram essencialmente acima de 0,6, à exceção do alemão ($h_Z = 0,59$).

Bíblia (idioma)	h_N	h_N ($n_{max} = 110$)
Bibelen pa hverdagsdansk (dinamarquês)	0,64	0,62
Det Norsk Bibelselskap 1930 (norueguês)	0,70	0,64
Het Boek (holandês)	0,64	0,58
Hoffnung fur Alle (alemão)	0,60	0,58
Icelandic Bible (islandês)	0,64	0,62
21st Century King James Version (inglês)	0,68	0,60
Nya Levande Bibeln (sueco)	0,60	0,60
Albanian Bible (albanês)	0,69	0,65
Hungarian Karoli (húngaro)	0,70	0,64
Hrvatski Novi Zavjet Rijeka 2001 (croata)	0,67	0,63
Ukrainian Bible (ucraniano)	0,68	0,64
Serbian New Testament Easy-to-Read Version (sérvio)	0,67	0,63
Russian New Testament Easy-to-Read Version (russo)	0,62	0,62
1940 Bulgarian Bible (búlgaro)	0,69	0,65
Almeida Revista e Corrigida 2009 (português)	0,69	0,64
Haitian Creole Version (crioulo haitiano)	0,64	0,61
La Bibbia della Gioia (italiano)	0,63	0,58
La Bible du Semeur (francês)	0,63	0,61
La Biblia de las Americas (espanhol)	0,70	0,65
Noua Traducere In Limba Romana (romeno)	0,68	0,65

Tabela 3.2: **Expoentes de Hurst para a série dos tamanhos**. A partir da função de flutuação para a série dos tamanhos, $F_N(n)$, o expoente de Hurst h_N foi obtido a partir da inclinação do ajuste linear da mesma função em escala logarítmica. O intervalo aqui utilizado na segunda coluna para o tamanho de janelas parte de $n = 4$ até, aproximadamente, $n = M/4$, em que M é o tamanho da série. Na terceira coluna, $n_{max} = 110$.

Por fim, é possível atribuir à série dos sinais S_i (tabela 3.4) um comportamento antipersistente, dado que o valor da média aplicada sobre seus expoentes é igual $\langle h_S \rangle = 0,43 < 0,5$, com um desvio padrão correspondente a $\sigma_{h_S} = 0,04$, para o caso em que não se impõe um n_{max} fixo. Para o caso em que se fixa $n_{max} = 110$, o comportamento negativamente correlacionado é ainda mais pronunciado ($\langle h_S \rangle = 0,32$, com $\sigma_{h_S} = 0,02$).

3.3 Correlações em livros bíblicos em vários idiomas

Dedica-se esta seção à continuidade da análise feita no capítulo anterior, relativa aos expoentes de Hurst para as séries temporais tomadas a partir dos livros do novo testamento, agora também em outras línguas. Para facilitar a visualização e avaliação dos dados, apenas a série dos tamanhos foi considerada neste momento. Para tal, dispõe-se a tabela 3.5

contendo os expoentes para cada um dos sete primeiros livros da bíblia para cada um dos 20 idiomas em estudo. Há de se ressaltar aqui que o livro Atos dos Apóstolos esteve ausente da base de dados para a bíblia em língua sérvia.

Bíblia (idioma)	h_Z	h_Z ($n_{max} = 110$)
Bibelen pa hverdagsdansk (dinamarquês)	0,60	0,60
Det Norsk Bibelselskap 1930 (norueguês)	0,64	0,64
Het Boek (holandês)	0,57	0,60
Hoffnung fur Alle (alemão)	0,57	0,59
Icelandic Bible (islandês)	0,58	0,63
21st Century King James Version (inglês)	0,63	0,66
Nya Levande Bibeln (sueco)	0,58	0,60
Albanian Bible (albanês)	0,64	0,66
Hungarian Karoli (húngaro)	0,64	0,67
Hrvatski Novi Zavjet Rijeka 2001 (croata)	0,65	0,66
Ukrainian Bible (ucraniano)	0,69	0,64
Serbian New Testament Easy-to-Read Version (sérvio)	0,62	0,62
Russian New Testament Easy-to-Read Version (russo)	0,61	0,62
1940 Bulgarian Bible (búlgaro)	0,66	0,67
Almeida Revista e Corrigida 2009 (português)	0,67	0,68
Haitian Creole Version (crioulo haitiano)	0,61	0,63
La Bibbia della Gioia (italiano)	0,57	0,61
La Bible du Semeur (francês)	0,59	0,60
La Biblia de las Americas (espanhol)	0,65	0,64
Noua Traducere In Limba Romana (romeno)	0,62	0,64

Tabela 3.3: **Expoentes de Hurst para a série dos módulos das diferenças**. A partir da função de flutuação para a série dos módulos, $F_Z(n)$, o expoente de Hurst h_Z foi obtido a partir da inclinação do ajuste linear da mesma função em escala logarítmica. O intervalo aqui utilizado na segunda coluna para o tamanho de janelas parte de $n = 4$ até, aproximadamente, $n = M/4$, em que M é o tamanho da série. Na terceira coluna, $n_{max} = 110$.

