
Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

# Identificação computacional de padrões interníveis em textos da literatura brasileira

Luciano Alves Machado Júnior

Feira de Santana

2024


Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Luciano Alves Machado Júnior

# Identificação computacional de padrões interníveis em textos da literatura brasileira

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Angelo Conrado Loula

Feira de Santana

2024

Ficha Catalográfica - Biblioteca Central Julieta Carteado - UEFS

M132e
Machado Júnior, Luciano Alves
Identificação computacional de padrões interníveis em textos da literatura brasileira / Luciano Alves Machado Júnior. – 2024.
125 p.: il.

Orientador: Angelo Conrado Loula.
Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Ciência da Computação, Feira de Santana, 2024.

1. Método computacional. 2. Análise textual. 3. Paralelismo textual. I. Loula, Angelo Conrado, orient. II. Universidade Estadual de Feira de Santana. III. Título.

CDU 004.02:806-90-5

Daniela Machado Sampaio Costa - Bibliotecária - CRB-5/2077

Luciano Alves Machado Júnior

# Identificação computacional de padrões interníveis em textos da literatura brasileira

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Feira de Santana, 07 de dezembro de 2023

## BANCA EXAMINADORA

gov.br
Documento assinado digitalmente
ANGELO CONRADO LOULA
Data: 11/03/2024 10:52:20-0300
Verifique em https://validar.iti.gov.br

---

Angelo Conrado Loula (Orientador(a))
Universidade Estadual de Feira de Santana

gov.br
Documento assinado digitalmente
ALVARO JOAO MAGALHAES DE QUEIROZ
Data: 21/03/2024 09:25:28-0300
Verifique em https://validar.iti.gov.br

---

Alvaro João Magalhaes de Queiroz
Universidade Federal de Juiz de Fora

gov.br
Documento assinado digitalmente
RODRIGO TRIPODI CALUMBY
Data: 11/03/2024 11:58:18-0300
Verifique em https://validar.iti.gov.br

---

Rodrigo Tripodi Calumby
Universidade Estadual de Feira de Santana

# Abstract

In literary texts, in poetry or prose, there is an intense presence of poetic devices and recurring linguistic resources at different linguistic levels. The identification of these linguistic resources with the help of computational tools can point out, through quantifiable analyses, patterns of relationships between these levels. The objective of this dissertation is to propose a computational method that allows the identification and correlation of textual patterns between linguistic levels in Brazilian literature texts. To this end, textual characteristics were extracted at different linguistic levels based on the quantification of occurrences through absolute and relative frequencies, both for the full text and for text excerpts, followed by correlation analysis of these quantified characteristics to identify patterns of parallelism between them. The results obtained in order to demonstrate this computational method were extracted from three Brazilian literary works (*Dom Casmurro*, *Macunaíma* and *Os Sertões*). These results contribute to the understanding of the various facets of the method, highlighting its ability to identify and correlate patterns at multiple linguistic levels, while demonstrating the variability of possible results, allowing a quantitative analysis of the patterns present. This research has the potential to open paths for studies in textual analysis, introducing a quantitative approach into a predominantly qualitative field.

**Keywords:** Digital Humanities, Computational textual analysis, Textual patterns, Text parallelism.

# Resumo

Em textos literários, em poesia ou prosa, há a presença intensa de dispositivos poéticos e recursos linguísticos recorrentes em diferentes níveis linguísticos. A identificação desses recursos linguísticos com o auxílio de ferramentas computacionais pode apontar, por meio de análises quantificáveis, padrões de relacionamento entre esses níveis. O objetivo desta dissertação é propor um método computacional que permita a identificação e correlação de padrões textuais entre níveis linguísticos em textos da literatura brasileira. Para isso, foi realizada a extração de características textuais em diferentes níveis linguísticos a partir da quantificação de ocorrências através de frequências absolutas e relativas, tanto para o texto completo quanto para trechos do texto, seguida da análise de correlação dessas características quantificadas para identificar padrões de paralelismo entre elas. Os resultados obtidos com o intuito de demonstrar esse método computacional foram extraídos de três obras literárias brasileiras (*Dom Casmurro*, *Macunaíma* e *Os Sertões*). Esses resultados contribuem para a compreensão das diversas facetas do método, destacando sua capacidade de identificar e correlacionar padrões em múltiplos níveis linguísticos, ao mesmo tempo em que demonstram a variabilidade de resultados possíveis, permitindo uma análise quantitativa dos padrões presentes. Esta pesquisa tem o potencial de abrir caminhos para estudos em análise textual, introduzindo uma abordagem quantitativa em um campo predominantemente qualitativo.

**Palavras-chave:** Humanidades Digitais, Análise textual computacional, Padrões textuais, Paralelismo textual.

# Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

A dissertação foi desenvolvida no Programa de Pós-Graduação em Ciência da Computação (PGCC), tendo como orientador o Prof. Dr. **Angelo Conrado Loula**.

# Agradecimentos

Primeiramente, quero expressar minha profunda gratidão a Deus por me observar, me preparar e me guiar desde o início desta jornada.

Aos meus pais, Clea e Luciano, seu amor e apoio ao longo de toda a minha vida são o alicerce que sustenta cada passo que dei até chegar a este momento. Vocês são verdadeiros heróis em minha história.

Aos meus irmãos, Lucio e Luigi, por representarem a palavra irmandade tão bem, amo vocês.

À minha avó Irene, cujo carinho e cuidado continuam a ser uma bênção em minha vida, dedico este agradecimento com carinho e gratidão.

À minha família, tanto àqueles que estão ao meu lado quanto àqueles que já não estão mais fisicamente neste plano, obrigado por me guiarem onde quer que estejam.

Agradeço ao meu orientador, professor Dr. Angelo Loula, por todo o apoio, orientação e compreensão ao longo de todo o meu processo acadêmico.

Também deixo meus agradecimentos a todos do corpo docente e técnico do Programa de Pós-Graduação em Ciência da Computação (PGCC) da Universidade Estadual de Feira de Santana (UEFS) e a meus colegas de mestrado por todo o suporte, contribuição e colaboração durante nossa árdua jornada.

Um agradecimento que não poderia faltar vai para minha antiga professora, agora colega e acima de tudo amiga, Lais Farias. Por todo o apoio desde a graduação, as mensagens de apoio e a troca de memes me ajudaram nos momentos de tempestade.

Aos meus colegas do IFBA Campus Irecê, por terem dado o apoio e suporte desde o meu ingresso ao mestrado até este momento. No mesmo espaço, onde antes eu fui aluno, hoje tenho alunos que me fazem entender todos os dias o porquê de ter escolhido trabalhar com educação.

A todos os meus amigos que compreenderam que nem sempre eu estaria disponível e mesmo assim me apoiaram e incentivaram a continuar fazendo o que gosto, meu mais sincero obrigado.

A todos os que me auxiliaram e incentivaram durante essa jornada, meu mais sincero obrigado.

*Aos meus pais, Clea e Luciano, por incentivarem minha busca por conhecimento e por me apoiarem nessa trajetória.*

# Sumário

# Alinhamento com a Linha de Pesquisa

**Linha de Pesquisa: Computação Inteligente**

A presente dissertação enquadra-se na linha de pesquisa Computação Inteligente pois busca contribuir com o campo das Humanidades Digitais aplicando tecnologias de Processamento de Linguagem Natural para a extração de características textuais na literatura brasileira e identificar correlações entre elas.

Esta dissertação tem por objetivo propor um método computacional com propósito de identificar e correlacionar padrões textuais em múltiplos níveis linguísticos.

Dessa forma espera-se que esta dissertação possa contribuir com pesquisas nas áreas de Linguística Computacional, Processamento de Linguagem Natural e Humanidades Digitais.

# Lista de Tabelas

# Lista de Figuras

# Capítulo 1

# Introdução

A linguagem é um sistema de comunicação humana, importante para a compreensão cultural, social, histórica e ideológica de uma determinada cultura. Ela apresenta uma riqueza de características, evidenciada pela presença de diferentes níveis linguísticos (Abaurre e Pontara, 2006):

- Nível fonológico: trata dos fonemas da língua, suas possibilidades de combinação em sílabas e a relação que eles mantêm com as letras na escrita alfabética.
- Nível morfológico: lida com as classes de palavras, suas flexões e processos de formação.
- Nível sintático: aborda as funções e relações das palavras nas sentenças da língua.
- Nível semântico: envolve a definição do significado das palavras, as relações de sentido que se estabelecem entre elas e sua organização em um léxico.
- Nível pragmático analisa a relação entre o sentido dos textos e o contexto em que são usados.

Para além do nível linguístico, a linguagem pode ser descrita a partir de suas funções. De acordo com Roman Jakobson (Jakobson e Pomorska, 1985), é possível identificar seis diferentes funções conforme o seu modo de utilização, dependendo da intenção do falante. Essas funções são: função referencial, função emotiva, função poética, função fática, função conativa e função metalinguística. Jakobson (2008) aponta que a função poética possui enfoque na mensagem e utiliza uma variedade de recursos linguísticos.

Dentre esses recursos linguísticos é possível observar a presença de elementos estruturais específicos, como padrões gramaticais, rítmicos, métricos, verbais e visuais. Esses elementos da poesia são definidos como dispositivos poéticos e podem ser demonstrados por meio de repetições, rimas, aliterações, anáforas e outras figuras de linguagem (Rustici, 1997).

Para Jakobson (Jakobson e Pomorska, 1985), a "estrutura da poesia é um paralelismo contínuo". O paralelismo é um fenômeno que se caracteriza pela presença de padrões literários de repetição, sendo definido por Jakobson como um artifício poético de retornos recorrentes.

Esses artifícios poéticos de retornos recorrentes podem ser exemplificados na primeira estrofe do poema *A Estrela*, de Manuel Bandeira:

> "Era uma estrela tão alta!
> Era uma estrela tão fria!
> Era uma estrela sozinha
> Luzindo no fim do dia."
> (Bandeira, 2019)

Nessa estrofe, é possível notar retornos recorrentes caracterizados pelas repetições do trecho "Era uma estrela" e a repetição fonética formando a rima fria/dia. Além disso, existe a presença de um padrão métrico, onde cada verso possui sete sílabas poéticas, sendo assim, versos heptassílabos (ou redondilha maior).

Segundo Jakobson (Jakobson e Pomorska, 1985), estes artifícios de retornos recorrentes, que caracterizam o paralelismo, não estão presentes apenas na linguagem poética, sendo encontrados também na prosa literária.

> "A prosa literária ocupa um lugar intermediário entre a poesia enquanto tal e a língua de comunicação comum, prática, não se devendo esquecer que é incomparavelmente mais difícil analisar um fenômeno intermediário, de transição, do que estudar fenômenos extremos".
> (Jakobson e Pomorska, 1985)

Sendo assim, o paralelismo é um fenômeno presente em uma variedade de produções textuais, incluindo obras literárias. A identificação desse fenômeno tem sido tradicionalmente realizada manualmente, através da leitura e análise crítica. De acordo com Ramsay (2013): "a revolução digital, apesar de todas as suas maravilhas, não penetrou na atividade central dos estudos literários". Entretanto, segundo Jockers e Underwood (2015), o crescente volume de dados digitais disponíveis e o desenvolvimento de novas ferramentas tornam possível a automação dessa análise, tendo a quantificação das características textuais como uma das formas para a identificação automatizada de padrões em textos. Dessa forma, a quantificação de fenômenos como o paralelismo tem o potencial de revolucionar a pesquisa literária, permitindo análises mais eficientes e abrangentes, em contraste com a interpretação qualitativa tradicional dos textos.

A utilização da computação literária (*Literary Computing*), subcampo das Humanidades Digitais pode identificar paralelismos em prosa com mais facilidade. A computação em humanidades envolve a automação de novos métodos de análise, abrangendo áreas como música, teatro, design, pintura e qualquer análise humanística possível. No entanto, seu núcleo permanece focado no discurso de textos escritos

(Busa, 2008). Com isso, torna-se viável o uso de ferramentas computacionais que facilitam a extração de características da linguagem. Uma análise textual realizada por computador pode identificar padrões, semelhanças e características que talvez passem despercebidas a um leitor humano.

Um estudo que utilizou a análise textual por computador foi realizado por Carvalho et al. (2020), no qual os autores propuseram o *MIVES*, um sistema para identificação automática de padrões métricos de versificação em prosa literária brasileira. Nesse estudo, foram revelados paralelismos de estruturas rítmicas variadas, como decassilábicas e dodecassilábicas, que foram identificadas por meio da aplicação computacional proposta.

Outro trabalho que utiliza análise textual computacional para identificação de paralelismos literários é o de Lima et al. (2021). Neste trabalho, os autores fazem a identificação computacional de paralelismos no nível fonético em textos da prosa brasileira e encontram pares de sílabas repetidas e reconhecendo os fonemas em diferentes livros.

Entretanto, os trabalhos acima não buscaram identificar o uso de paralelismos entre os diferentes níveis literários. O trabalho de Carvalho et al. (2020) identificou paralelismos em estruturas métricas, enquanto o de Lima et al. (2021) identificou paralelismos a partir da contagem das repetições de sílabas fonéticas. É sabido que o paralelismo pode acontecer em diferentes níveis linguísticos, e alguns estudos anteriores já abordaram a análise computacional em múltiplos níveis textuais, mas fora do contexto do paralelismo em textos literários. O estudo de Eder et al. (2017), por exemplo, utilizou um protótipo computacional para encontrar padrões estilométricos em diferentes níveis textuais, porém sem estabelecer ou propor relações entre níveis textuais e sem citar a utilização do paralelismo.

A partir disso, vale observar que, apesar da extração de características e paralelismos textuais em diferentes níveis de linguagem já terem sido estudados e relatados em trabalhos anteriores, esses estudos não abordam a relação entre a identificação de paralelismo em um determinado nível de linguagem e outro, e, por isso, não extraem esse potencial relacionamento durante as análises textuais. Algumas perguntas que podem ser respondidas a partir de um estudo sobre o relacionamento entre níveis são: Existe relação entre os paralelismos métricos como os encontrados por Carvalho et al. (2020) e os paralelismos fonéticos encontrados por Lima et al. (2021)? E existe relação entre um determinado sentimento e a presença de padrões estilísticos?

Assim, o objetivo desta dissertação é propor e demonstrar um método computacional de identificação de paralelismos textuais interníveis, a partir de padrões de correlação entre múltiplos níveis linguísticos em obras literárias. Para isso, textos são transformados em sequências numéricas pela extração de características textuais em diferentes níveis linguísticos para realização posterior de análises de correlação entre essas sequências numéricas.

Como consequência desta dissertação, almeja-se que esses novos métodos possam

auxiliar pesquisadores da área de linguística em estudos literários mais aprofundados, com a oportunidade de elaboração de novas perguntas para a área a partir do método computacional proposto, que auxilia na busca das respostas por meio de análises quantificáveis. Este trabalho abre possibilidade de novas análises textuais entre os níveis linguísticos, ofertando uma ferramenta de análise quantitativa em uma área predominantemente qualitativa. Além disso, o trabalho contribui para a difusão de conhecimento acerca das técnicas de Humanidades Digitais, Processamento de Linguagem Natural, Linguística Computacional e Análise Textual assistida por computador.

Dessa forma, foram levantadas as seguintes questões de pesquisa:

- (i) Como a utilização de correlações entre características textuais de diferentes níveis linguísticos pode ser utilizada para a identificação de paralelismos textuais?
- (ii) Quais as implicações da variação dos tipos de correlação na identificação de paralelismos?
- (iii) Como a abordagem quantitativa introduzida por esse método pode auxiliar na análise de textos literários?
- (iv) Quais são as possíveis limitações do método proposto e como essas limitações podem ser superadas?

Os demais capítulos estão organizados da seguinte maneira:

- No **Capítulo 2** é apresentada a **Fundamentação** do trabalho, onde são expostos conceitos e técnicas importantes para a elaboração do mesmo.
- No **Capítulo 3** é apresentada a **Metodologia** do trabalho.
- No **Capítulo 4** são apresentados os **Resultados**, com discussões acerca dos diferentes paralelismos identificados.
- Finalizando no **Capítulo 5**, onde são apresentadas as **Conclusões** do trabalho.

# Capítulo 2

# Fundamentação

Neste capítulo, serão apresentados os conceitos fundamentais para a elaboração da dissertação. Inicialmente, serão introduzidos os conceitos de paralelismo e padrões textuais da poesia e prosa literária. Após isso, será abordado o campo de Humanidades Digitais e sua evolução como campo de estudo. Adicionalmente, este capítulo aborda a Análise Textual e a utilização do Processamento de Linguagem Natural (PLN) com objetivo de extrair características textuais.

Sendo assim, este capítulo abordará a aplicação do PLN em:

- Análises estilométricas: para agrupar documentos e identificar similaridades e diferenças entre textos e identificar autoria;
- Reconhecimento de Entidades Nomeadas: para identificar personagens, lugares e organizações;
- Análise de sentimentos: para identificar e classificar expressões emocionais presentes em diversos tipos de conteúdo textual
- Modelagem de tópicos: para a definição de tópicos e a distribuição de palavras associadas a cada tópico.

Em seguida, será explorado o conceito da análise de correlação, utilizando métodos como gráficos de dispersão e os coeficientes de correlação de Pearson e Spearman.

Ao final do capítulo, será apresentado um conjunto diversificado de estudos acadêmicos relacionados à extração de padrões e características textuais em diferentes níveis linguísticos, com foco em pesquisas que tiveram como objetivo a identificação de padrões em textos literários, a aplicação de diferentes níveis linguísticos na análise textual e a identificação de características relevantes utilizadas para esse propósito.