A partir de um olhar à tabela 3.5, vê-se que as tendências encontradas para os valores do expoente de Hurst em português se mantêm, em geral, nas outras línguas. Constata-se, também, que o livro de menor tamanho, Romanos, com 451 sentenças, apresenta o maior grau de correlação, em média, enquanto que Lucas, de 1.136 sentenças, o maior deles, tem o expoente que atesta o comportamento menos correlacionado.

Na subseção 2.4.1, foi trazida à discussão a possibilidade de um intervalo de valores para as janelas da DFA diferente alterar o valor do expoente de Hurst, e, também, sobre a plausibilidade do argumento de que as séries utilizadas devem ter seus n_{max} comparáveis

para que também sejam comparáveis seus expoentes de Hurst. Tendo sido verificado que tais inferências são aplicáveis ao novo testamento da bíblia em português, estendeu-se a análise às demais línguas.

Bíblia (idioma)	h_S	h_S ($n_{max} = 110$)
Bibelen pa hverdagsdansk (dinamarquês)	0,39	0,32
Det Norsk Bibelselskap 1930 (norueguês)	0,43	0,33
Het Boek (holandês)	0,45	0,32
Hoffnung fur Alle (alemão)	0,34	0,29
Icelandic Bible (islandês)	0,45	0,34
21st Century King James Version (inglês)	0,39	0,32
Nya Levande Bibeln (sueco)	0,45	0,32
Albanian Bible (albanês)	0,45	0,33
Hungarian Karoli (húngaro)	0,47	0,34
Hrvatski Novi Zavjet Rijeka 2001 (croata)	0,41	0,30
Ukrainian Bible (ucraniano)	0,35	0,31
Serbian New Testament Easy-to-Read Version (sérvio)	0,47	0,30
Russian New Testament Easy-to-Read Version (russo)	0,41	0,32
1940 Bulgarian Bible (búlgaro)	0,41	0,33
Almeida Revista e Corrigida 2009 (português)	0,52	0,36
Haitian Creole Version (crioulo haitiano)	0,40	0,32
La Bibbia della Gioia (italiano)	0,44	0,33
La Bible du Semeur (francês)	0,42	0,34
La Biblia de las Americas (espanhol)	0,46	0,31
Noua Traducere In Limba Romana (romeno)	0,44	0,34

Tabela 3.4: **Expoentes de Hurst para a série dos sinais das diferenças**. A partir da função de flutuação para a série dos sinais, $F_S(n)$, o expoente de Hurst h_S foi obtido a partir da inclinação do ajuste linear da mesma função em escala logarítmica. O intervalo aqui utilizado na segunda coluna para o tamanho de janelas parte de $n = 4$ até, aproximadamente, $n = M/4$,em que M é o tamanho da série. Na terceira coluna, $n_{max} = 110$.

Quando se comparam os expoentes presentes na tabela 3.2 com aqueles encontrados na tabela 3.5, nota-se, de uma maneira geral, que os primeiros indicam mais fortemente a presença de comportamento positivamente correlacionado (ou persistente) do que os últimos[1]. Partindo-se da ideia já apresentada na seção 2.4, calculou-se a média sobre os expoentes de Hurst relativos ao tamanho máximo de janelas igual a $n_{max} = 110$ (como na seção 2.4), utilizando-se a série dos tamanhos do novo testamento da bíblia em todos os idiomas em estudo. Tentou-se responder à questão: a implementação deste procedimento (de redução do intervalo do tamanho das janelas) leva, sensivelmente, à redução ou ao

[1]O livro Romanos parece ser uma forte exceção, ainda que seja o menor livro dos sete considerados.

aumento do expoente de Hurst? A figura 3.5 dá algumas pistas do comportamento do expoente neste contexto.