## 2.1 Padrões textuais e paralelismo

A linguagem poética é marcada por dispositivos poéticos, incluindo figuras de linguagem, que podem ser encontradas na poesia e que caracterizam seu ritmo, seu

significado e seus sons ou fonemas. Esses dispositivos poéticos estão presentes em diferentes níveis linguísticos que possuem figuras de linguagem que caracterizam o nível: fonológico (caracterizam o som), sintático (caracterizam a construção da comunicação), semântico (caracterizam a forma e o significado das palavras) e pragmático (caracterizam os pensamentos, implícitos ou explícitos) (Bechara, 2012).

Algumas dessas figuras de linguagem são:

- (i) Anáfora: usa repetições de palavras ou trechos para dar ênfase em um determinado sentimento ou ponto. Dispositivo poético encontrado no nível semântico.

  Ex:

  > "**É preciso** casar João,
  > **é preciso** suportar, Antônio,
  > **é preciso** odiar Melquíades
  > **é preciso** substituir nós todos."
  > Andrade (2004)

- (ii) Aliteração: repetição de um mesmo som consonantal, é encontrado no nível fonológico.

  Ex:

  > "**V**o**z**es **v**elada**s**, **v**eludo**s**a**s** **v**o**z**e**s**,
  > **V**olúpia**s** do**s** **v**iolõe**s**, **v**o**z**e**s** **v**elada**s**
  > **V**agam no**s** **v**elhos **v**órtice**s** **v**elo**z**e**s**
  > Do**s** **v**ento**s**, **v**iva**s**, **v**ã**s**, **v**ulcani**z**ada**s**."
  > Sousa (1900)

- (iii) Assonância: enquanto a aliteração é caracterizada pela repetição de sons consonantais, a assonância é definida pela repetição de sons vocálicos e também pode ser encontrado no nível fonológico.

  Ex:

  > "V**o**zes veladas,velud**o**sas v**o**zes,
  > V**o**lúpias d**o**s vi**o**l**õ**es, v**o**zes veladas
  > Vagam n**o**s velh**o**s v**ó**rtices vel**o**zes
  > D**o**s vent**o**s, vivas, vãs, vulcanizadas."
  > Sousa (1900)

- (iv) Metáfora: comparação implícita em um determinado trecho, sem utilizar expressões comparativas. Está presente no nível semântico.

  Ex:

  > "Oh, como não me alegra
  > ter este **coração de pedra**!"
  > Meireles (1960)

- (v) Onomatopéia: ocorre quando as palavras representam sons de objetos ou coisas por meio de fonemas. Esse recurso está presente no nível fonológico.

  Ex:

  > "E o sino bem alto
  > **delém-dem**
  > **delém-dem**
  > **delém-dem**
  > **delém-dem!**"
  > Lisboa (2018)

- (vi) Prosopopeia: realiza a personificação de algo inanimado dando características ou ações humanas. Está presente no nível pragmático.

  Ex:

  > "**O cravo caiu doente,**
  > **A rosa o foi visitar;**
  > **O cravo deu um desmaio,**
  > **A rosa pôs-se a chorar.**"
  > Romero (1954)

- (vii) Antítese: apresenta ideias contrárias em um determinado trecho de texto. Esse dispositivo está presente no nível pragmático.

  Ex:

  > "Alma minha gentil, que te partiste
  > Tão cedo desta vida descontente
  > Repousa **lá** no Céu eternamente
  > E viva eu **cá** na terra sempre triste"
  > de Camões (1998)

Além dos exemplos mencionados, existem outros dispositivos poéticos e figuras de linguagem que podem ser encontrados em poesias e poemas. Por esse motivo, Cavanagh et al. (2014) afirma que, ao ouvirmos ou lermos uma poesia, direcionamos nossa atenção a vários níveis linguísticos. Isso nos permite compreender diferentes aspectos, como os estruturais, gramaticais, rítmicos, métricos, verbais e visuais. Esses elementos, em conjunto, moldam a ideia e o conceito do poema.

Entretanto, não é apenas na poesia que encontramos esses dispositivos poéticos, eles podem estar presentes em qualquer gênero textual, inclusive na prosa. Isso reforça a perspectiva de Roman Jakobson, que argumenta que a prosa literária ocupa um lugar intermediário entre a poesia e a língua da comunicação, sendo mais difícil analisar um fenômeno de transição (dispositivos poéticos em prosa) do que um fenômeno extremo (dispositivos poéticos em poesias e poemas)(Jakobson e Pomorska, 1985).

Dispositivos poéticos como a anáfora e a aliteração, que envolvem a repetição de sílabas, fonemas ou palavras no início de versos ou sentenças, podem criar uma

estrutura paralela que destaca aspectos específicos do texto, gerando ritmo e ênfase. Jakobson afirma que a poesia é um paralelismo contínuo e define esse fenômeno como "um artifício poético de retornos recorrentes", e

De acordo com Frog (2014): "O paralelismo tem sido considerado uma característica fundamental do discurso poético". Porém, Jakobson destaca que o paralelismo não se restringe exclusivamente à poesia (Jakobson e Pomorska, 1985). J. Fox (2014) expõe: "O paralelismo descreve a tendência comum de recorrer à ligação de palavras e frases para dar ênfase, autoridade ou significado a uma expressão de ideias", o que também reafirma que o paralelismo não é algo exclusivo da poética, já que em textos da prosa literária também são encontradas repetições e ligações de palavras para ênfase em ideias.

No entanto, de acordo com Jakobson, há uma diferença entre o paralelismo encontrado na poesia e o paralelismo encontrado na prosa. Enquanto na poesia, o verso é quem dita o paralelismo através de estruturas rítmicas, métricas e sonoras de repetição, na prosa são as unidades semânticas que organizam as estruturas paralelas. Essas estruturas influenciam na história, na caracterização dos personagens e nas cenas durante a narração. Além disso, o paralelismo pode ser caracterizado pela presença de um ou mais artifícios linguísticos de retornos recorrentes em um texto, seja ele em poesia ou prosa.

Os paralelismos presentes na prosa literária podem ser encontrados com mais facilidade com a ajuda da computação, através dos estudos no campo das Humanidades Digitais. Já que, para a obtenção desses retornos recorrentes, deve haver uma análise minuciosa e detalhada de um texto com milhares de palavras, tarefa essa que pode demandar dias quando feita manualmente, enquanto computacionalmente esse tempo pode ser reduzido para horas.

## 2.2 Humanidades Digitais

O campo de estudo das Humanidades Digitais teve como um dos principais marcos o trabalho desenvolvido por Roberto Busa (Busa, 1980), quando criou um *index verborum* (índice de palavras) de todas as palavras nas obras de São Tomás de Aquino. Essa ideia resultou em um total de cerca de 11 milhões de palavras em latim. Ao considerar que uma máquina pudesse ajudá-lo, Busa conseguiu transferir os textos para cartões perfurados. Além disso, foi desenvolvido um programa de concordância para o projeto.

De acordo com Hockey (2004), após o trabalho de Busa (1980), diversos pesquisadores iniciaram estudos no campo das Humanidades Digitais. Alguns desses trabalhos se concentraram no estudo de autoria, a exemplo dos trabalhos de Ellegård (1962); Mosteller e Wallace (1963), que visam identificar a autoria de um determinado texto a partir de suas características.

Com o avanço tecnológico, o campo das Humanidades Digitais também progrediu. O

trabalho de Busa (1980), que inicialmente utilizou cartões perfurados, foi transferido para fitas magnéticas e, por fim, publicado em CD-ROM (Busa, 1992).

Atualmente, com a internet, as Humanidades Digitais estão ganhando novos adeptos (Hockey, 2004). Isso se deve ao fato de que estudantes que se especializam nesse campo de estudo podem exercer cargos em publicação eletrônica, tecnologias educacionais e desenvolvimento de multimídia. Além disso, esses alunos também podem se envolver em novas pesquisas dentro desse campo em constante evolução.

Apesar do início das Humanidades Digitais ter sido focado apenas em estudos textuais, hoje, os estudos desse campo estão mais amplos, abordando áreas das ciências humanas e provocando interdisciplinaridades. Conforme Schreibman et al. (2004), o campo das Humanidades Digitais continua interessado em texto, mas, os avanços tecnológicos possibilitaram capturar, manipular e processar outras mídias. Dessa forma, o campo se redefiniu para abranger uma abordagem multimídia.

A partir disso, compreendeu-se que o campo das Humanidades Digitais é responsável pela junção dos estudos em computação com os estudos em ciências humanas.

Drucker (2021) descreve que, para isso, os projetos em Humanidades Digitais iniciam com materiais ou dados, como textos, imagens, sons e arquivos de mídia, podendo incluir a passagem de dados físicos para dados digitais. Posteriormente, esses dados passam por um processamento computacional, seja uma análise estatística conduzida por computador ou uma mineração de dados. Por fim, esses dados processados são analisados e apresentados por pesquisadores do campo através de publicações em *blogs* ou em artigos e revistas das áreas de Humanidades Digitais, computação ou ciências humanas.

Ainda de acordo com Drucker (2021), as pesquisas em Humanidades Digitais envolvem uma série de decisões e tarefas que possibilitam o processamento de dados humanos de diversas áreas, como biologia, ecologia, história, geografia, artes, política, economia, linguística e filosofia, sendo a análise textual uma das abordagens utilizadas nesse contexto.

### 2.2.1 Leitura aproximada e leitura distante

A tecnologia tornou possível a criação de textos em formatos digitais. Essa capacidade foi um dos fundamentos para o desenvolvimento das Humanidades Digitais. Entretanto, essa mesma tecnologia também resultou em um aumento significativo na quantidade de informações que recebemos no nosso dia a dia. Consequentemente, métodos e técnicas para tornar a informação mais compreensível estão se tornando cada vez mais cruciais (Saito et al., 2010).

De acordo com Hawthorn (2000), a técnica chamada leitura aproximada (*close reading*) tem sido utilizada desde o século XX. Ela envolve a interpretação e avaliação de um determinado trecho do texto, abrangendo a análise de: (i) indivíduos, eventos

e ideias, assim como o seu desenvolvimento e interação; (ii) palavras e frases utilizadas; (iii) estrutura e estilo do texto; e (iv) padrões de argumento (Richardson, 2004).

A Figura 2.1 apresenta um exemplo de leitura aproximada em um livro digital por meio da página web do *eMargin* (Kehoe e Gee, 2013). Nessa imagem, é visível a existência de um espaço destinado a anotações, bem como uma barra lateral que viabiliza a inserção de marcações coloridas no texto. Essas funcionalidades permitem que o usuário destaque diferentes ideias encontradas ao longo de sua leitura.

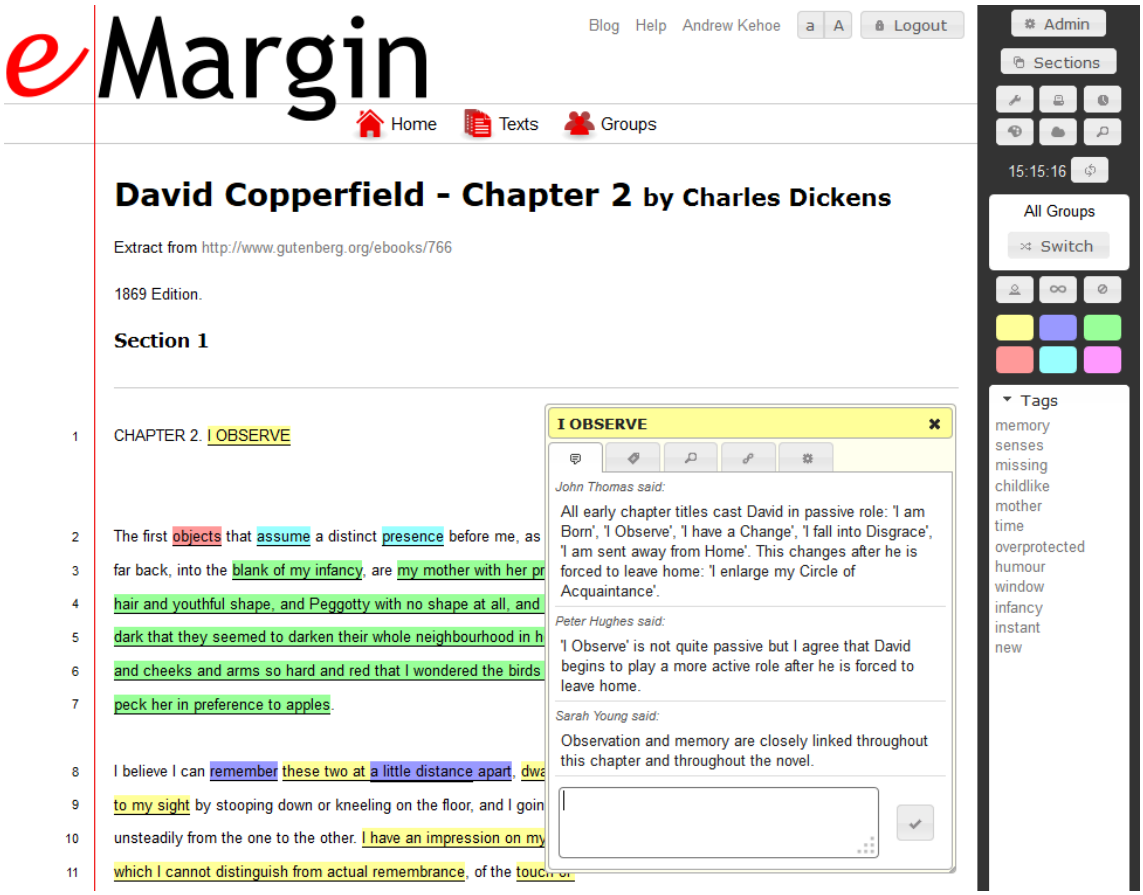


Figura 2.1: Demonstração do funcionamento do sistema web *eMargin* (Kehoe e Gee, 2013) para leitura aproximada

Além disso, a técnica de leitura aproximada possibilita a utilização de diferentes tamanhos de fontes, cores variadas, glifos e conexões diversas para destacar os trechos de interesse. Isso enriquece a análise textual e facilita a identificação de padrões.

Apesar das técnicas de leitura aproximada serem utilizadas desde o século XX, a computação possibilitou o desenvolvimento de outras abordagens para auxiliar na visualização das análises textuais. Em 2005, com a publicação do livro *Graphs, Maps, Trees: Abstract Models for a Literary History* (Moretti, 2005), a técnica de

leitura distante (*distant reading*) se tornou conhecida por apresentar a análise textual por meio de representações gráficas, como grafos, mapas e árvores.

A técnica de leitura distante tem como objetivo criar uma visão abstrata do texto, deslocando o foco da análise do conteúdo textual para a visualização dos elementos presentes em um ou em vários textos (Jänicke et al., 2017).

A Figura 2.2 ilustra um exemplo de leitura distante na forma de um grafo. Esse tipo de técnica de análise textual oferece uma representação abstrata de um texto literário. Na figura, que apresenta o trabalho de Jänicke et al. (2015), é possível identificar as conexões entre algumas entidades nomeadas e palavras encontradas no Novo Testamento. Além disso, através do tamanho da fonte, notam-se as entidades e palavras que mais aparecem no texto.

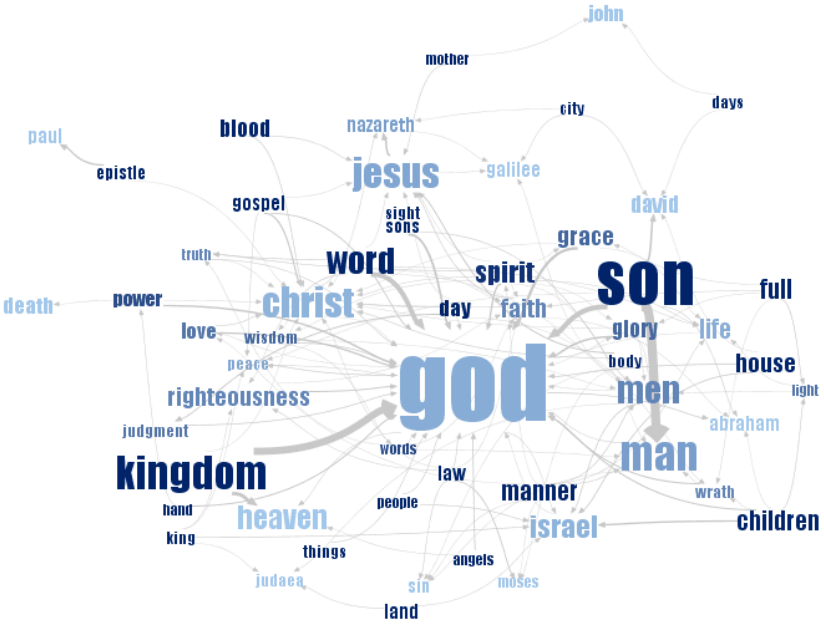


Figura 2.2: Utilização da técnica de leitura distante através de grafo (Jänicke et al., 2015)

Dessa forma, a técnica de leitura distante possibilita a visualização da informação por meio de representações complexas e de grande escala. Essas representações podem ser elaboradas por meio de diversos métodos, como mapas de calor, nuvens de *tags*, mapas, linhas do tempo e/ou grafos.