	Mat.	Marc.	Luc.	Jo.	At. dos Ap.	Rom.	I Cor.
Dinamarquês	0,61	0,59	0,61	0,59	0,59	0,72	0,60
Norueguês	0,54	0,56	0,52	0,64	0,53	0,75	0,70
Holandês	0,60	0,63	0,54	0,59	0,62	0,63	0,55
Alemão	0,54	0,57	0,59	0,58	0,58	0,57	0,55
Islandês	0,61	0,61	0,56	0,57	0,59	0,73	0,65
Inglês	0,59	0,57	0,5	0,60	0,49	0,74	0,55
Sueco	0,60	0,59	0,58	0,60	0,59	0,68	0,51
Albanês	0,62	0,62	0,56	0,62	0,60	0,68	0,59
Húngaro	0,57	0,62	0,53	0,59	0,53	0,74	0,66
Croata	0,58	0,56	0,56	0,58	0,55	0,71	0,63
Ucraniano	0,63	0,61	0,56	0,63	0,62	0,67	0,59
Sérvio	0,59	0,64	0,59	0,63		0,76	0,63
Russo	0,64	0,57	0,67	0,59	0,57	0,68	0,57
Búlgaro	0,61	0,60	0,56	0,60	0,54	0,72	0,52
Português	0,57	0,58	0,58	0,58	0,57	0,73	0,61
Crioulo haitiano	0,62	0,65	0,60	0,57	0,64	0,64	0,59
Italiano	0,55	0,58	0,51	0,66	0,61	0,57	0,52
Francês	0,61	0,63	0,59	0,63	0,56	0,62	0,63
Espanhol	0,59	0,62	0,53	0,58	0,52	0,71	0,61
Romeno	0,59	0,61	0,59	0,60	0,55	0,74	0,68
Média	0,59	0,60	0,57	0,60	0,54	0,69	0,60
Desvio Padrão	0,03	0,03	0,04	0,02	0,10	0,06	0,05

Tabela 3.5: **Expoentes de Hurst para livros da bíblia em vários idiomas**. Os expoentes foram tomados a partir das séries temporais (dos tamanhos) derivadas dos livros mencionados, e a média e o desvio padrão referente a eles também foram expostos. O livro Atos dos Apóstolos, em língua sérvia, não esteve presente na base de dados, justificando a ausência do expoente na exposição desta tabela. Seguindo a mesma padronização utilizada na seção 2.4, utilizou-se n_{max}=110.

Por meio de uma leitura do gráfico da figura 3.5, pode-se ver que, para quase todos os idiomas, à exceção do russo e do ucraniano, manteve-se o que foi verificado para o texto em língua portuguesa: limitar o tamanho máximo de janelas, cujas possíveis correlações são aqui investigadas via DFA, leva à diminuição do expoente de Hurst e, assim, a um indício de maior consistência entre o grau de correlação encontrado individualmente para cada um dos sete primeiros livros do novo testamento e o mesmo texto de maneira integral.