Ademais, existe a possibilidade da combinação das duas técnicas para uma análise mais profunda. Essa análise combinada pode ser realizada através de sistemas interativos e imersivos, nos quais os usuários têm a capacidade de se "aproximar" e se "distanciar" de determinadas partes do texto para identificar conexões entre elas.

## 2.3 Análise Computacional de Textos

Embora as Humanidades Digitais atualmente adotem abordagens interdisciplinares que não se limitam necessariamente a textos, inicialmente, os estudos na área tinham uma ênfase, talvez restrita, na análise textual Presner (2010). Entretanto, essa análise não se limita a textos escritos, possibilitando a realização de pesquisas que procuram extrair e analisar diversas formas textuais, incluindo multimídia, a exemplo do trabalho de Sun et al. (2020), que utiliza a correlação canônica para analisar as relações ocultas entre texto, áudio e vídeo.

A análise textual realizada pela leitura frase a frase ou palavra a palavra, diretamente pelo pesquisador, é de difícil execução. Afinal, a análise de um grande volume de textos consome muito tempo dos pesquisadores da área de linguagens.

Com isso, a computação, no âmbito das Humanidades Digitais, pode auxiliar na tarefa de análise textual.

> "Os computadores são excepcionalmente adequados para esse tipo de análise, e apenas a intuição e o discernimento humanos, em combinação com o poder computacional bruto das máquinas programadas para agir como ferramentas eletrônicas altamente especializadas, podem tornar alguns textos ou problemas textuais acessíveis aos estudiosos."
> (Rommel, 2004)

Essa afirmação de Rommel (2004) respalda o desenvolvimento de ferramentas computacionais para auxiliar no processamento de texto e nas análises estatísticas de padrões textuais.

Além disso, de acordo com Burrows (2004), quando um grande volume de dados é analisado, certos padrões textuais são mais evidentes.

Como a utilização da computação em problemas de Análise Textual possibilita a análise de textos maiores, é possível obter análises mais aprofundadas. Para isso, faz-se necessária a utilização do Processamento de Linguagem Natural e outras técnicas computacionais.

### 2.3.1 Linguística Computacional e Processamento de Linguagem Natural

Para Vieira e Lima (2001):

> "A linguística computacional é a área de conhecimento que explora as relações entre a linguística e a computação, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural."
> (Vieira e Lima, 2001)

A linguagem natural é aquela utilizada pelos seres humanos através da fala, escrita ou sinais (Libras, por exemplo). Desse modo, a linguística computacional utiliza de técnicas computacionais para a manipulação da linguagem humana.

O Processamento de Linguagem Natural (PLN) é uma área de pesquisa e aplicação que explora como os computadores podem ser usados para entender e manipular texto ou fala em linguagem natural visando realizar tarefas práticas como a criação de *chatbots*, sugestões de textos, reconhecimento de voz, traduções em tempo real, assistentes de voz e outras funcionalidades (Chowdhury, 2003).

Além disso, com o Processamento de Linguagem Natural, é possível utilizar técnicas para classificar, analisar e reconhecer padrões em textos. Entretanto, de acordo com Kannan et al. (2014), para a utilização dessas técnicas, os textos podem precisar passar por uma etapa de pré-processamento. Essa etapa é essencial, pois alguns dados textuais contêm caracteres especiais, numéricos, datas etc., tornando necessária uma filtragem antes de serem processados.

Segundo Kannan et al. (2014), esse pré-processamento pode incluir:

- (i) *Tokenização*: quebra do texto em frases, palavras, símbolos ou outros elementos textuais. Esses pedaços de texto são chamados de *tokens*;
- (ii) Remoção de *stop words*: remoção de palavras que são frequentes em um texto e que não possuem significância na compreensão do conteúdo. É uma tarefa considerada difícil, pois a remoção de palavras pode impactar durante o treinamento de um algoritmo de aprendizado de máquina ou uma análise textual. Além disso, a lista de *stop words* varia de um texto para outro ou de uma língua para outra;
- (iii) *Stemização*: É o processo de juntar variantes de uma palavra em seu radical. Por exemplo: as palavras: gato, gata, gatos e gatas, unem-se ao mesmo radical: gat.

Ademais, Manning et al. (2010) descreve outro exemplo de pré-processamento, conhecido como lematização. Essa técnica consiste em retornar as palavras para sua forma base. Por exemplo, as palavras “gato”, “gata”, “gatos” e “gatas”, são lematizadas para o mesmo lema: “gato”. Já as palavras “subia”, “subiu” e “subistes”, são lematizadas para o mesmo lema: “subir”.

Após a etapa de pré-processamento do texto, torna-se possível empregar diversas técnicas e ferramentas computacionais com o objetivo de extrair características e identificar padrões presentes no texto.

De acordo com Indurkhya e Damerau (2010), uma das técnicas utilizadas para a realização do processamento do texto se chama *Part of Speech* (PoS) *tagging*, que atribui a cada palavra do texto uma marcação correspondente à sua classe gramatical. Essa técnica é importante para alguns tipos de análises textuais, pois os algoritmos de aprendizado de máquina podem utilizar, em certa medida, a identificação das classes gramaticais e a categorização das palavras durante o treinamento.

Ainda de acordo com Indurkhya e Damerau (2010), algumas das técnicas aprendizado de máquina em que pode ser aplicadas a etapa de *PoS-tagging* são: *Support Vector Machines* (SVM), Redes Neurais, Árvores de Decisão, Algoritmos Genéticos e Teoria dos conjuntos difusos (*Fuzzy*).

Uma das aplicações do Processamento de Linguagem Natural é a identificação de padrões textuais, que pode ser utilizada nas técnicas de:

- Análise de estilometria: busca identificar estilos textuais, podendo ser utilizada na identificação de autoria do texto de acordo com os padrões estilísticos encontrados (Ramnial et al., 2016);
- Reconhecimento de entidades nomeadas: tem como objetivo a identificação de entidades no texto, a identificação de entidades no texto, abrangendo categorias como lugares, pessoas, datas, objetos e cidades (Indurkhya e Damerau, 2010);
- Análise de sentimentos: busca identificar, classificar e analisar o teor sentimental de uma sentença (Indurkhya e Damerau, 2010).
- Modelagem de Tópicos: possui como foco principal a identificação de tópicos em um texto a partir de algoritmos de aprendizado profundo (Blei et al., 2003)

Ademais, para realizar o Processamento de Linguagem Natural podem ser necessários diversos subsistemas, já que Vieira e Lima (2001) expõem que a língua natural possui diferentes aspectos: sons, palavras, sentenças, sem contar os seus diferentes níveis linguísticos.

Essas técnicas e subsistemas podem ser utilizados para a extração de características textuais em diferentes textos, literários e não literários.

### 2.3.2 Características Textuais

A extração de características textuais ocorre em diversos níveis linguísticos (Lagutina et al., 2019). Além disso, é possível identificar, do ponto de vista da linguística, os seguintes níveis básicos:

- Nível Fonético: entonação e melodia através da extração de padrões de repetição silábica (conforme o artigo de Lima et al. (2021)), rimas, tonicidade silábica, através da extração da quantidade particular de sílabas de uma sentença (estruturas métricas) (vide o trabalho de Carvalho et al. (2020)), etc.;
- Nível Morfológico: frequência da palavra, *stop words*, sinônimos, antônimos, neologismos, etc.;
- Nível Sintático: classes gramaticais (*part of speech*), reconhecimento de entidades nomeadas, classificação e análise de sentimentos, tipos de coesão sintática, paralelismo sintático, comprimento de sentença, pontuações, etc.;.
- Nível Semântico: relações entre partes do texto, modelagem de tópicos, etc.;

Nos diversos estudos sobre análise textual (Eder et al., 2017; Dell'Orletta et al., 2013; Wanner et al., 2017; Lagutina et al., 2020; Carvalho et al., 2020; Lima et al., 2021), que serão abordados na Seção 2.5, a escolha de características textuais e as formas de extração variam de acordo com os objetivos da pesquisa.

**Análise de estilometria**

Os métodos estilométricos têm a capacidade de serem aplicados a diversas tarefas nas ciências humanas (Eder et al., 2017). Eles são úteis para agrupar documentos com base em suas características, identificar similaridades e diferenças entre textos individuais e conjuntos de documentos, além de rastrear a evolução temporal de documentos anotados.

A análise de estilometria é comumente utilizada nos estudos e pesquisas que objetivam à identificação de autoria e detecção de plágio.

Conforme mencionado por Eve (2022), um dos avanços mais importantes para os estudos de análise estilométrica foi o trabalho de Mosteller e Wallace (1963). Esse estudo se concentrou na autoria dos artigos Federalistas, escritos sob pseudônimos entre 1787 e 1788, com o objetivo de persuadir os cidadãos do estado de Nova Iorque, nos Estados Unidos da América, a ratificar a Constituição Americana.

Em seu trabalho, Mosteller e Wallace (1963) conduziram uma investigação sobre a distribuição de palavras (dentre elas: artigos, pronomes, conjunções e etc.) nesses artigos. Os pesquisadores chegaram a conclusões semelhantes às dos historiadores anteriores no que diz respeito à autoria dos artigos. No entanto, adotaram uma abordagem fundamentada em probabilidades estatísticas e análise bayesiana para sua análise.

O estudo realizado por Mosteller e Wallace (1963) demandou três anos para ser finalizado. No entanto, de acordo com Juola (2013), os computadores são capazes de executar esse tipo de análise em questão de segundos.

Portanto, é correto afirmar que a computação desempenha um papel crucial na implementação e estudo das análises estatísticas. Como resultado, diversos estudos nesse campo foram realizados, incluindo trabalhos de análise estilométrica para identificação de autoria. Esses trabalhos podem aproveitar as ferramentas e técnicas computacionais disponíveis para realizar análises mais eficientes e abrangentes.

Um dos casos mais repercutidos sobre a identificação de autoria, envolveu o trabalho de Juola (2013). Nesse caso, o autor relata ter recebido um e-mail de um repórter do *Sunday Times* de Londres descrevendo uma dica de que J. K. Rowling, autora da série de livros *Harry Potter*, possivelmente tinha escrito secretamente um romance policial intitulado *The Cuckoo's Calling* (traduzido como *O chamado do Cuco* no Brasil), publicado sob o pseudônimo de Robert Galbraith.

Em seu artigo, Juola (2013) descreve como o seu programa JGAAP (*Java Graphical Authorship Attibution Program*) foi capaz de identificar a autoria do livro ao quanti-

ficar o grau de similaridade entre os livros de J. K. Rowling e o livro publicado sob o pseudônimo de Robert Galbraith. Após a confirmação e divulgação da descoberta, a autora decidiu reconhecer a autoria do romance, gerando uma cobertura midiática mundial sobre a identificação de autoria, conforme ilustrado na Figura 2.3.

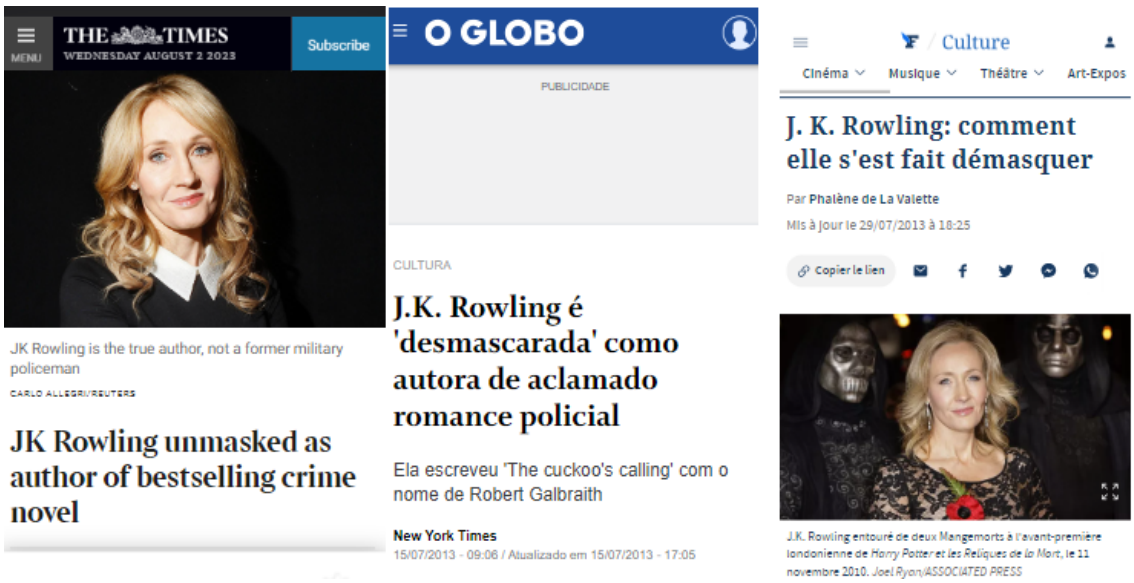


Figura 2.3: Jornais britânico, brasileiro e francês descrevendo como a autoria do livro *The Cuckoo's Calling* foi atribuída à J.K. Rowling

Por outro lado, Ramnial et al. (2016), descrevem a análise estilométrica como uma forma de análise estatística de textos escritos. Essa abordagem vai além da identificação de autoria, buscando identificar padrões literários ou características peculiares de linguagem a partir dos elementos presentes no texto analisado.

É importante ressaltar que os resultados de uma análise estilométrica são dependentes das características textuais a serem utilizadas. Isso ocorre porque a extração e associação de estilos textuais requer a identificação cuidadosa das características textuais a serem consideradas na análise.

Segundo Lagutina et al. (2019), em seus estudos sobre estilometria para identificação de autoria: "A escolha das características estilométricas do texto é a fase de estudo mais importante." Isso ocorre devido ao fato de que essa escolha influencia diretamente nos resultados obtidos.

Algumas das características utilizadas com mais frequência nos estudos sobre estilometria são o comprimento médio da sentença, frequência de conjunções, pronomes e preposições, frequência das letras, frequência das palavras e frequência dos sinais de pontuação.

Outras características, como as figuras de linguagem, também podem ser extraídas para a análise estilométrica. Exemplos dessas características incluem anáforas, anadiploses, diácopes, epanalepsis, epíforas, epizeuxis, polissíndetos e símploces. Essas

características podem ser relevantes em análises de poesias e poemas, dependendo do propósito do artigo ou trabalho.

A extração de características possibilita uma análise quantitativa de textos. Isso ocorre porque os dados extraídos podem ser quantificados com o intuito de representar a quantidade de identificações de figuras de linguagem encontradas em um texto ou em um trecho dele. Por exemplo: em um texto de 400 palavras, foram identificados 45 verbos e 15 anáforas.

Essa abordagem permite perceber que a utilização de diversas características pode impactar na elaboração de análises estilométricas mais sofisticadas.

**Reconhecimento de Entidades Nomeadas**

O Reconhecimento de Entidades Nomeadas é a tarefa do Processamento de Linguagem Natural responsável pela extração e identificação de nomes pessoais, nomes de organizações, nomes de locais, entre outros (Nadeau e Sekine, 2007).

Para realizar o Reconhecimento de Entidades Nomeadas, são aplicados modelos de Aprendizado de Máquina em diversos textos. Por sua vez, para realizar o treinamento supervisionado faz-se necessária a utilização da Engenharia de Características (*Feature Engineering*), destacada por Sharnagat (2014) como uma tarefa essencial do Reconhecimento de Entidades Nomeadas para todos os classificadores

A partir da aplicação do Reconhecimento de Entidades Nomeadas, é possível identificar e quantificar personagens, lugares e organizações, por exemplo. Dessa forma, assim como a extração de características estilométricas, o Reconhecimento de Entidades Nomeadas permite uma análise textual de forma quantitativa.

A extração de entidades nomeadas pode ser quantificada como características textuais, como por exemplo: o número de personagens e o número de personagens falantes, a contagem de menções do personagem, a contagem de verbos de fala, a contagem de menção ao personagem, bem como a frequência bruta dos personagens na narrativa.

Com isso, a utilização do Reconhecimento de Entidades Nomeadas em pesquisas de análise textual pode proporcionar uma compreensão mais profunda do texto em análise e introduzir novas formas de quantificação de textos. Entretanto, a sua extração pode ser imprecisa, visto a dificuldade computacional para identificação de entidades nomeadas. A ambiguidade das palavras (por exemplo, “rio” pode referir-se a um curso d’água, ao nome de uma cidade ou ao verbo rir), a variação na representação das entidades e a detecção de entidades desconhecidas ou raras que não constam no conjunto de treinamento dificultam a extração e identificação precisa das entidades nomeadas.

**Análise de sentimentos**

A análise de sentimentos é uma área de estudo computacional que busca compreender as opiniões, avaliações, atitudes e emoções das pessoas em relação a diversas entidades, indivíduos, problemas, eventos, tópicos e seus atributos (Liu e Zhang, 2012).

Esse tipo de análise utiliza técnicas de processamento de linguagem natural e aprendizado de máquina para identificar, extrair e classificar as expressões de sentimentos presentes em textos, como comentários em redes sociais, análises de produtos, *feedbacks* de clientes, entre outros.

A análise de sentimentos, quando aplicada a partir de algoritmos de aprendizado de máquina supervisionados, desenvolve modelos preditivos de sentimento usando dados anotados. Em tal abordagem, cada texto é representado por um vetor de características que quantifica aspectos específicos ou frequências de palavras. Os modelos de aprendizado de máquina são treinados e validados com textos de referência já anotados. Dentre esses modelos supervisionados destacam-se o *Naive Bayes* e o *Support Vector Machine* (SVM).