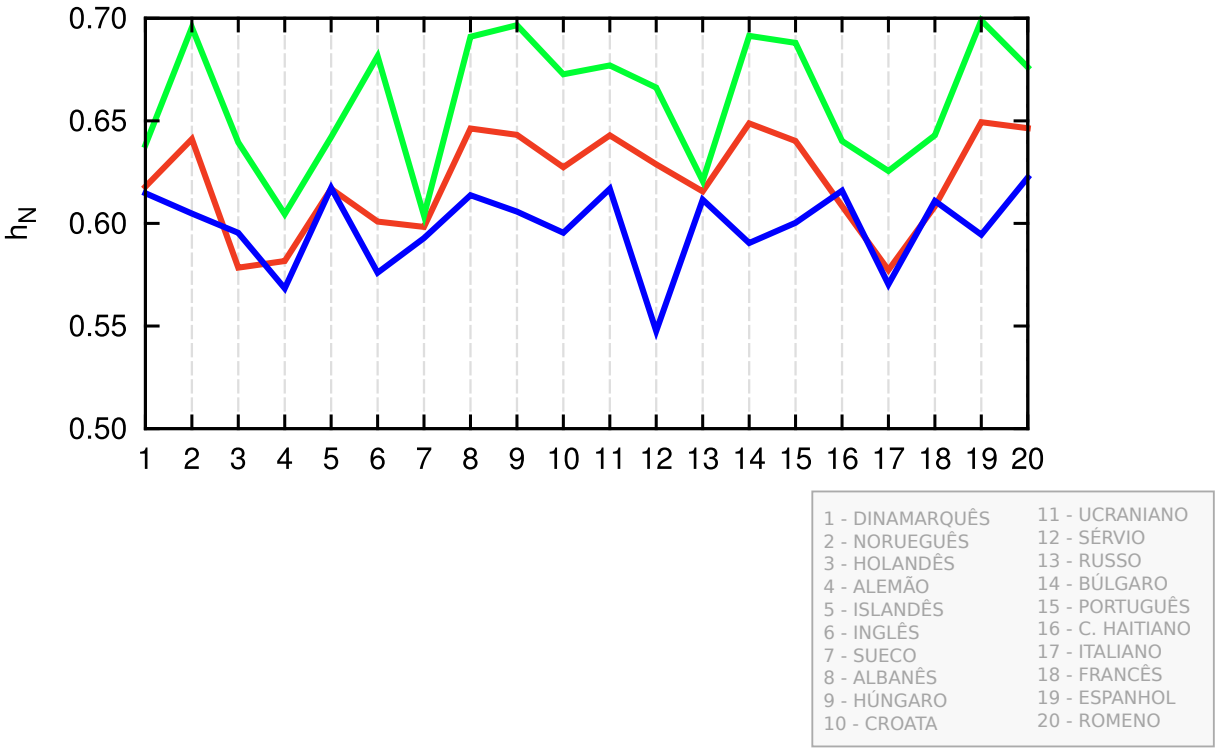

Figura 3.5: **Expoentes de Hurst obtidos via três maneiras para a série dos tamanhos**. Os vértices das linhas verdes correspondem aos expoentes de Hurst obtidos originalmente para a série dos tamanhos relativa ao novo testamento, mantidos seus $n_{max} = M/4$. Aqueles das linhas azuis correspondem à média dos expoentes referentes aos sete primeiros livros do novo testamento, todos eles restritos a $n_{max} = 110$. Os novos expoentes, obtidos a partir da restrição do tamanho máximo de janela a $n_{max} = 110$ à série composta por todo o novo testamento da bíblia, são representados pelos encontros dos segmentos de linhas vermelhos. As vinte línguas são consideradas neste esquema.

Capítulo 4

Conclusão

É importante pontuar os objetivos propostos neste trabalho a fim de confrontá-los com os resultados. O trabalho todo baseou-se na investigação da existência de correlações em textos. O texto escolhido para tal fim foi o novo testamento da bíblia, dado que a diversidade de estilos de escrita encontrada ao longo deste texto é relativamente grande. Isso se deve à quantidade de autores, que viveram em épocas bem espaçadas temporalmente. Esta obra ainda se apresenta num vasto número de traduções [40], o que permite à análise recair também sobre as línguas, bem como sobre se a escolha de um determinado idioma em especial pode mudar os resultados acerca de correlações.

Para que fosse viabilizada essa análise, algumas metodologias foram definidas. Análises prévias da função de correlação já têm sido consideradas há algum tempo [27]. Supôs-se, então, que as séries poderiam apresentar a característica de não-estacionariedade, e um processo de se retirar as tendências temporais foi empregado. Ainda no mesmo capítulo dedicado à apresentação dos dados (capítulo 1), explica-se como foram executados esses passos, comparando-se as séries temporais a caminhadas aleatórias. O expoente de Hurst, o indicativo de correlações desta dissertação, foi encontrado a partir da DFA aplicada a três séries temporais diferentes.

A primeira dessas três séries analisadas foi a própria série temporal do tamanho de sentenças, e as outras duas, derivadas daquela: uma delas, a dos módulos e a outra, a dos sinais. A dos módulos das diferenças consistia no valor absoluto da diferença de tamanho

de duas sentenças consecutivas, enquanto que a dos sinais consistia no sinal (valor dividido pelo módulo) dessas mesmas diferenças. Em suma, o expoente de Hurst para cada uma das séries deveria fornecer uma informação diferente. Em ordem das séries aqui listadas, foram feitas as seguintes: há comportamento correlacionado ou anticorrelacionado para (*i*) os tamanhos de frases ao longo dos textos; (*ii*) a magnitude da variação dos tamanhos de frases; e (*iii*) o sinal dessas mesmas variações? Adicionalmente, ainda se pergunta também se há diferença de comportamento com respeito aos idiomas (vinte foram considerados aqui); se a diferença nos critérios de pontuação empregados na definição de sentenças atestam diferentes expoentes de Hurst; e se é possível a detecção de mais de um expoente de Hurst em uma série de frases.

O primeiro resultado apresentado nesta pesquisa foi o expoente de Hurst para a série original (dos tamanhos) com respeito à língua portuguesa (capítulo 2). O seu valor aferido em $h_N = 0,69$ indicou comportamento positivamente correlacionado para esta série. A conclusão que daí se tira é que há grande probabilidade de que frases grandes sejam seguidas por frases também grandes, o mesmo acontecendo às pequenas. Na realidade, é esperado que as frases se agrupem, ao longo do texto, de acordo com seu tamanho.

Em suma, para a bíblia em português e a série original dela extraída, quando foi analisado o efeito da mudança da definição de frase segundo outros critérios de pontuação, não foram detectadas grandes mudanças[1]. Como já explicado, há a possibilidade de que o fato se justifique com base na competição entre dois fatores: tamanho das frases e coesão entre elas. Esses dois fatores poderiam, quando intensificados, anularem-se mutuamente, atestando mudanças pouco sensíveis no expoente de Hurst.

A análise seguiu com o uso das duas outras séries dentro da mesma metodologia. A série dos módulos, sob a DFA, mostrou-se positivamente correlacionada ($h_Z = 0,67$), enquanto a dos sinais, negativamente correlacionada (em geral, $h_S < 0,5$). A primeira constatação diz respeito à correlação na magnitude das variações de frases ao longo do texto. Isso quer dizer que uma variação relativamente grande de uma frase em relação à

[1]Vale lembrar que a mudança de critério a partir dos três pontos básicos (ponto final, interrogação e exclamação) para um outro que incluísse também os outros sinais (dois pontos, ponto-e-vírgula, ou os dois) foi considerável. No entanto, se analisadas as diferenças entre os expoentes cujos critérios incluem os três sinais e algum dos outros dois sinais (ou os dois), constata-se que são relativamente pequenas. Toda esta observação recém feita diz respeito às DFAs limitadas a $n_{max} = M/4$. Para o caso em que $n_{max} = 110$, ainda menos sensíveis foram as diferenças encontradas.

sua seguinte é provável de ser seguida por uma outra variação comparavelmente grande, já que o expoente se apresenta acima de $h = 0,5$. A constatação de comportamento anticorrelacionado para a série dos sinais indica que há intermitência em relação às probabilidades de que uma frase seja seguida por uma maior ou por uma menor. Juntando-se as duas premissas, uma conclusão possível é que, ao longo do texto, é consistente esperar que as variações de frases (em relação às consecutivas) sejam coesas em tamanho, ainda que se alternem decréscimos e acréscimos de forma intermitente.

Com isso, faz-se importante deixar claro que as conclusões acima obtidas são melhor visualizadas com a técnica de se restringir o tamanho máximo, medido em função do número de elementos, das janelas sobre as quais se aplicou a DFA. Essa padronização foi útil para que as correlações entre séries de diversos tamanhos fossem analisadas de forma razoavelmente comparável. O valor utilizado como padrão para este tamanho máximo foi de $n_{max} = 110$. Isso é justificado com base na recorrência de se aplicar a DFA sobre um intervalo do tamanho das janelas com um máximo igual a aproximadamente 1/4 do tamanho total da série. Uma vez que se consideram livros individuais dentro da bíblia como unidades mais coesas do que todo o novo testamento, utilizado nas presentes análises, empregou-se o $n_{max} = 110$ relativo ao menor dos livros (Romanos), que, nesta versão em português, contém 451 frases.

As análises que acabaram de ser descritas foram repetidas para mais dezenove idiomas, no capítulo 3. Quando analisada a série original, todos os expoentes de Hurst sobre o gráfico obtido por meio da DFA apresentaram-se acima ou iguais a 0,6, sendo que $\langle h_N \rangle = 0,62$ e $\sigma_{h_N} = 0,02$, apontando para correlação persistente.

Quanto ao estudo da série dos módulos, percebe-se que há, assim como na série original, uma tendência positivamente correlacionada, ainda que mais fraca ($0,57 \leq h_Z \leq 0,69$). Para esta série, tirarem-se conclusões acerca do comportamento do expoente no gráfico em log-log da função de flutuação destendenciada é mais difícil, dada uma considerável flutuação entre os valores. Já para a série dos sinais, nota-se que praticamente todos os valores dificilmente ultrapassam 0,5, $\langle h_S \rangle = 0,32$ e $\sigma_{h_N} = 0,02$, como observado inicialmente em português.

Uma conclusão importante neste trabalho é que, após o cálculo de todos os 60 ex-

poentes (20 línguas vezes três tipos de série), não se notaram diferenças significativas conforme a língua ou conforme grupos linguísticos. Desta forma, as correlações que aqui se analisam não devem depender de uma estrutura gramatical específica de determinado idioma. Quando se analisam, por outro lado, DFAs referentes a séries bem menores que todo o novo testamento, como um livro em particular, os expoentes de Hurst variam significantemente em relação ao todo, indicando que, talvez, autores imprimam sinais de correlação diferentes às suas obras em relação a outros.

Por fim, com a elaboração de dois gráficos cujos pontos correspondessem às línguas em estudo, pretendia-se encontrar tendências nas variáveis extraídas dos textos. Os gráficos apresentados relacionavam a média com o desvio padrão (dos tamanhos de frases, figura 3.3) e a média com o número de frases (que compunham o novo testamento da bíblia em determinado idioma, figura 3.4). Como resultado, obtiveram-se dois conjuntos de pontos bem coesos. A tendência dos pontos do gráfico da média pelo desvio padrão se apresentou linearmente crescente, enquanto que o gráfico da média pelo número de frases foi aproximado por uma reta decrescente. Apesar de o número de frases da bíblia em um dado idioma poder ser muito diferente de um relativo a outra língua, ainda sim, as variáveis aqui apresentadas (média, desvio padrão, número absoluto de frases) se dispunham uniformemente correlacionadas por linearidade. Também é digno de análise o caráter decrescente do tamanho do texto todo considerado (o novo testamento da bíblia em vários idiomas) em relação ao tamanho de suas sentenças, em conformidade com a lei de Menzerath, que postula que objetos, sobretudo linguísticos, têm o seu tamanho diminuído (em função da quantidade de partes) assim que o tamanho das partes individuais aumenta.