Entretanto, a utilização de algoritmos de aprendizado de máquina não é a única forma de extração de características textuais para análise de sentimentos. Segundo Soleymani et al. (2017), para realizar estimativas, a associação de uma palavra ou frase a um sentimento, emoção ou polaridade frequentemente é incorporada em um léxico. Esses léxicos podem ser criados por meio de anotação manual, solicitando a anotadores humanos que interpretem o significado das palavras. Isso permite a extração de características de diversas maneiras, dependendo do objetivo da análise e do texto a ser utilizado.

Para Pang et al. (2008), a análise de sentimentos pode ser aplicada em textos de mídias sociais para a extração de opiniões sobre eventos, figuras públicas ou produtos, usando a polaridade da palavra (positiva, negativa ou neutra).

Dessa forma, é possível quantificar e classificar as sentenças ou trechos de um determinado livro a partir da sua polaridade, identificando, entre outras coisas: trechos com maior carga emocional (considerando a emoção, positiva ou negativa, como uma única coisa), trechos com maior polaridade negativa, trechos com maior polaridade positiva, contagem de palavras positivas e negativas.

Além disso, existe a possibilidade de explorar mais do que os sentimentos positivos e negativos. Mohsin e Beltiukov (1905) descreve a utilização da roda das emoções de Plutchik (1982) para a inclusão de oito emoções, indo além dos sentimentos positivos e negativos em um texto.

Em sua Teoria das Emoções, Plutchik (1982) descreve que as emoções básicas do ser humano são: alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza. Além disso, por serem emoções básicas, elas podem derivar outras emoções. Dessa forma, Plutchik (1982) descreve a roda das emoções, que pode ser observada na Figura 2.4.

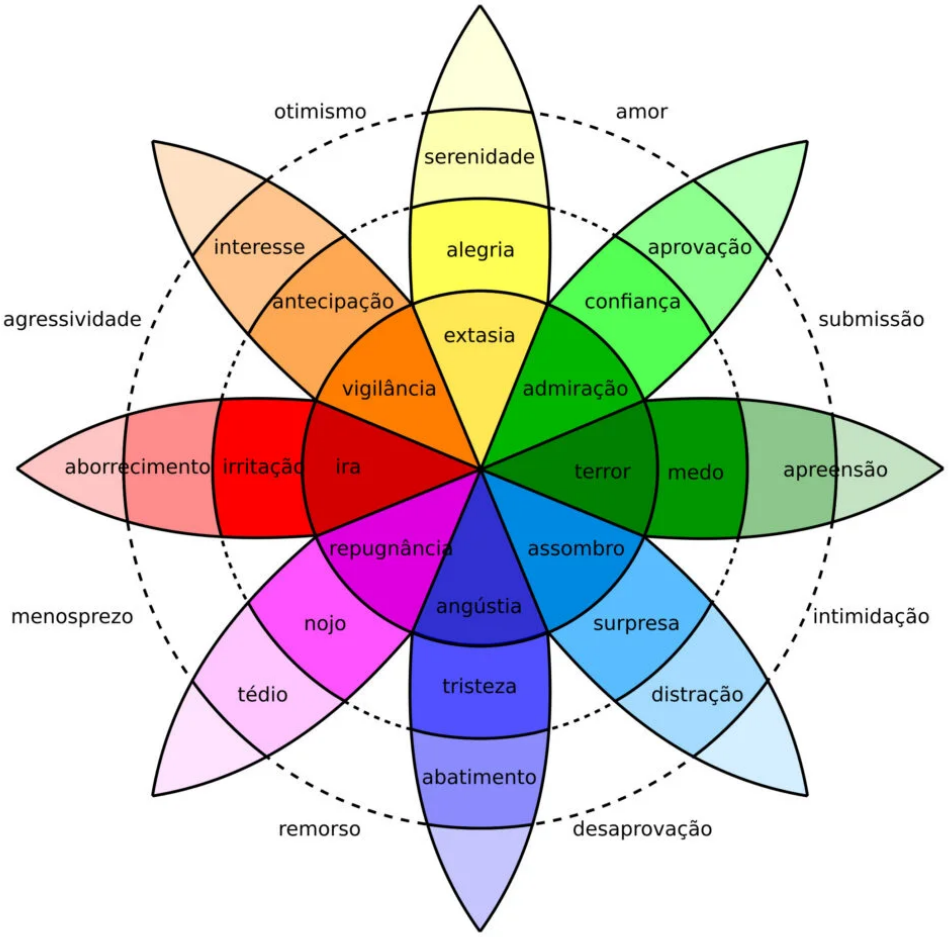


Figura 2.4: Roda das emoções, Plutchik (1982)

Essas emoções são utilizadas de forma complementar à polaridade da palavra, possibilitando análises textuais de maior profundidade. Afinal, em textos literários, é possível observar a existência de uma série de sentimentos associados ao enredo, às situações em que os personagens estão expostos ou mesmo a associações com o cotidiano do leitor. Portanto, a aplicação da análise de sentimentos em textos literários possibilita identificar diversas características associadas aos sentimentos e emoções presentes no texto.

Sendo assim, a análise de sentimentos também pode ser utilizada para a extração de características textuais, o que viabiliza análises mais detalhadas e a criação de novas medidas de quantificação.

### Modelagem de Tópicos

Abordagens probabilísticas, como a modelagem de tópicos, oferecem uma compreensão mais ampla e profunda de documentos nos quais essa abordagem é aplicada, sendo uma técnica elaborada para lidar com grandes volumes de dados textuais e pode ser utilizada para identificar padrões nos textos (Costa e Duarte, 2019).

Na modelagem de tópicos, cada tópico é representado por uma distribuição de palavras. Ao aplicar algoritmos de modelagem de tópicos, como o LDA (*Latent Dirichlet Allocation*), é possível identificar esses tópicos e suas respectivas distribuições em um determinado documento ou conjunto de documentos.

De acordo com Blei et al. (2003):

> "A alocação latente de Dirichlet (LDA) é um modelo probabilístico gerador de um corpus. A ideia básica é que os documentos sejam representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras."
>
> (Blei et al., 2003)

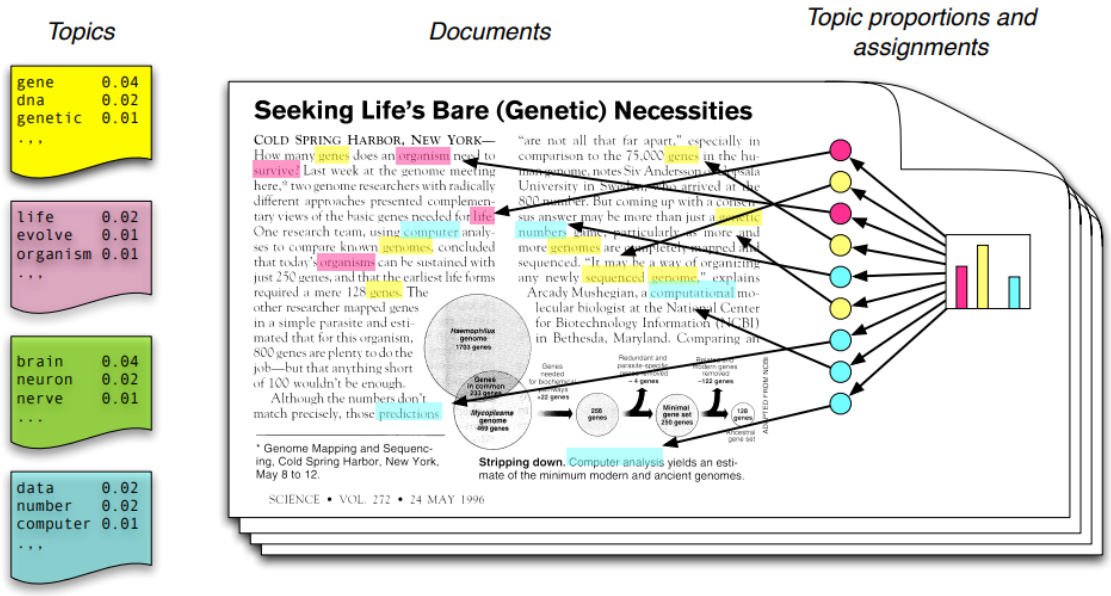


Figura 2.5: As intuições por trás do LDA, por Blei (2012)

A Figura 2.5, elaborada por Blei (2012), demonstra o funcionamento do LDA em um documento. Com ela, é possível observar os tópicos identificados no texto a partir de suas palavras. Além disso, é possível observar que as palavras relacionadas a análise de dados estão destacadas em azul, as palavras relacionadas a biologia estão destacadas em rosa e as palavras relacionadas a genética estão destacadas em amarelo. Esses destaques representam os tópicos identificados no texto em questão: análise de dados, biologia e genética.

Além disso, na Figura 2.5 é possível identificar a forma de extração e a ordem de processamento da modelagem de tópicos a partir do LDA. Inicialmente, é feita a seleção da distribuição de tópicos (representada pelo histograma à direita). Depois, para cada palavra do texto é feita uma atribuição de tópico (representada por círculos coloridos). Por fim, nos cartões coloridos à esquerda e nos destaques coloridos do texto é possível observar as palavras mais relevantes para cada tópico.

Outra forma representativa do modelo pode ser observada na Figura 2.6, também elaborada por Blei (2012). Nesse modelo, cada nó é uma variável aleatória, e é rotulado de acordo com sua função no processo. Os nós ocultos, que representam os tópicos, suas proporções e atribuições, estão na cor branca, enquanto os nós observados, que representam as palavras nos documentos, estão em cinza. No esquema, o retângulo N denota as palavras presentes nos documentos, e o retângulo D denota os próprios documentos utilizados para a extração dos tópicos.

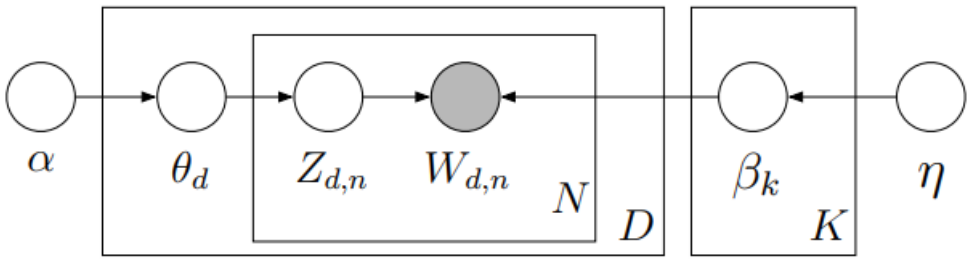


Figura 2.6: O modelo gráfico do LDA, por Blei (2012)

A partir da aplicação do LDA, várias análises podem ser conduzidas para extrair características relevantes dos textos. Algumas delas incluem:

- Identificação de tópicos predominantes: Isso permite determinar os tópicos mais relevantes abordados em uma coleção textual.
- Análise de coocorrência: Ao relacionar palavras-chave presentes nos tópicos identificados, pode-se compreender as associações e relações semânticas entre elas. Um exemplo é o trabalho de Shao e Qin (2014), que realiza a análise de coocorrência de palavras para examinar a correlação semântica dos temas presentes em um texto. Essa abordagem oferece uma medição aprimorada de similaridade de texto, combinando o modelo de tópico oculto com a análise de coocorrência de palavras.
- Agrupamento de documentos: Com a modelagem de tópicos, é possível agrupar documentos similares, o que facilita a organização e a categorização de grandes conjuntos de textos.

Com a modelagem de tópicos é possível revelar padrões ocultos nos textos, identificar o tópico do texto, analisar a variação da relevância dos tópicos ao longo do texto,

identificar palavras-chave e sua frequência no texto, além de identificar gírias. Essas identificações podem ser quantificadas para uma extração de características.

## 2.4 Análise de correlação

De acordo com Larson e Farber (2015), uma correlação é uma conexão entre duas variáveis. Essas duas variáveis podem ser representadas usando pares de valores ($x$, $y$), onde $x$ é a variável independente (também chamada de variável explanatória), e $y$ é a variável que pode ser influenciada por $x$ (também chamada de variável resposta).

Consequentemente, a criação de um gráfico de dispersão entre as variáveis $x$ e $y$ é uma maneira eficaz de entender a natureza da correlação. Nele, os pares ordenados são colocados em um plano cartesiano, permitindo a visualização da relação existente entre essas duas variáveis.

Dessa forma, um diagrama de dispersão retrata as variáveis em forma de pares ordenados ($x$, $y$) como pontos em um plano. A variável independente $x$ é representada no eixo horizontal, enquanto a variável dependente $y$ é mostrada no eixo vertical. A partir do diagrama, é possível analisar se há uma conexão entre as duas variáveis, ajudando a determinar se uma correlação está presente.

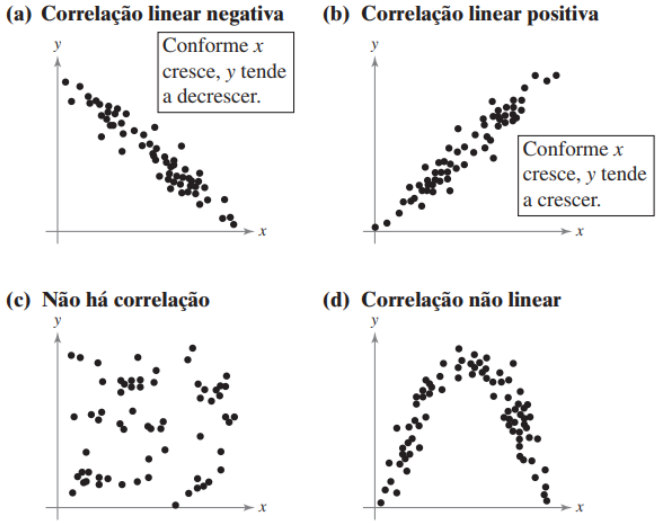


Figura 2.7: Indicativos sobre correlação em gráficos de dispersão (Larson e Farber, 2015)

A Figura 2.7 representa como as correlações lineares podem ser visualizadas em um gráfico de dispersão. Na Figura 2.7a, é possível observar uma correlação linear negativa, onde, quando a variável $x$ aumenta, $y$ tende a diminuir. Na Figura 2.7b, é possível observar uma correlação linear positiva, onde, quando a variável $x$ aumenta, $y$ tende a aumentar. Já na Figura 2.7c, não há uma correlação linear visível,

podendo-se concluir que não há relação entre as variáveis $x$ e $y$. Por fim, na Figura 2.7d, é possível observar uma correlação não linear, já que existe uma relação visível entre $x$ e $y$; entretanto, essa relação não pode ser traçada em uma linha reta.

A análise de correlação por meio de um diagrama de dispersão pode ser suscetível a interpretações subjetivas e enviesadas. Para obter uma avaliação mais precisa da direção e intensidade de uma correlação linear entre duas variáveis, é apropriado calcular o coeficiente de correlação. Esse coeficiente fornece uma abordagem objetiva para compreender a conexão entre as variáveis, representando tanto a intensidade quanto a direção de uma relação linear entre elas. Isso ajuda a evitar possíveis interpretações subjetivas e fornece uma medida numérica mais confiável da correlação.

Para calcular o coeficiente de uma correlação linear é possível utilizar a correlação de Pearson:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{2.1}$$

A Equação 2.1 descreve o cálculo da correlação de Pearson, representada como a variável $r$, entre as variáveis $x$ e $y$. Nessa equação, $n$ representa a quantidade de pares de dados analisados.

Dessa forma, o coeficiente de correlação varia entre -1 e 1, abrangendo todos os valores intermediários. Quando as variáveis $x$ e $y$ apresentam uma forte correlação linear positiva, o valor aproxima-se de 1. Em casos de uma correlação linear negativa, ele tende a se aproximar de -1. Quando $x$ e $y$ demonstram uma relação linear perfeitamente positiva ou negativa, o coeficiente assume os valores 1 ou -1, respectivamente. Na ausência de uma correlação linear, o valor se aproxima de 0.

Entretanto, vale ressaltar que um valor de coeficiente de correlação próximo ou igual a 0 não necessariamente implica na ausência de relação entre as variáveis, apenas indica a inexistência de uma relação linear.

A Figura 2.8 representa como as correlações lineares podem ser visualizadas em um gráfico de dispersão, desta vez com adição do coeficiente de correlação de Pearson calculado.

Na Figura 2.8a, é possível observar uma correlação linear positiva perfeita, com $r$=1, significando que existe uma relação entre $x$, representando o número de ingressos comprados, e $y$, representando o custo total da compra.

Já na Figura 2.8b, é possível observar uma correlação linear positiva forte, com $r$=0.81, onde, quando a altura de um indivíduo (variável $x$) aumenta, o tamanho do seu calçado (variável $y$) tende a aumentar. No entanto, essa correlação não possui uma linearidade tão visível quanto na Figura 2.8a, indicando que nem sempre a altura afeta o tamanho do calçado.