Tomando como inspiração estudos prévios acerca de correlações em textos, estes capítulos foram em busca de conclusões sobre o papel da linguagem nessas análises. Por exemplo, os sinais de pontuação foram estudados exclusivamente para a bíblia em português, mas um estudo futuro com um escopo diferente poderá, por exemplo, investigar essas mudanças em outras línguas, dado que a variação das regras gramaticais pode promover dinâmicas diferentes sobre os padrões linguísticos percebidos nos textos. Seja como forem procedidos os estudos, é claro que a matéria-prima a partir de onde se tiram os dados será sempre *texto* e, dentro deste contexto, outras obras podem ser analisadas a fim de se estender a análise também aos gêneros linguísticos. As possibilidades são

muitas, inclusive quanto à metodologia: uma análise multifractal dos textos analisados nesta dissertação poderia ser, também, elucidativa (estendendo-se, por exemplo, um estudo multifractal realizado para textos somente em língua inglesa [28]). Pode-se perceber, então, que não apenas em termos de dados númericos é que se compõe a utilidade deste trabalho, mas também na sua proeminência dentro da caminhada investigativa rumo à integração da linguística às análises de correlação.

Referências Bibliográficas

[1] L. Boltzmann, *Über die mechanische Bedeutung des zweiten Hauptsätzes der Wärmetheorie*, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften, Vol. 53, 195-220 (1866).

[2] S. R. Dahmen, *A obra de Boltzmann em física*, Revista Brasileira de Ensino Física, Vol.28, ISSN 1806-1117 (2006).

[3] R. Lopez-Pena, R. Capovilla, R. Garcia-Pelayo e H. Waelbroeck, *Complex system and binary networks*, Springer, Berlim (1995).

[4] N. Boccara, *Modeling complex systems*, Springer-Verlag, Nova Iorque (2004).

[5] C. Castellano, S. Fortunato e V. Loreto, *Statistical physics of social dynamics*, Review of Modern Physics, Vol. 81, 591-646 (2009).

[6] Linguística quantitativa (`http://www.sfs.uni-tuebingen.de/en/ql/research.html`).

[7] O que é linguística? (`http://linguistics.ucsc.edu/about/what-is-linguistics.html`)

[8] S. Eroglu, *Menzerath–Altmann law: statistical mechanical interpretation as applied to a linguistic organization*, Journal of Statistical Physics, Vol. 152, 392-405 (2014).

[9] J. Baixeries, H. Hernández-Fernández, N. Forns e R. Ferrer-i-Cancho, *The parameters of the Menzerath-Altmann law in genomes*, Journal of Quantitative Linguistics, Vol. 20, 94-104 (2013).

[10] R. Ferrer-i-Cancho, N. Forns, A. Hernández-Fernández, Bel-enguix e J. Baixeries, *The challenges of statistical patterns of language: The case of Menzerath's law in genomes*, Complexity, Vol. 18, 11-17 (2013).

[11] R. Ferrer-i-Cancho, A. Hernández-Fernández, J. Baixeries, Ł. Debowski e J. Mačutek, *When is Menzerath-Altmann law mathematically trivial? A new approach*, Statistical Applications in Genetics and Molecular Biology, Vol. 13, 633-644 (2014).

[12] F. Font-Clos, G. Boleda e A. Corral, *A scaling law beyond Zipf's law and its relation to Heaps' law*, New Journal of Physics, Vol. 15, 93033-93048 (2013).

[13] V. Bochkarev, E. Lerner e A. Shevlyakova, *Deviations in the Zipf and Heaps laws in natural languages*, Journal of Physics: Conference Series, Vol. 490 (2014).

[14] X. Yan e P. Minnhagen, *Randomness versus specifics for word-frequency distributions*, Physica A, Vol. 444, 828–837 (2016).

[15] R. Perline, *Zipf's law, the central limit theorem, and the random division of the unit interval*, Physical Review E, Vol. 54, 220-223 (1996).

[16] G. Miller e E. Newman, *Tests of a statistical explanation of the rank-frequency relation for words in written English*, American Journal of Psychology, Vol. 71, 209-218 (1958).

[17] G. Miller, E. Newman e E. Friedman, *Length-frequency statistics for written English*, Information and Control, Vol. 1, 370-38 (1958).

[18] R. Rousseau e Q. Zhang, *Zipf's data on the frequency of Chinese words revisited*, Scientometrics, Vol. 24, 201-220 (1992).