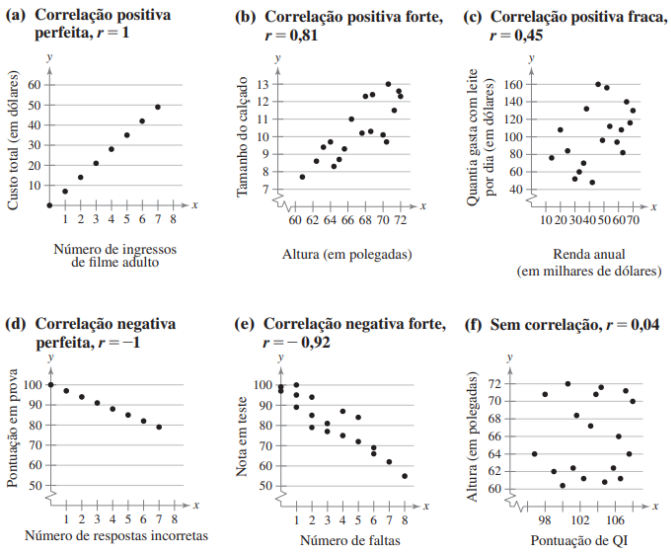


Figura 2.8: Indicativos sobre correlação e seus respectivos valores de $r$ em gráficos de dispersão (Larson e Farber, 2015)

Na Figura 2.8c, há uma correlação linear fraca, com $r$=0.45, indicando que em certas ocasiões a variável $x$ influencia a variável $y$.

Na Figura 2.8d, é possível observar uma correlação linear negativa perfeita, com $r$=-1, indicando que quanto maior a quantidade de respostas erradas ($x$), menor a nota ($y$).

A Figura 2.8e, indica uma correlação linear negativa forte, com $r$=-0.92, indicando que quando a quantidade de faltas de um indivíduo (variável $x$) aumenta, a sua nota (variável $y$) tende a diminuir, porém não possuindo uma linearidade tão visível quanto na Figura 2.8d, indicando que nem sempre a quantidade de faltas afeta na nota.

Por fim, a Figura 2.8f retrata que não há uma correlação entre as variáveis $x$ (pontuação de QI de um indivíduo) e $y$ (altura de um indivíduo), indicando que essas variáveis são independentes, uma não impacta a outra, já que o valor do seu coeficiente de correlação é próximo a 0 ($r$=0.04).

Para cálculos de correlação não lineares, outras fórmulas além de Pearson podem ser utilizadas, a exemplo da correlação de Spearman.

O coeficiente de correlação de Spearman mensura a intensidade de uma relação monotônica. Isso requer que os dados estejam relacionados de forma monotônica. Em essência, isso significa que quando uma variável aumenta, a outra variável também aumenta; da mesma forma, quando essa variável diminui, a outra também diminui.

Ao contrário de uma relação linear em que as variáveis mudam na mesma direção com uma taxa constante, uma relação monotônica implica que as variáveis se movem na mesma direção, porém, não necessariamente mudam sua constância.

O coeficiente de correlação de Spearman é calculado considerando a posição de cada elemento nos pares de dados a serem analisados, ordenados em ordem decrescente. Sua fórmula de cálculo é a seguinte:

$$r = 1 - \frac{6 \sum (d^2)}{n^3 - n} \tag{2.2}$$

A Equação 2.2 descreve o cálculo da correlação de Spearman, representada como a variável $r$. Nessa equação, $n$ representa a quantidade de pares de dados analisados e $d$ representa a diferença entre as posições dos valores do par.

As correlações de Pearson e Spearman mostram apenas um valor para a correlação entre as variáveis, sendo esse valor obtido a partir da correlação ao longo de toda a sequência de dados.

Entretanto, em alguns casos, faz-se necessária a visualização da variação da correlação ao longo de toda a série temporal das variáveis a serem analisadas. Dessa forma, pode-se utilizar uma análise de correlação em série temporal utilizando janelas deslizantes.

Com isso, é possível observar se em correlações detectadas como fortes a partir dos coeficientes de Pearson e Spearman, o valor de seu coeficiente é alto ao longo de toda a série temporal ou se existe alguma variação de valor em determinado momento. O mesmo princípio é válido para correlações fracas ou inexistentes, já que podem haver momentos na série temporal onde o valor de $x$ influencia fortemente o de $y$.

Entretanto, a presença de uma correlação forte entre duas variáveis não automaticamente indica uma relação de causa e efeito entre elas, sendo essencial uma análise mais aprofundada para estabelecer se essa relação de causa e efeito realmente existe entre as variáveis em questão.

## 2.5 Trabalhos Relacionados

Nesta seção, serão abordados trabalhos relacionados ao tema de extração de padrões e características textuais em diferentes níveis linguísticos. Inicialmente, quatro trabalhos serão apresentados, os quais estão relacionados à identificação de padrões estilométricos em textos (Eder et al., 2017; Dell'Orletta et al., 2013; Wanner et al., 2017; Lagutina et al., 2020). Essa abordagem é relevante para esta dissertação, já que a detecção de padrões estilométricos pode ser realizada por meio de análises textuais em diversos níveis linguísticos, demonstrando a extração de características com relações interníveis. Além disso, será discutido o artigo de Waumans et al. (2015), sobre a extração de redes sociais em textos literários e os padrões estilométricos que podem ser extraídos a partir delas utilizando o Reconhecimento de Entidades Nomeadas. Em seguida, serão abordados dois trabalhos sobre a Análise de Sentimentos (Jacobs, 2019; Min e Park, 2019), que utilizam diferentes técnicas de identificação

de padrões emotivos nos textos. Por fim, dois trabalhos que aplicam técnicas de extração de características em textos da prosa literária brasileira serão explorados (Carvalho et al., 2020; Lima et al., 2021). Um desses trabalhos se concentra na identificação de estruturas métricas, enquanto o outro foca na extração de repetições de sílabas fonéticas. Ambos os trabalhos trazem relevância a esta dissertação, uma vez que esta propõe um método computacional capaz de identificar paralelismos na prosa literária brasileira em diferentes níveis, possibilitando estabelecer conexões com esses estudos e outras publicações já existentes.

No artigo de Eder et al. (2017), os autores utilizaram uma análise em diferentes níveis linguísticos (morfológico, sintático e semântico) para encontrar, classificar ou agrupar padrões estilométricos em textos literários. Esses padrões estilométricos foram usados para inferir a autoria dos textos e posteriormente agrupá-los de acordo com a autoria encontrada.

Para isso, os autores propuseram o *WebSty*, um sistema não supervisionado para atribuição de autoria, onde os documentos de entrada são agrupados automaticamente em conjuntos que devem incluir textos do mesmo autor.

Algumas das características extraídas para o estudo de Eder et al. (2017) foram:

- Morfológicas: comprimento de parágrafos e frases, frequência de palavras e sinais de pontuação;
- Gramaticais: lemas e *Part of speech tagging* (PoS *tagging*) (agrupamento das suas classes gramaticais);
- Semânticas: classes de nomes próprios, significados lexicais, domínios temáticos.

Os dados foram divididos em conjuntos de treino e teste e aplicados em técnicas de aprendizado de máquina como *Naive Bayes* e *Support Vector Machine* (SVM), utilizando uma validação cruzada 10-*fold cross-validation* para avaliar os resultados. Por fim, Eder et al. (2017) relataram que o *WebSty* obteve sucesso na identificação de estilometria e agrupamento de textos de acordo com a autoria.

Além disso, Eder et al. (2017) apresentaram que a utilização de diferentes níveis linguísticos desempenhou um papel crucial no resultado final da ferramenta. Isso evidencia que a utilização de diferentes níveis linguísticos em problemas de análise textual e reconhecimento de padrões pode contribuir para uma análise mais profunda e um desempenho aprimorado. Ademais, esses padrões estilométricos podem estar relacionados a outros padrões já identificados, e é responsabilidade do método apresentado nesta dissertação explorar a tentativa de descobrir essas possíveis relações.

No artigo de Dell'Orletta et al. (2013), os autores conduziram um estudo de caso com o propósito de identificar os traços distintivos presentes em textos da prosa italiana. Eles examinaram características linguísticas que poderiam caracterizar esses textos e as testaram em duas tarefas de classificação: classificação por gênero e classificação

por legibilidade. As características foram extraídas de quatro categorias principais, baseadas em diferentes níveis de análise linguística. A lista abaixo demonstra alguma dessas características e as quatro categorias utilizadas por Dell'Orletta et al. (2013):

- Características de Texto Bruto: Comprimento de sentença e palavras;
- Características Lexicais: Frequência das palavras, *Type/Token Ratio* (TTR) que mede a variedade lexical em um texto;
- Características Morfossintáticas: Distribuição de *Part-Of-Speech* (PoS), densidade léxica, modo, tempo e pessoa verbal;
- Características Sintáticas: Distribuição de tipos de dependência (sujeito, objeto direto), Predicados verbais, Subordinações.

A análise morfossintática utilizada por Dell'Orletta et al. (2013) foi realizada com a marcação de PoS e análise de dependência através do *DeSR parser* (Attardi, 2006), utilizando também um algoritmo de *Support Vector Machines* (SVM). Os resultados apontaram que as características identificadas podem ser confiavelmente utilizadas na classificação de textos por gênero e nível de legibilidade.

Em sua pesquisa, Wanner et al. (2017) argumentam que a maioria das abordagens para identificação de autoria textual está principalmente focada em características lexicais. Nesse contexto, os autores visam mostrar que a incorporação de dependência sintática e características de discurso pode reduzir o número total de características necessárias, ao mesmo tempo em que mantém um desempenho competitivo usando técnicas de classificação padrão. Isso destaca a importância de considerar diferentes aspectos linguísticos na análise de autoria textual.

Para isso, Wanner et al. (2017) utilizaram o *Natural Language Toolkit* (NLTK) (Bird et al., 2009) para análise de dependência e análise de discurso, extraindo cerca de 180 características, sendo algumas delas:

- Características baseadas em caracteres: índices de pontuações, dígitos, sinais;
- Características baseadas em palavras: número de letras por palavra, riqueza de palavras;
- Características baseadas em sentenças: número de palavras por sentença, desvio padrão de palavras por sentença;
- Características baseadas em dicionários léxicos: índices de marcadores de recurso, interjeições, abreviações, polaridade das palavras (palavras positivas e negativas de acordo com o sentimento associado);
- Características sintáticas: características de *Part-of-Speech*(PoS), características de dependência e características a partir de grafos.

Como resultados, Wanner et al. (2017) apresentam que a dependência sintática e as características do discurso desempenham um papel significativo na tarefa de

identificação de gênero e autor, verificação de autoria e detecção de plágio. Dessa forma, algumas dessas características são utilizadas no método computacional aqui proposto.

Já no artigo de Lagutina et al. (2020), os autores comparam distintos tipos de características estilométricas: características de nível mais básico (que englobam aquelas baseadas em caracteres e em palavras) e características de nível mais avançado, baseadas no ritmo.

Para essa pesquisa, Lagutina et al. (2020) escolheram as seguintes características:

- Características fundamentadas em caracteres: comprimento médio da sentenças em letras e pontuações, frequências de letras e pontuações;
- Características fundamentadas em palavras: comprimento médio da sentenças em palavras, comprimento médio da palavra em letras;
- Características fundamentadas em ritmos: ocorrência de figuras de linguagem (anáfora, anadiplose, diácope, epanalepsis, epífora, epizeuxis, polissíndeto e símploce), fração de palavras únicas e fração de *Part-of-Speech*(PoS).

Essas características estilométricas foram submetidas a quatro algoritmos de classificação para fins de comparação: o *AdaBoost*, o *Random Forest*, o *Bidirectional Long Short Term Memory* (biLSTM) e o *Gated Recurrent Unit* (GRU).

Os experimentos realizados com os algoritmos mencionados revelaram que a utilização de todos os tipos de características resultou em melhorias significativas em comparação com a utilização de apenas um tipo e, na maioria das situações, com dois tipos de características.

Dessa forma, o artigo de Lagutina et al. (2020) é mais uma referência para o trabalho de análise textual computacional aqui proposto, visto que aborda características de diferentes níveis e demonstra a importância da incorporação dessas características em análises inter-níveis..

Waumans et al. (2015) definiram a autoria da história de um romance através da análise topológica da rede social de personagens. Para isso, eles desenvolveram uma ferramenta automatizada que analisa os diálogos nos romances, identifica os personagens e calcula as relações deles ao longo da história. Isso permite avaliar a evolução da rede durante o desenvolvimento do romance.

Waumans et al. (2015) propuseram uma análise de rede social que indicou semelhanças na forma como as redes sociais são apresentadas nos romances e como os sociólogos estudam as redes sociais reais. Além disso, esta análise de rede forneceu uma série de características relacionadas à história do livro e ao estilo do autor. Entre essas características, destacam-se:

- Frequência de espaçamento de diálogo: indica a presença relativa de narração nas conversas

- Taxa de identificação do locutor: indica o quanto o autor lembra seus leitores da identidade dos falantes

No entanto, os padrões estilométricos propostos por Waumans et al. (2015) não foram avaliados em conjunto com a evolução das redes sociais ao longo da história, nem com outros padrões textuais. Essa lacuna pode ser suprida pelo método computacional proposto nesta dissertação, visto que ele busca identificar a correlação entre características de diferentes níveis textuais de forma quantitativa.

Apesar dos estudos na área de análise estilométrica trazerem luz a esta pesquisa, esse tipo de análise não é o único que precisa ser considerado ao objetivar a extração de características de diferentes níveis para posterior análise.

Jacobs (2019) teve como propósito introduzir e avaliar a ferramenta chamada *SentiArt*, destinada a calcular a carga emocional e o perfil de personalidade de personagens em obras literárias. Essa ferramenta foi empregada em dois estudos.

O primeiro estudo teve como propósito verificar a viabilidade do *SentiArt* em pesquisas de poética computacional abrangendo múltiplos idiomas. Para atingir tal objetivo, foram empregados classificadores de Aprendizado de Máquina e Redes Neurais Artificiais: *Adaboost* em conjunto com Árvores de Decisão, *k-Nearest Neighbors* (kNN), Regressão Logística, *Naïve Bayes* e *MultiLayerPerceptron* (MLP). Esses classificadores foram utilizados para prever categorias de sentimentos com base em entradas obtidas de diferentes ferramentas de análise de sentimentos computacionais. Os resultados desse primeiro estudo indicaram que o *SentiArt*, quando combinado com regressão logística, demonstrou um desempenho superior em comparação com outras ferramentas de análise de sentimentos computacionais.

O segundo estudo abordou o cálculo de perfis de figuras emocionais e perfis de personalidade para os personagens presentes em textos literários. Os resultados obtidos indicaram que o *SentiArt* gerou previsões coerentes em relação aos perfis emocionais e de personalidade dos personagens fictícios. Além disso, a ferramenta alcançou uma notável precisão em termos de validação cruzada ao categorizar os personagens nas seguintes classes: amigável, hostil, bom, mau, caridoso e desalmado.

Ainda que não tenha explorado a aplicação de diversos níveis linguísticos, Jacobs (2019) evidencia que ao empregar a Análise de Sentimentos, é viável extrair atributos textuais adicionais, além das características convencionais frequentemente identificadas em várias pesquisas.

Min e Park (2019) apresentam uma estrutura fundamentada em rede visando modelar uma narrativa, com ênfase nas personagens e suas interações. Para testar o sistema proposto, os autores utilizaram uma versão em inglês do livro "Os Miseráveis".

Quanto à metodologia, Min e Park (2019) optaram por representar a narrativa como um conjunto de linhas temporais de personagens em interação. Esse enfoque permitiu a construção de uma rede gradual de personagens conforme a trama avança.

Para quantificar os sentimentos, os autores adotaram o LIWC (*Linguistic Inquiry and Word Count*), um software que fornece valores distintos para sentimentos positivos e negativos, combinados em uma única métrica denominada SPI (Índice de Polaridade de Sentimento).

Através da sua implementação, Min e Park (2019) identificaram que diversos personagens, especialmente os protagonistas, apresentaram variações notáveis nos sentimentos ao longo da narrativa e se mostraram agentes impactantes no tom emocional da história, exercendo influência nas flutuações emocionais de outros personagens na trama.

O artigo de Min e Park (2019) realiza uma análise de sentimentos focada nos personagens e estabelece uma conexão entre esses sentimentos e o tom da narrativa, evidenciando que a identificação de personagens e seus sentimentos podem ser utilizados em análises textuais de diferentes fins.

Até aqui, os trabalhos apresentados nessa seção abordaram a extração de características para textos em diversos idiomas, entretanto, existem trabalhos focados na extração de características em textos da literatura brasileira.

Carvalho et al. (2020) apresentaram o *MIVES* (*Mining Verse Structure*), um sistema elaborado para a identificação e classificação das estruturas métricas de versificação na prosa literária de língua portuguesa.

O funcionamento do sistema ocorre em quatro etapas:

- (i) Extração de sentenças do texto: A partir de um texto digital os sistema segmenta as sentenças, identificando-as a partir marcadores chamados pontos-finais da sentença.
- (ii) Separação silábica e marcação de sílabas tônicas: Após a segmentação das sentenças, elas passam por uma separação silábica e marcação de sílabas tônicas.
- (iii) Escansão e identificação das estruturas métricas de versificação: Nessa etapa são identificados fenômenos intra-vocabulares (crase, sinérese, diérese) e inter-vocabulares (crase, sinalefa, elisão).
- (iv) Visualização e análise de resultados: Os autores utilizaram uma interface gráfica com técnicas de leitura aproximada (*close reading*) para destacar, no texto original, as sentenças métricas encontradas.

Ao submeter um livro ao sistema, é possível obter os seguintes resultados: identificação das estruturas, frequência absoluta e relativa de ocorrência, valores de distância entre as sentenças, desvio padrão e distribuição de frequência absoluta das estruturas, além de um arquivo externo contendo as estruturas encontradas e sua posição no texto.

No artigo, Carvalho et al. (2020) apresentam a extração de padrões literários na prosa através da análise de fenômenos linguísticos da poética. Contudo, o foco

reside exclusivamente na extração de características métricas do texto, não havendo tentativa de relacionar essa identificação com outros padrões linguísticos.

Já Lima et al. (2021) introduziram um sistema computacional voltado para a identificação de paralelismos de natureza fonológica em prosa literária. Esse sistema se baseia na detecção de elementos poéticos, como aliteração, assonância, consonância e rima.

A primeira etapa na extração dessas características foi a fase de pré-processamento textual, na qual Lima et al. (2021) realizaram a *tokenização* das palavras do texto, seguida pela aplicação de separadores silábicos e pela conversão de grafemas em fonemas. Em seguida, a versão fonética do texto foi submetida a algoritmos a fim de derivar as sequências de sílabas de interesse.

Em sequência, o texto foi submetido a dois algoritmos. O primeiro deles empregou uma janela deslizante com o objetivo de detectar pares de sílabas repetidas. O segundo algoritmo fez uso dos pares de sílabas identificados para construir sequências completas de texto. Como resultado final, foi produzida uma lista contendo sequências de sílabas fonéticas repetidas, acompanhada da indicação da posição de cada ocorrência, permitindo ao usuário visualizar o texto de maneira específica. Finalmente, os autores ressaltam que a análise computacional identificou uma frequência considerável desses fenômenos fonológicos na prosa literária.

Semelhante ao trabalho de Carvalho et al. (2020), o artigo de Lima et al. (2021) demonstra a detecção de padrões literários na prosa de língua portuguesa através da análise de fenômenos linguísticos relacionados à poética. No entanto, os autores se concentram exclusivamente em fenômenos fonológicos e métricos. Nesse contexto, o método computacional proposto neste estudo busca a identificação de relações entre esses padrões previamente encontrados.

Os estudos apresentados nesta seção são cruciais para entender a importância de uma pesquisa que busque associar as características textuais de diferentes níveis linguísticos. Com efeito, esses estudos demonstram o uso de características provenientes de níveis variados para tarefas como autoria, estilo textual, padrões emocionais, métricos e fonológicos. Contudo, é válido notar que os autores não exploram a ligação entre níveis distintos ou entre diferentes características em suas análises.

Dessa forma, os estudos relacionados se conectam diretamente com a pesquisa que esta dissertação apresenta, já que este trabalho busca analisar, associar e correlacionar as características extraídas em diferentes níveis linguísticos, seguindo as abordagens discutidas nesses estudos correlatos para a sua fase inicial de extração de características.

# Capítulo 3

# Metodologia

Nos textos literários, a linguagem desempenha um papel que vai além da mera comunicação. Ela se torna uma forma de expressão artística, onde dispositivos poéticos podem ser empregados para aprofundar a experiência do leitor. A função poética da linguagem pode ser notada uma vez que os escritores exploram, através desses dispositivos poéticos, o significado das palavras, a sonoridade do texto, seu ritmo e sua estrutura. Nesse contexto, os paralelismos surgem através dos retornos recorrentes (repetições) desses dispositivos poéticos.

Por sua vez, os paralelismos podem estar presentes em distintos níveis de linguagem e também em padrões de correlação entre os níveis, onde as repetições, simetrias e assimetrias se manifestam.

A identificação de um paralelismo textual internível envolve uma avaliação minuciosa do texto, que visa extrair padrões em cada um dos níveis linguísticos envolvidos para que possam ser identificadas as relações entre eles. A introdução de métodos computacionais com a capacidade de quantificar e correlacionar características extraídas de diferentes níveis linguísticos possibilita evidenciar novos de padrões de paralelismo.

A correlação entre características de diferentes níveis pode ser vista como um paralelismo uma vez que o paralelismo pode implicar em repetição e criação de estruturas paralelas ao longo do texto. Essa ordem na inter-estruturação de diferentes níveis pode se manifestar em padrões identificáveis durante a análise textual. Por exemplo, ao analisar um trecho da obra com uma alta quantidade de emoções negativas e de entidades nomeadas, é possível identificar uma relação entre determinados personagens e emoções negativas, resultando em correlações entre essas variáveis de análise textual.

Sendo assim, o paralelismo em textos literários pode levar a correlações entre variáveis de análise textual devido à consistência e repetição de certos padrões linguísticos. Essa relação reflete a forma como os autores constroem e organizam seus textos para transmitir significados e efeitos estilísticos específicos.

Propomos, portanto, um método computacional para a identificação de paralelismos interníveis com base na análise de correlação entre as características extraídas de diferentes níveis e visamos oferecer uma abordagem quantitativa para a identificação desses padrões interdimensionais a partir da quantificação dessas correlações.

Os trabalhos relacionados apresentados oferecem uma variedade de características textuais em diferentes níveis, permitindo a quantificação da presença ou ocorrência de elementos como sílabas fonéticas, entidades nomeadas, polaridades emocionais, padrões métricos, entre outros.

Definimos operacionalmente que ocorre um paralelismo internível quando há uma forte correlação entre valores de duas características extraídas. Isso pode ocorrer quando altos valores de uma característica co-ocorrem com altos valores de outra característica, considerando todo o texto ou um trecho do texto. Além disso, o paralelismo internível também pode ocorrer em casos de correlações negativas fortes, onde altos valores de uma característica estão relacionados a baixos valores da outra.

O fluxo do método computacional proposto, desde a entrada do texto até a detecção das correlações entre suas características e a subsequente visualização dos resultados, pode ser observado na Figura 3.1.

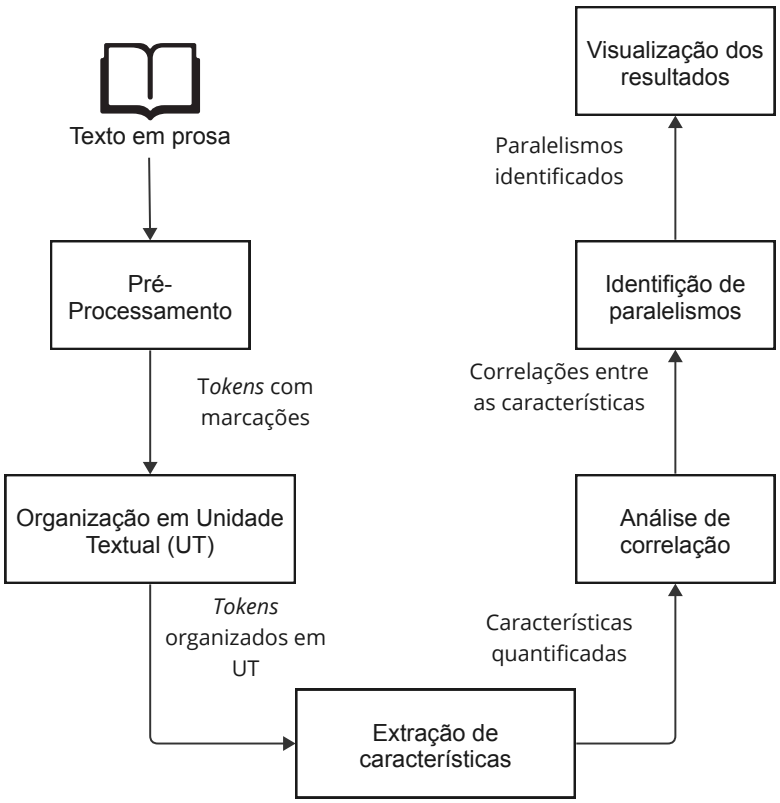


Figura 3.1: Fluxograma do funcionamento do método

Com a entrada de um texto, é iniciado o seu pré-processamento com a *tokenização*

das palavras e pontuações.

Em seguida, é realizada a marcação desses *tokens* a partir de sua classe gramatical, entidade nomeada, lema, polaridade e emoção. Posteriormente, esses *tokens* são organizados em Unidades Textuais (UTs), sendo cada UT, um trecho do texto a ser analisado de maneira individualizada. Essa Unidade Textual pode variar em tamanho, cobrindo desde um trecho curto, como uma sentença ou parágrafo, até uma parte maior do texto, como um capítulo ou seção. A escolha do tamanho da UT depende dos objetivos da análise e das características específicas do estudo.

Com uma UT de tamanho 1, por exemplo, o livro é dividido em partes, cada uma contendo 1 sentença. No caso de uma UT de tamanho 100, o livro é dividido em partes, cada uma contendo 100 sentenças.

A extração de características constitui a base deste método, pois visa capturar diversas propriedades presentes nos textos literários. Essa extração é realizada transformando o conteúdo textual, composto por sequências de palavras e sentenças, em uma série de valores quantitativos descritivos do texto a partir de suas UTs.

A cada UT é associado um valor para cada característica, representando sua quantificação na forma de frequência absoluta ou relativa e no caso da modelagem de tópicos a relevância (em porcentagem) de cada um dos tópicos para a UT. Essas características quantitativas oferecem uma representação numérica das propriedades textuais, possibilitando uma análise comparativa no texto completo ou em diferentes trechos do texto.

Nesse processo foram extraídas:

- Características estilométricas: extraídas com o objetivo de proporcionar uma visão sobre a linguagem e o estilo presentes nos textos literários.
- Análise de sentimentos: aplicada para avaliar o tom emocional presente nos textos literários.
- Características poéticas: permitem a exploração de traços poéticos nos textos literários.
- Modelagem de tópicos: para extração e identificação dos tópicos presentes nos textos literários.

A etapa de análise de correlação avalia a relação entre as características textuais de diferentes níveis dentro de um texto. Cada nível se refere a um conjunto de características textuais que foram selecionadas e extraídas para análise. Ao final dessa etapa são quantificadas todas as correlações entre características.

Os valores das correlações são utilizados na etapa de identificação de paralelismos textuais com o objetivo de verificar a ocorrência de paralelismos entre as características no texto completo ou em diferentes trechos do texto. Esses paralelismos são identificados automaticamente com base em um limiar de correlação, um valor mínimo de associação entre características obtidas para o texto.

Esta identificação de paralelismos pode ser realizada de forma global (texto completo) ou segmentada (trecho do texto). Com o intuito de identificar paralelismos em uma obra inteira, a identificação é realizada de maneira global, considerando a série completa de valores das características extraídas para todo o livro. Já a análise segmentada proporciona a identificação de paralelismos em trechos textuais e utiliza uma quantidade delimitada de UTs para o cálculo de correlação.

Por fim, as correlações, as características quantificadas e os paralelismos são exibidos a partir de tabelas e gráficos de dispersão e de linhas.

## 3.1 Extração de características

Este estudo aborda a identificação de paralelismos textuais entre diferentes níveis linguísticos com características variadas, como a frequência de palavras, a análise de sentimentos e a modelagem de tópicos, a partir de medidas de correlação. Para esta análise quantitativa, a extração de características serve como um meio de traduzir elementos textuais em dados mensuráveis, possibilitando uma análise computacional dos textos para identificar padrões textuais em diferentes níveis de linguagem.

Objetivando realizar um levantamento das características textuais a serem extraídas para realizar a identificação dos paralelismos, foi feita uma busca por trabalhos relacionados à extração de características textuais para finalidades diversas. Estes trabalhos foram buscados nas bases ACM (Association for Computing Machinery), ACL (Association for Computational Linguistics), IEEE (Institute of Electrical and Electronics Engineers) e Google Scholar resultando na seleção de artigos das áreas de linguística computacional e processamento de linguagem natural priorizando publicações que empregaram textos literários em suas pesquisas e que exploraram múltiplos níveis textuais para a extração de características.

A realização desse levantamento de características possibilita uma análise das características em vários níveis textuais, permitindo a identificação das mais frequentemente encontradas em estudos de análise textual:

- Contagens de palavras: extraídas a partir da quantidade de palavras únicas no texto, da quantidade de lemas únicos no texto, das frequências relativa e absoluta das palavras no texto.
- Distribuição de classes gramaticais (do inglês: *Distribution of Part-Of-Speech* ou *PoS Distribution*): extraídas utilizando técnicas de *PoS Tagging* e podem ser definidas como a classe gramatical de cada palavra/*token*. Aqui, as medidas quantitativas referem-se à contagem de *tokens* em cada uma das classes gramaticais e à frequência relativa das classes gramaticais em relação ao número de palavras/*tokens*.
- TTR (*Type-Token Ratio* ou Relação Tipo-*Token*): trata-se da divisão entre a quantidade de palavras distintas e a quantidade total de palavras. Essa relação está associada à riqueza lexical de um determinado trecho ou texto.

As características desse levantamento podem ser observadas em trabalhos relacionados a essa pesquisa, disponíveis na Seção 2.5. A Tabela A.1 no Anexo A apresenta o levantamento de 118 características organizadas em 21 grupos, o nível de cada grupo e os artigos que utilizam ao menos uma das características do grupo. A partir da tabela é possível identificar as principais características dos artigos e trabalhos relacionados.

Apoiado aos artigos selecionados, inicia-se a fase de seleção de características, suas utilizações e suas formas de extração, onde foram encontrados sistemas e ferramentas que são aplicados para a extração de características textuais em diversos níveis:

- *NLPyPort* (Ferreira et al., 2019)- um *pipeline* de PLN, em Python, voltado para a língua portuguesa e desenvolvido a partir de recursos já existentes, como o NLTK (Bird et al., 2009), e possibilita tarefas de *tokenização*, marcação de PoS, lematização e reconhecimento de entidades nomeadas;
- Os dicionários léxicos para extração da polaridade e emoções dos *tokens* de palavras:
  - LIWC (*Linguistic Inquiry and Word Count*) (Pennebaker et al., 2001)- realiza o agrupamento de palavras em categorias, o que possibilita a análise de características em textos, incluindo a análise de sentimentos, traduzido e adaptado para português em (Balage Filho et al., 2013);
  - *OpLexicon* (*OpinionLexicon*) (Souza e Vieira, 2011)- construído para tarefas de análise de sentimentos, onde agrupa as palavras em português de acordo com sua polaridade (positiva ou negativa);
  - *SentiLex-PT 02* (Silva et al., 2012)- construído para aplicações de mineração de opinião em português e possibilita a detecção e classificação de sentimentos e opiniões direcionados a entidades humanas.
  - *NrcEmolex* (Mohammad e Turney, 2013)- um léxico de palavras associadas a oito emoções básicas: alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza.
- *MIVES* (Carvalho et al., 2020)- um sistema que realiza a escansão para a identificação computacional de estruturas métricas de versificação em prosa de língua portuguesa.
- *ALLPRO* (Lima et al., 2021)- um método computacional de análise textual para identificação de paralelismos fonéticos, especificamente sequências de repetições de sílabas fonéticas.
- Gensim (Řehůřek e Sojka, 2010)- uma biblioteca desenvolvida para o processamento e extração de tópicos em textos brutos. Utiliza algoritmos de aprendizado de máquina não supervisionados.

As ferramentas e sistemas acima, foram selecionadas pela possibilidade de extração de características em textos escritos em português. Além disso, a utilização dessas ferramentas possibilita a extração de características em vários níveis linguísticos.

### 3.1.1 Características estilométricas

Alguns dos artigos encontrados utilizam da extração de características estilométricas para diversos tipos de análise textual. Nesse contexto, uma das etapas presentes neste trabalho envolve a extração dessas características estilométricas para identificar os paralelismos nos textos literários da prosa brasileira.

Para isso, o presente método computacional segue a proposta do trabalho de Ferreira et al. (2019) para extrair as características estilométricas a partir de suas classes gramaticais (artigos, substantivos, pronomes, verbos e etc.) e entidades nomeadas (pessoas, lugares, organizações e etc.) além da extração do lema da palavra.

Para essa tarefa é realizada a *tokenização*, que se refere à divisão de um texto em unidades menores chamadas *tokens*, nesta pesquisa os *tokens* são as palavras e pontuações presentes no texto.

Contudo, Ferreira et al. (2019) aplicam em seu trabalho o uso das contrações de palavras e a substituição de clíticos durante a etapa de *tokenização*. Em uma contração a palavra "da" é desmembrada nos *tokens* "de" e "a", dado que, segundo as regras gramaticais da língua portuguesa, "da" é a combinação da preposição "de" com o artigo "a". A substituição de clíticos, pode ser exemplificada pela transformação da palavra "comprei-o" nas palavras "eu", "comprei" e "ele".

Segundo Ferreira et al. (2019), a substituição das contrações e dos clíticos aprimora o processo de marcação de classes gramaticais (PoS *tagging*) e o Reconhecimento de Entidades Nomeadas. Entretanto, a utilização dessas técnicas faz com que o número de palavras aumente, além disso, as contrações e clíticos podem significar marcas estilísticas do autor, impactando diretamente na análise literária aqui exposta. Por conseguinte, no método computacional aqui proposto, essas técnicas não são utilizadas.

Uma comparação da *tokenização* utilizando a contração de palavras e a substituição de clíticos e outra sem uso de contrações e substituição clíticos pode ser observada na Tabela 3.1.

Tabela 3.1: Comparação das *tokenizações* a partir da primeira sentença do livro *Macunaíma*.

| **Sentença** | **No fundo do mato-virgem nasceu Macunaíma, herói de nossa gente.** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Tokenização* por Ferreira et al. (2019) | No | fundo | de | o | mato-virgem | nasceu | Macunaíma | , | herói | de | nossa | gente | . |
| *Tokenização* utilizada | No | fundo | do | | mato-virgem | nasceu | Macunaíma | , | herói | de | nossa | gente | . |

Após a etapa de *tokenização* inicia-se a marcação dos *tokens* em:

- Classes Gramaticais: envolve a marcação de cada *token* como substantivo, adjetivo, advérbio, artigo, conjunção (por tipo), pronome (por tipo), nome próprio, preposição, pontuação ou verbo (por tipo);

- Entidades nomeadas: busca identificar cada *token* como uma entidade nomeada ou não. Caso o *token* seja reconhecido como uma entidade nomeada ele é marcado como: abstração, acontecimento, coisa, local, obra, organização, outro, pessoa, tempo ou valor.