[19] W. Li, *Random texts exhibit Zipf's-law-like word frequency distribution*, IEEE Transactions on Information Theory, 1842-1845 (1992).

[20] R. Ferrer-i-Cancho, *Euclidean distance between syntactically linked words*, Physical Review E, Vol. 70, 056135 (2004).

[21] D. Zanette, *Self-similarity in the taxonomic classification of human languages*, Advances in Complex Systems, Vol. 4, 281-286 (2001).

[22] T. Schürmann e P. Grassberger, *The predictability of letters in written english*, Fractals, Vol. 4, 1-5 (1996).

[23] E. Pechenick, M. Danforth e P. Dodds, *Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution*, PLoS ONE, Vol. 10, 1-24 (2015).

[24] D. Hernández, D. Zanette e I. Samengo, *Information-theoretical analysis of the statistical dependencies among three variables: Applications to written language*, Phys. Rev. E, Vol. 92, 022813 (2015).

[25] C. Cuskley, F. Colaiori, C. Castellano, V. Loreto, M. Pugliese e F. Tria, *The adoption of linguistic rules in native and non-native speakers: Evidence from a Wug task*, Journal of Memory and Language, Vol. 84, 205-223 (2015).

[26] G. Cocho, J. Flores, C. Gershenson, C. Pineda, S. Sánchez, *Rank diversity of languages: generic behavior in computational linguistics*, PLoS ONE, Vol. 10, 1-12 (2015).

[27] W. Ebeling e A. Neiman, *Long-range correlations between letters and sentences in texts*, Physica A, Vol. 215, 233-241 (1995).

[28] I. Grabska-Gradzińska, A. Kulig, J. Kwapień, P. Oświecimka e S. Drożdż, *Multifractal analysis of sentence lengths in English literary texts*, *ArXiv e-prints*, abs/1212.3171, 2012.

[29] E. Magnone, *A novel graphical representation of sentence complexity: the description and its application*, Scientometrics, Vol. 98, 1301-1329 (2014).

[30] S. Furuhashi e Y. Hayakawa, *Lognormality of the distribution of japanese sentence lengths*, Journal of the Physical Society of Japan, Vol. 81, 034004 (2012).

[31] M. Esposti, *Mathematical models of textual data: a short review*, Mathematical Models and Methods for Planet Earth, Vol. 6, 99-110 (2014).

[32] M. Montemurro e P. Pury, *Long-range fractal correlations in literary corpora*, Fractals, Vol. 10, 451-461 (2002).

[33] M. Montemurro, *Quantifying the information in the long-range order of words: Semantic structures and universal linguistic constraints*, Cortex, Vol. 55, 5-16 (2014).

[34] E. Altmann, G. Cristadoro e M. Esposti, *On the origin of long-range correlations in texts*, Proceedings of the National Academy of Sciences, Vol. 109, 11582-11587 (2012).

[35] M. Montemurro e D. Zanette, *Universal entropy of word ordering across linguistic families*, PLoS ONE, Vol. 6, 1-9 (2011).

[36] C. Bian, R. Lin, X. Zhang, Q. Ma e P. Ivanov, *Scaling laws and model of words organization in spoken and written language*, Europhysics Letters, Vol. 113, 18002 (2016).

[37] M. Ausloos, *Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series*, Physical Review E, Vol. 86, 031108 (2012).

[38] M. Ausloos, *Measuring complexity with multifractals in texts. Translation effects*, Chaos, Solitons, and Fractals, Vol. 45, 1349-1357 (2012).

[39] Y. Ashkenazy, P. Ivanov, S. Havlin, C.-K. Peng, A. Goldberger e H. Eugene Stanley, *Magnitude and sign correlations in heartbeat fluctuations*, Physical Review Letters, Vol. 86, 1900-1903 (2001).

[40] Bíblias pelo mundo (`http://worldbibles.org/`).

[41] Almeida revista e corrigida (`http://www.biblegateway.com/versions/Almeida-Revista-e-Corrigida-2009-ARC/#booklist`).

[42] A evolução das línguas europeias a partir do proto-indoeuropeu (*http://www.phil.muni.cz/linguistica/art/blazek/bla-003.pdf*).

[43] Python (`https://www.python.org/`).

[44] Mathematica (`https://www.wolfram.com/mathematica/`).

[45] F. Reif, *Fundamentals of statistical and thermal physics*, McGraw-Hill Book Company, 1965.

[46] R. N. Mantegna e H. E. Stanley, *An introduction to econophysics: correlations and complexity in finance*, Cambridge University Press, New York, EUA, 2000.

[47] R. Metzler e J. Klafter, *The random walk's guide to anomalous diffusion: a fractional dynamics approach*, Physics Reports, Vol. 339, 1-77 (2000).

[48] J. H. Vuolo, *Fundamentos da teoria de erros*, Edgard Blucher, (1996).

[49] M. H. DeGroot e M. J. Schervish, *Probability and statistics*, Addison Wesley, 2012.

[50] R. M. Bryce e K. B. Sprague, *Revisiting detrended fluctuation analysis*, Scientific Reports, Vol. 2, 315 (2012).

[51] C.-K. Peng, S. V. Buldyrev, S. Havlin, H. E. Stanley, A. L. Goldberger, *Mosaic organization of DNA nucleotides*, Phyisical Review E, Vol. 49, 1685-1689 (1994).

[52] J. W. Kantelhardt, E. Koscielny-Bunde, H. A. Rego, S. Havlin, A. Bunde, *Detecting long range correlations with detrended fluctuation analysis*, Physica A, Vol. 295, 441-454 (2001).

[53] S. Picoli Jr *Física estatística dos sistemas complexos: aplicações interdisciplinares.* 2007. 115 f. Tese (Doutorado em Física). Departamento de Física, Universidade Estadual de Maringá, Maringá. 2007.

[54] P. K. Janert, *Data Analysis with open source tools*, Gravenstein O' Reilley Media, 2010.

[55] Estimation of the Hurst parameter of long range dependent time series (`https://www.eecis.udel.edu/~mills/fractal/tr137.pdf`).

[56] I. Simonsen, A. Hansen e O. Magnar Nes, *Determination of the Hurst exponent by use of wavelet transforms*, Physical Review E, Vol. 58, 2779-2787 (1998).

[57] B. B. Mandelbrot e J. R. Wallis, *Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence*, Water Resources Research, Vol. 3, 69-83 (1969).

[58] C. Papadimitriou, K. Karamanos, F. Diakonos, V. Constantoudis e H. Papageorgiou, *Entropy analysis of natural language written texts*, Physica A, Vol. 389, 3260-3266 (2010).

[59] S. Piantadosi, *Zipf's word frequency law in natural language: a critical review and future directions*, Psychonomic Bulletin & Review, Vol. 21, 1112–1130 (2014).

[60] J. Kantelhardt, S. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, H. Stanley, *Multifractal detrended Fluctuation analysis of nonstationary time series*, Physica A, Vol. 316, 87-114 (2002).

[61] R. Mantegna, S. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, M. Simons e H. Stanley, *Linguistic features of noncoding DNA sequences*, Physical Review Letters, Vol. 73, 3169-3172 (1994).

[62] R. Voss, *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences*, Physical Review Letters, Vol. 68, 3805-3808 (1992).

[63] C.-K. Peng, S. Buldyrev, A. Goldberg, S. Havlin, F. Sciortino, M. Simons e H. Stanley, *Long-range correlations in nucleotide sequences*, Nature, Vol. 356, 3805-3808 (1992).

[64] M. Ausloos e K. Ivanova, *Dynamical model and nonextensive statistical mechanics of a market index on large time windows*, Physical Review E, Vol. 68, 046122 (2003).

[65] E. Koscielny-Bunde, A. Bunde, S. Havlin, H. Eduardo Roman, Y. Goldreich, e H.-J. Schellnhuber, *Indication of a universal persistence law governing atmospheric variability*, Physical Review Letters, Vol. 81, 729-732 (1998).

[66] J. M. Hausdorff , S. L. Mitchell, R. Firtion, C.-K. Peng , M.E. Cudkowicz, J. Y. Wei e A. L. Goldberger, *Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease*, Journal of Applied Physiology, Vol. 82, 262-269 (1997).

[67] L. M. Stadler, B. Sepiol, B. Pfau, J. W. Kantelhardt, R. Weinkamer e G. Vogl, *Detrended fluctuation analysis in x-ray photon correlation spectroscopy for determining coarsening dynamics in alloys*, Phyiscal Review E, Vol. 74, 041107 (2006).

[68] S. Bahar, J. W. Kantelhardt, A. Neiman, H. H. A. Rego, D.F. Russell, L. Wilkens, A. Bunde e F. Moss, *Long-range temporal anti-correlations in paddlefish electroreceptors*, Europhysics Letters, Vol. 56, 454 (2001).

[69] E. S. dos Santos, *Controle postural como um sistema complexo: análise da distribuição das velocidades do centro-de-pressão*. 2015. 111 f. Tese (Doutorado em Física). Departamento de Física, Universidade Estadual de Maringá, Maringá. 2015.

[70] A. I. da Silva, *Atividade psicomotora, epidemias e lideranças*. 2015. 113 f. Tese (Doutorado em Física). Departamento de Física, Universidade Estadual de Maringá, Maringá. 2015.

[71] Base de dados (bíblias em várias línguas) (`www.biblegateway.com`).