Para fins de organização, as marcações derivadas do Reconhecimento de Entidades Nomeadas (REN) também estão incluídas no conjunto de características estilométricas. Isso se deve ao fato de que alguns trabalhos relacionados utilizam essas características na identificação de autoria, como exposto na seção de Trabalhos Relacionados (Seção 2.5).

Além disso, também é realizada a identificação do lema da palavra, que representa sua forma base. Conforme abordado no Capítulo 2, a lematização objetiva reduzir palavras flexionadas a uma forma comum para que elas possam ser tratadas de forma consistente em análises de texto.

Portanto, cada *token* recebe uma marcação de classe gramatical, que pode ser substantivo, adjetivo, advérbio, artigo, conjunção, pronome, nome próprio, preposição, pontuação ou verbo. Também é atribuída a marcação de entidade nomeada, que pode ser abstração, acontecimento, coisa, local, obra, organização, outro, pessoa, tempo ou valor. Por fim, o lema da palavra também é identificado nesse mesmo processo.

Em resumo, um exemplo concreto da marcação dessas características está presente na Tabela 3.2. Nessa tabela, é possível examinar a identificação das marcações de PoS, a lematização e o REN na primeira sentença do livro *Macunaíma*.

Tabela 3.2: Extração de características estilométricas em uma sentença do livro Macunaíma.

| Sentença nº1 | No fundo do mato-virgem nasceu Macunaíma, herói de nossa gente. | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Token* | No | fundo | do | mato-virgem | nasceu | Macunaíma | , | herói | de | nossa | gente | . |
| PoS *tagging* | ADP | n | ADP | N | v-fin | N | punc | n | prp | pron-det | n | punc |
| Lema | no | fundo | do | mato-virgem | nascer | Macunaíma | , | herói | de | nosso | gente | . |
| REN | Não identificadas na sentença. | | | | | | | | | | | |

O resultado da identificação de classes gramaticais consiste na atribuição de rótulos a cada um dos *tokens*, indicando sua classe gramatical. Na sentença de exemplo, é possível observar as seguintes marcações:

- ADP: adposição;
- n e N: substantivos comuns e próprios, respectivamente;
- v-fin: verbo finito;
- punc: pontuação;
- prp: preposição;

- pron-det: pronome determinante.

Adicionalmente, é viável notar o lema associado a cada *token* e o REN, que categoriza cada *token* como uma Entidade Nomeada. Nota-se que, mesmo a palavra "Macunaíma" sendo uma Entidade Nomeada, o método atual não a reconhece como tal. Essa limitação decorre das dificuldades que métodos computacionais enfrentam para identificar corretamente entidades nomeadas, conforme abordado no Capítulo 2.

A partir deste pré-processamento, para completar as etapas de *tokenização* ao nível das palavras, a marcação de classes gramaticais, a identificação de lemas e o REN, é necessário reunir e, em seguida, agrupar esses *tokens* em forma de sentenças para realizar a extração de características.

Uma sequência de *tokens* é considerada como uma única sentença abrangendo desde o seu início até um sinal de pontuação de fim de sentença (! ou ? ou . ou .. ou !? ou ?! ou . . . ). Isso viabiliza a contagem das sentenças presentes no texto, a identificação de *tokens* que as compõem e a organização da extração de características por sentença.

Este agrupamento das palavras/*tokens* em sentenças, por sua vez, dá origem à criação de um parâmetro ajustável denominado Unidade Textual (UT). Uma UT de tamanho 1 é equivalente a uma sentença extraída. Ao ajustar esse parâmetro, por exemplo, para 100, o resultado é a divisão de todo o texto em partes contendo no máximo 100 sentenças cada.

Por fim, cada UT é constituída por uma quantidade específica de *tokens*, o que implica que cada UT contém uma série de características quantificáveis. Essas características podem incluir a contagem de artigos, a quantidade de palavras/*tokens* presentes, o número de entidades nomeadas identificadas, entre outros aspectos semelhantes.

As características estilométricas extraídas são:

- Comprimento da Sentença: O número de palavras em cada sentença.
- Unicidade de palavras: Contagem das palavras únicas, Contagem dos lemas únicos, TTR de palavras, TTR de lemas.
- Classes Gramaticais (*PoS Tagging*): Contagens das classes gramaticais das palavras, incluindo adjetivos, advérbios, artigos, conjunções, interjeições, numerais, pontuação, pronomes, preposições, substantivos e verbos.
- Entidade nomeada: Contagens de entidades nomeadas, incluindo pessoas, locais, combinações de pessoa e local, e todas as entidades nomeadas em geral.

Dependendo do tamanho da UT em termos de quantidade de palavras, a utilização dessas características quantificadas em valores absolutos pode influenciar na análise. Por exemplo, uma UT com 100 palavras pode possuir uma maior quantidade de adjetivos em comparação a uma UT com 50 palavras. Além disso, como as demais

características também são afetadas pelo tamanho da UT, uma correlação entre elas pode ser influenciada por esse fator. Um exemplo é a correlação entre o par de características contagem de palavras e contagem de substantivos, já que a quantidade de palavras de uma UT pode influenciar a quantidade de substantivos.

Para possibilitar análises não influenciadas pelo tamanho das UTs, são extraídas as frequências relativas de algumas características. Isso é feito para que a quantidade de palavras não se torne uma variável de confusão na avaliação de correlação entre características. Para isso, as frequências relativas são calculadas dividindo a quantidade da característica específica na UT pelo número total de palavras na mesma UT, proporcionando uma perspectiva mais equilibrada e ajustada em relação ao tamanho da UT.

Com isso, além das características de frequências absolutas listadas acima são extraídas as seguintes características estilométricas de frequências relativas:

- Frequências de PoS: adjetivo, advérbios, artigos, conjunções, interjeições, numerais, pontuação, pronomes, preposições, substantivos, verbos;
- Frequências de Entidades Nomeadas: Pessoa, Local, Pessoa+Local, Todas as entidades nomeadas;

As características que foram extraídas e identificadas nesta subseção possibilitam a identificação de padrões literários paralelos e o estabelecimento de medidas de correlação entre diferentes características, ampliando a capacidade de análise e proporcionando novas medidas para a pesquisa em questão.

### 3.1.2 Características de Análise de Sentimentos

Os trabalhos relacionados expõem a extração de características textuais utilizando a Análise de Sentimentos. Para garantir uma análise textual abrangente e em diversos níveis linguísticos, o método computacional realiza a extração de características a partir da Análise de Sentimentos.

Sendo assim, após a extração de *tokens* dos textos analisados, foram utilizados dicionários léxicos (Pennebaker et al., 2001; Souza e Vieira, 2011; Silva et al., 2012) para a identificação da polaridade de cada *token*.

A polaridade de um *token* refere-se à sua classificação como positivo ou negativo. Em outras palavras, os *tokens* de palavras presentes no texto são marcados conforme a polaridade associada aos dicionários léxicos. Os *tokens* positivos são associados à polaridade positiva, enquanto os negativos são associados à polaridade negativa.

Para associar essa polaridade, cada *token* é submetido a uma busca nos dicionários léxicos mencionados. Se o *token* for encontrado, a identificação da polaridade é realizada. Caso contrário, o lema do *token*, extraído no processo anterior, é utilizado. Por último, se nem o *token* nem o seu lema são encontrados em nenhum dos dicionários léxicos utilizados, a polaridade é considerada neutra para o *token* em questão.

Isso resulta em cada *token* sendo associado a uma das três polaridades: negativa, neutra ou positiva.

Além dos dicionários léxicos empregados na associação da polaridade, também é utilizado o dicionário de Mohammad e Turney (2013) para associar os *tokens* a diferentes emoções: alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza. Essas emoções são alinhadas com a teoria das oito emoções básicas proposta por Plutchik (1982).

O processo de associação dessas emoções segue o método de associação de polaridade descrito anteriormente. Inicialmente, busca-se o *token* no dicionário, e se esse não for localizado, recorre-se à busca pelo lema do *token*. Quando o *token* ou seu lema são encontrados, as emoções associadas àquele *token* são associadas, caso contrário, aquele *token* recebe apenas a marcação de polaridade.

Isso resulta em cada *token* sendo associado a uma polaridade (negativo, neutro ou positivo), podendo ainda ser associado a uma das emoções possíveis: alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza.

Através da extração da polaridade e das emoções dos *tokens*, torna-se viável a identificação de valores quantificáveis dessas características tanto em sentenças individuais quanto em UTs.

Um exemplo da extração da polaridade e da emoção pode ser observado na Tabela 3.3. Nesta tabela, é possível observar a polaridade e a(s) emoção(ões) associada(s) a cada *token*. Na sentença de exemplo, houve a identificação da polaridade negativa em duas palavras: "incitavam" e "Ai". Além disso, essas mesmas palavras foram as únicas com emoções identificadas. Vale destacar que todas essas emoções associadas são efetivamente negativas, proporcionando uma visão mais abrangente do conteúdo além da simples informação de polaridade.

Tabela 3.3: Associação de características de Análise de Sentimentos em uma sentença do livro Macunaíma.

| Sentença nº 7 | Sio incitavam a falar exclamava: If — Ai! | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Tokenização* | Sio | incitavam | a | falar | exclamava | : | If | — | Ai | ! |
| Polaridade | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | -1 | 0 |
| Emoção(ões) | - | Expectativa<br>Medo<br>Raiva | - | - | - | - | - | - | Medo<br>Nojo<br>Tristeza | - |

Dessa maneira, semelhante às características estilométricas abordadas na subseção anterior, as características derivadas da Análise de Sentimentos também caracterizam níveis para identificação de paralelismos neste trabalho.

Com isso, as características selecionadas e extraídas a partir da Análise de Sentimentos são:

- Análise de sentimento: Contagens de palavras relacionadas à positividade, negatividade, neutralidade, bem como emoções específicas como alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza.
- Sentimento associado a classe gramatical: Contagens de adjetivo positivo, adjetivo negativo, advérbio positivo e advérbio negativo.
- Polaridades e Cargas Emocionais: Polaridade da Sentença, Carga Emocional da sentença, Polaridade de adjetivos, Polaridade de advérbios, Carga Emocional de adjetivos e Carga Emocional de advérbios;

Além das características de frequências absolutas, listadas acima, foram extraídas as seguintes características de frequências relativas:

- Frequências de Sentimentos: Positividade, Negatividade, Neutralidade, Alegria, Confiança, Expectativa, Medo, Nojo, Raiva, Surpresa, Tristeza;

### 3.1.3 Características poéticas

As características extraídas de figuras de linguagem e elementos poéticos também podem ser utilizadas em uma análise textual em prosa, a exemplo dos trabalhos de Carvalho et al. (2020); Lima et al. (2021).

Algumas características poéticas podem ser extraídas por meio de sistemas computacionais que realizam a escansão dos textos, possibilitando a identificação de versos a partir das suas sentenças. Com esse intuito, o método computacional aqui proposto apoia-se no trabalho de Carvalho et al. (2020) para identificar e classificar estruturas métricas obtidas a partir da escansão.

Para atingir esse objetivo, Carvalho et al. (2020) propõem as seguintes etapas:

- Extração de segmentos frásicos: a partir da cópia do texto é feita a segmentação do mesmo em sentenças.
- Módulo de separação silábica: com a extração de sentenças, cada palavra da sentença sofre a separação silábica e a marcação de sílaba tônica.
- Escansão e identificação de estruturas de versificação: é realizada a detecção de fenômenos inter-vocabulares ou intra-vocabulares, considerando que fenômenos como crase, sinérese, diérese e elisão podem impactar a escansão. Com isso, torna-se viável a extração das estruturas métricas, podendo ser feita a partir de sentenças completas, do início ou do final de uma sentença. Além disso, é possível definir um tamanho mínimo e máximo para as estruturas métricas, abrangendo desde monossílabos (uma sílaba) até dodecassílabos (doze sílabas).

Com essa abordagem, torna-se viável extrair características métricas de diferentes maneiras: de sentenças completas, no início de sentenças e no final de sentenças.

Além disso, o tamanho das estruturas métricas selecionado para o método computacional aqui proposto varia de um mínimo de heptassílaba (sete sílabas) a um máximo de dodecassílaba (doze sílabas), resultando na inclusão de novas características para a análise textual.

A Tabela 3.4 oferece um exemplo de como a sentença de número 3013 do livro *Os Sertões* foi submetida à escansão e classificação. Nesse processo, as características métricas foram extraídas a partir de sentenças completas. Na tabela, é possível observar a identificação de cada sílaba, separada por uma barra (/), bem como a marcação da sílaba tônica, indicada por um símbolo de cerquilha (#)

Tabela 3.4: Exemplo de sentença escandida pelo *MIVES*.

| | |
|---|---|
| Sentença nº 3013 | Entram também de certo modo na luta. |
| Escansão | #En/tram/ tam/b#ém/ de/ c#er/to/ m#o/do/ na/ l#u/ta. |
| Classificação | Hendecassílabo (11 sílabas) |

No exemplo apresentado na Tabela 3.4, o verso extraído foi classificado como hendecassílabo, o que significa que ele contém 11 sílabas.

Outra característica poética que pode ser extraída de textos em prosa é a identificação de sequências de repetições de sílabas fonéticas. Ao identificar essas repetições fonéticas, torna-se possível quantificá-las em diferentes partes do texto, permitindo a identificação de trechos que contenham um maior número de repetições fonéticas.

Com o propósito de alcançar esse objetivo, Lima et al. (2021) propuseram um método computacional contendo as seguintes etapas:

- *Tokenização*: A partir do envio de um texto é feita a *tokenização* das palavras do mesmo. Em seu trabalho, Lima et al. (2021) realiza a *tokenização* a partir da separação de todas as palavras do texto dos sinais de pontuação e dos caracteres de escape. Esse processo é realizado a partir da expressão regular: [\w]+|['()-.,!:?;\n\t]. Dessa forma a *tokenização* aplicada por Lima et al. (2021) para a identificação de repetições de sílabas fonéticas difere da utilizada até o momento neste trabalho.

- Divisão Silábica e Conversão para Fonemas: Após a tokenização, cada palavra é dividida em sílabas grafemáticas. Em seguida, utilizando um dicionário de divisão silábica, ocorre a conversão em sílabas fonéticas.

- Busca por Repetições de Sílabas Fonéticas: Após a divisão das palavras em sílabas fonéticas, é realizada a busca por pares de sílabas fonéticas similares. Isso possibilita a construção e filtragem das sequências de repetição.

- Construção e Filtragem de Sequências de Repetições de Sílabas Fonéticas: A construção das sequências de repetição acontece ao identificar pares que se repetem em um intervalo de palavras, conhecido como janela de busca, conforme definido por Lima et al. (2021). Essa janela de busca pode ser ajustada para

obter uma variedade de resultados. A filtragem das sequências ocorre com base em um número mínimo de repetições. Esse valor mínimo representa o comprimento de uma sequência de repetição, ou seja, a quantidade total de repetições presentes em cada janela de busca.

| S | L | Macunaíma | | Os Sertões | | Triste Fim | |
|---|---|---|---|---|---|---|---|
| | | abs | rel | abs | rel | abs | rel |
| 2 | 3 | 200 | 0.46% | 1829 | 1,15% | 851 | 1.27% |
| 2 | 4 | 39 | 0.09% | 383 | 0,24% | 188 | 0.28% |
| 2 | 5 | 16 | 0.04% | 100 | 0,06% | 49 | 0.07% |
| 3 | 3 | 449 | 1.03% | 3354 | 2,11% | 1525 | 2.28% |
| 3 | 4 | 92 | 0.21% | 923 | 0,58% | 388 | 0.58% |
| 3 | 5 | 38 | 0.09% | 288 | 0,18% | 124 | 0.19% |
| 4 | 3 | 782 | 1.79% | 4461 | 2,90% | 2205 | 3.30% |
| 4 | 4 | 179 | 0.41% | 1572 | 0,99% | 646 | 0.97% |
| 4 | 5 | 64 | 0.15% | 564 | 0,35% | 236 | 0.35% |
| 5 | 3 | 1170 | 2.68% | 6671 | 4,19% | 2836 | 4.24% |
| 5 | 4 | 297 | 0.68% | 2295 | 1,44% | 912 | 1.36% |
| 5 | 5 | 118 | 0.27% | 925 | 0,58% | 381 | 0.57% |

Figura 3.2: Impacto da variação dos parâmetros para extração de repetições silábicas, (Lima et al., 2021)

Com isso, Lima et al. (2021) demonstra como a variação desses parâmetros pode afetar os resultados obtidos. A Figura 3.2 apresenta como as variações dos parâmetros de tamanho da janela de busca (S) e de comprimento mínimo da sequência (L) impactam nos resultados obtidos em valores absolutos e relativos (em relação ao número de palavras do texto) de diferentes livros em prosa da literatura brasileira.

Essa variação de parâmetros e resultados também é aplicada no método computacional proposto, buscando analisar novas características textuais e avaliar a correlação e análise textual em diferentes níveis. Com o objetivo de realizar uma análise textual mais abrangente, foram utilizados os tamanhos S=2 com L=3, S=3 com L=4 e S=4 com L=4 para obter os resultados nos livros a serem analisados.

Consequentemente, é possível identificar: a sílaba fonética que se repete na sequência, as posições textuais das sílabas fonéticas que se repetem, o comprimento da sequência (quantidade de repetições de sílabas fonéticas) e a quantidade total de sequências de repetições de sílabas fonéticas no texto.

A Tabela 3.5 ilustra as repetições identificadas na sentença de número 1247 do livro *Os Sertões*, indicando a sílaba repetida que é destacada entre colchetes ([ ]).

No exemplo da Tabela 3.5, foram identificadas três repetições silábicas diferentes, duas da sílaba [a] e uma da sílaba [tSi]. Por esse processo de extração de repetições não necessitar da utilização do conceito de sentença, quando a primeira repetição identificada se encontra entre a sentença destacada e sua sentença anterior, a sentença mais afetada pela repetição extraída é identificada e marcada com a referida

Tabela 3.5: Exemplos de repetições identificadas em uma sentença pelo *ALLPRO*.

| | |
|---|---|
| Sentença nº 1247 | A sua evolução psíquica, por mais demorada que esteja destinada a ser, tem, agora, a garantia de um tipo fisicamente constituído e forte. |
| Rep. Sílabica [ a ] | a conquistar um di[a].<br>[a] su[a] evolução psíquica, |
| Rep. Silábica [a] | tem, [a]gora, [a] garanti[a] de um tipo |
| Rep. Silábica [tSi] | garantia de um [ti]po fisicamen[te] cons[ti]tuído e forte. |

repetição. Esse processo pode ser visualizado na primeira repetição silábica [a] da Tabela 3.5, onde existe um trecho da sentença anterior na repetição, entretanto, essa repetição afeta mais a sentença em questão (nº 1247), por conter duas repetições da sílaba [a] no trecho analisado.

As características selecionadas e extraídas a partir dos trabalhos de Carvalho et al. (2020) e Lima et al. (2021) são:

- Repetições de Sílabas fonéticas: Contagens onde foram consideradas as variações no tamanho da janela de busca (S) e no comprimento mínimo da sequência (L), como S2L3, S3L4 e S4L4.
- Sentenças métricas: Foram identificadas e somadas as sentenças métricas de heptassílabo, octossílabo, eneassílabo, decassílabo, hendecassílabo e dodecassílabo a partir do início de sentenças, do final de sentenças e de sentenças completas, conforme o método proposto por Lima et al. (2021).

Consequentemente, ao acrescentar essas características poéticas às características estilométricas e de análise de sentimentos, é possível realizar a análise de padrões paralelos entre os diversos níveis textuais expostos.

### 3.1.4 Modelagem de Tópicos

A modelagem de tópicos possibilita a inclusão de características de nível semântico nas análises textuais a partir da extração e identificação dos tópicos presentes em um texto, bem como a avaliação da relevância de cada tópico ao longo do texto.

Os algoritmos utilizados na modelagem de tópicos possibilitam a descoberta automática da estrutura semântica dos documentos. Isso é alcançado por meio de análises estatísticas, que identificam padrões de coocorrência durante o treinamento com o *corpus* do documento. Essa abordagem permite, entre outras tarefas, a extração de tópicos com base na semântica do texto.

Para essa tarefa de extração de tópicos, o método computacional proposto utiliza o algoritmo *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003), por ser o mais simples de ser utilizado, de acordo com Blei (2012).A utilização do LDA possibilita a escolha da quantidade de tópicos a serem extraídos para análise, gerando diferentes resultados conforme essa quantidade aumenta ou diminui.

A utilização do algoritmo LDA no texto é realizada a partir da segmentação do mesmo em Unidades Textuais (UTs). Como o tamanho da UT é um parâmetro

ajustável ele permite a obtenção de resultados variados para a análise textual de um mesmo livro. O algoritmo LDA considera cada uma dessas UTs como um documento distinto, determinando tópicos e suas respectivas relevâncias para cada uma delas.

Assim, a modelagem de tópicos é realizada apenas com valores de UT maiores que 1. Quando UT é igual a 1, representa apenas uma única sentença do texto, o que torna a aplicação da modelagem de tópicos difícil devido à falta de conteúdo significativo nessa extensão reduzida. Por outro lado, valores mais altos permitem agrupar o texto em segmentos, conferindo à UT uma estrutura mais robusta.

A seguir, como etapa de pré-processamento, anterior ao treinamento do algoritmo, há a remoção das *stop words* (palavras que não possuem relevância para o treinamento). A lista de *stop words* pode ser conferida no Anexo B.

Além disso, o pré-processamento do texto envolve outras etapas, como a remoção de caracteres especiais, a conversão de letras maiúsculas para minúsculas e a transformação das palavras em *tokens*. Embora seja possível considerar a remoção de acentos, é importante reconhecer que na língua portuguesa, a acentuação gráfica desempenha um papel vital ao diferenciar palavras e segue regras e normas específicas. Portanto, a remoção dos acentos pode ter um impacto direto nos resultados e por isso não é utilizada.

Após o pré-processamento do texto, o algoritmo é treinado com o corpus do texto, sendo a quantidade de tópicos definida como parâmetro. Após uma análise da variação desse parâmetro, chegou-se à conclusão de que, para a análise aqui exposta, a quantidade ideal de tópicos é de 5. Valores acima de 5 resultam em estruturas de tópicos semelhantes, levando à repetição de alguns tópicos, enquanto valores abaixo de 5 não proporcionam uma diversidade adequada de tópicos para a análise.

Essa abordagem possibilita a extração de informações sobre a variação desses tópicos ao longo do texto e a relevância de cada tópico em cada Unidade Textual. Como resultado, cada tópico é associado a um valor quantitativo de relevância em todo o texto.

A Figura 3.3 ilustra um gráfico de linhas no qual cada linha corresponde a um tópico identificado ao longo do livro *Os Sertões*, de Euclides da Cunha. O eixo $X$ representa as UTs de tamanho 100, enquanto o eixo $Y$ mostra os valores de relevância de cada tópico ao longo das Unidades Textuais.

As características selecionadas e extraídas a partir da modelagem de tópicos resultou em cinco características para a análise internível, onde cada tópico possui um valor de relevância em cada Unidade Textual. Nesse caso, a quantidade de características equivale à quantidade de tópicos extraídos (5 no caso desta pesquisa). Essas características foram extraídas apenas com UTs de tamanhos acima de 1.

Desse modo, os tópicos extraídos se transformam em características para a análise textual, sendo a relevância de cada tópico em cada UT a quantificação dessas características.

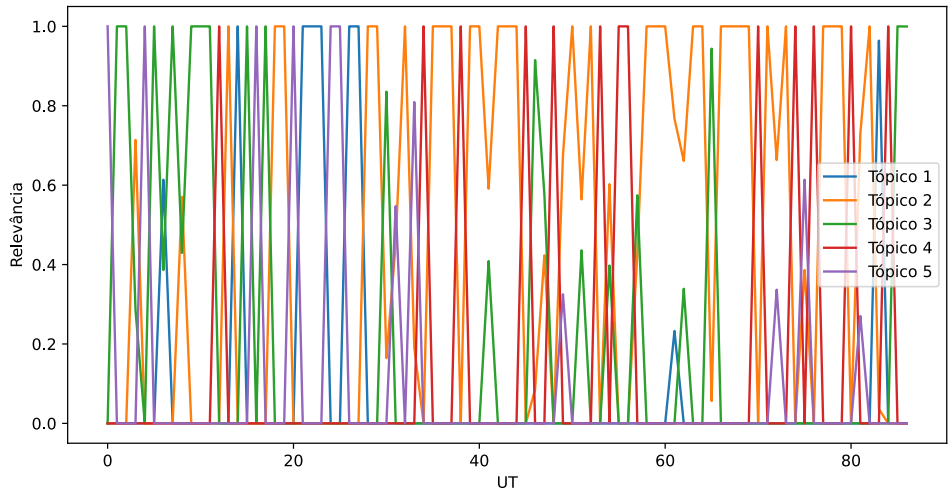


Figura 3.3: Variação de tópicos e seus valores de relevância ao longo do livro *Os Sertões*, de Euclides da Cunha, com UT de tamanho 100

## 3.2 Divisão das características em níveis

O método computacional proposto busca a identificação de paralelismos textuais interníveis, sendo assim, existe a necessidade de dividir as características extraídas, organizando-as em diferentes níveis.

As características extraídas foram divididas nos seguintes níveis:

- Nível Fonético: Neste nível, estão incluídas as características relacionadas à fonética do texto, como a repetição de sílabas fonéticas.
- Nível Métrico: Este nível se concentra na métrica do texto, ou seja, na estrutura de versificação. Inclui as características relacionadas aos padrões métricos, sejam eles de sentenças completas, de ínicio das sentenças ou de final das sentenças.
- Nível Lexical: Aqui, estão organizadas as características relacionadas ao vocabulário do texto, como a contagem de palavras, palavras únicas, lemas únicos e a *Type-Token Ratio* (TTR).
- Nível Gramatical: Neste nível estão as características relacionadas às categorias gramaticais das palavras, como adjetivos, advérbios, substantivos, etc.
- Nível Sentimental: Aqui, são consideradas as características relacionadas à análise de sentimentos, incluindo polaridade de palavras a carga emocional e emoções específicas, como alegria, medo, nojo, etc.
- Nível Gramatical e Sentimental: Este nível agrupa as características que fazem uma junção dos dois níveis e inclui características mistas, como adjetivos

positivos, carga emocional de advérbios, polaridade de adjetivos, etc.

- Nível de Entidades Nomeadas: Neste nível, estão incluídas as características relacionadas ao reconhecimento de entidades nomeadas, como nomes de pessoas e locais.
- Nível de Tópicos: Aqui, são abordadas as características relacionadas à modelagem de tópicos, permitindo a análise de como os tópicos discutidos no texto podem estar relacionados entre diferentes partes do texto.

Essa divisão em níveis agrupa características com finalidades semelhantes e fornece uma visão de como diferentes aspectos do texto podem estar relacionados em termos de paralelismo. Essa divisão, entretanto, é apenas uma das possibilidades de organização das características, podendo ser alterada de acordo com os interesses da pesquisa.

A lista completa das características e seus respectivos níveis pode ser encontrada na Tabela C.1, disponível no Anexo C.

## 3.3 Análise de correlação e identificação de paralelismos

A identificação de paralelismos no método computacional proposto baseia-se na análise de correlação entre as características extraídas. Após a extração dessas características, o método computacional investiga suas relações, permitindo destacar os padrões presentes nos textos literários em vários níveis de linguagem.

O método computacional realiza essa análise de correlação utilizando tanto o coeficiente de correlação de Pearson quanto o de Spearman para avaliar as relações entre as sequências de valores de distintas características extraídas em diferentes níveis.

Para avaliar a correlação global para cada par de características, são considerados os valores quantificados das características em todo o livro, organizados em UTs a partir de um parâmetro denominado tamanho da UT, que delimita o tamanho máximo (em quantidade de sentenças) de cada UT para análise.

Entretanto, os paralelismos textuais não podem ser deduzidos apenas pela análise global do livro, uma vez que esses padrões podem ocorrer em trechos específicos do texto, em vez de abranger toda a obra.

Portanto, uma correlação global não implica necessariamente que essas características sejam paralelas ao longo de toda a narrativa. Além disso, características que possuem valores de correlação próximos a zero em uma análise global podem estar correlacionadas em momentos particulares do livro, que podem possuir valores de correlação próximos de 1 ou -1, caracterizando uma correlação forte.

Por esse motivo, o método computacional proposto utiliza a detecção de paralelismos em segmentos da obra, empregando a correlação a partir de janelas deslizantes. Isso

possibilita identificar pontos de correlação e paralelismos na obra, em vez de se basear exclusivamente na correlação global do livro.

Essa possibilidade implica que a correlação deslizante nem sempre corresponde à correlação global do livro, e permite a identificação de paralelismos em pontos específicos do texto.

Por sua vez, a identificação de paralelismos pelo método proposto é dependente de um parâmetro denominado limiar, que delimita um valor mínimo absoluto para identificar quais pares de características possuem paralelismos a partir do seu coeficiente de correlação. Por padrão, o limiar possui um valor absoluto de 0.75. Esse valor de 0.75 foi escolhido pois caracteriza uma correlação forte (Larson e Farber, 2015), e por ser absoluto permite a identificação de correlações fortes positivas e fortes negativas. Com esse valor padrão de 0.75 correlações iguais ou superiores a 0.75 (positivos ou negativos) são consideradas como pontos de paralelismo na obra. Vale ressaltar que esse valor padrão pode ser alterado para atender as necessidades da análise, ou ainda, baseado em um estudo prévio de textos da prosa não literária e de textos de poesias.

Para a realização dessa identificação de paralelismos não foram consideradas correlações entre características do mesmo nível, uma vez que o objetivo principal deste trabalho é identificar correlações entre diferentes níveis. Além disso, as correlações entre níveis de frequência absoluta e suas respectivas frequências relativas não foram incluídas na análise pelo mesmo motivo.

Outro parâmetro ajustável é o tamanho da janela deslizante, responsável por delimitar a quantidade de UTs dentro de cada janela para o cálculo da correlação segmentada. O valor mínimo e padrão para esse parâmetro é 10 UTs, podendo ser modificado de acordo com as necessidades da análise. Esse valor mínimo foi escolhido porque janelas menores que 10 podem não demonstrar valores aceitáveis para validar uma correlação, considerando que esse valor estabelece a quantidade de pares utilizados para o cálculo do coeficiente.

Para ilustrar como os parâmetros funcionam na prática, consideremos o livro *Os Sertões*, que consiste em um total de 8634 sentenças. Com um tamanho de UT configurado como 100 este livro resulta em 87 UTs, cada uma contendo até 100 sentenças. Portanto, com a configuração padrão de tamanho 10 para a janela deslizante, a correlação deslizante será inicialmente calculada usando como recorte as UTs de 0 a 9, depois as UTs de 1 a 10, e assim por diante.

O resultado desse processo é uma tabela que lista cada par de características (uma combinação simples de todas as características de diferentes níveis tomadas de 2 em 2). Para cada par, são identificados os valores absolutos de correlação que atendem ou excedem o limiar configurado, os pares que não atingem ou ultrapassam o limiar são desprezados por essa identificação mas podem ser visualizados posteriormente. O arquivo também registra o método de correlação em que o paralelismo foi identificado (seja o de Pearson, o de Spearman ou ambos), os valores de coeficiente de correlação

global, coeficientes mínimo e máximo das correlações segmentadas encontradas para esse par, a quantidade de pontos/trechos (em janelas de UTs) onde esses paralelismos foram identificados e os seus índices.

A quantidade de pares de características que podem ser analisadas pelo método proposto é de 4268, valor dado a partir do número de combinações possíveis entre as características selecionadas de níveis diferentes multiplicado por dois, já que o cálculo de correlação pode ser feito a partir dos coeficientes de Pearson e Spearman, podendo ocorrer a identificação de paralelismos em ambos.

Além disso, existe uma quantidade máxima de paralelismos que podem ser identificados, essa quantidade é dada pela multiplicação da quantidade de pares de características pela quantidade de janelas. Para um livro com 50 UTs e com a janela deslizante de tamanho 10 teremos ao todo 40 janelas, com isso, a quantidade máxima de paralelismos que podem ser encontrados pelo método proposto é 170720.

## 3.4 Visualização dos resultados

Após a identificação de paralelismos, é possível visualizar os resultados por meio da estatística descritiva permitindo leitura distante, utilizando tabelas e gráficos que demonstram como essas características variam ao longo do texto que está sendo analisado. Essa representação visual também ajuda a identificar os momentos no texto em que essas características se destacam, tornando mais fácil localizar e visualizar esses trechos específicos dentro do próprio texto, permitindo uma leitura aproximada.

### 3.4.1 Formas de apresentação dos resultados

As representações visuais proporcionam uma melhor compreensão das análises e identificações realizadas. As formas de visualização utilizadas são as seguintes:

- Tabelas Descritivas: As tabelas descritivas são usadas para apresentar de forma organizada os resultados quantitativos dos paralelismos identificados. Cada linha da tabela representa um paralelismo identificado pelo método. As colunas representam as informações sobre o paralelismo encontrado, como o par de características onde o paralelismo foi encontrado, o método de cálculo de correlação (Pearson ou Spearman), o coeficiente da correlação global (considerando todo o livro), os coeficientes máximo e mínimo da correlação segmentada e a quantidade de paralelismos identificados utilizando a correlação segmentada.
- Matriz de correlação: As matrizes de correlação representam os valores dos coeficientes de correlação global entre características de diferentes níveis. As cores mais avermelhadas indicam valores mais próximos de 1, enquanto as cores mais azuladas representam valores mais próximos de -1.

- Gráficos de Linha: Os gráficos de linha são usados para dois propósitos: representar a variação da quantificação de um par de características ao longo do livro ou representar a variação da correlação segmentada ao longo do livro. No primeiro caso, cada linha no gráfico corresponde a uma característica específica, o eixo $x$ representa as UTs (trechos do texto), enquanto o eixo $y$ mostra os valores da característica. No segundo caso, uma linha representa o coeficiente de correlação, o eixo $x$ representa as janelas de UTs, enquanto o eixo $y$ mostra os valores do coeficiente de correlação.
- Gráficos de Dispersão: Os gráficos de dispersão são usados para ilustrar as relações entre duas características específicas. Cada ponto no gráfico representa uma UT, com um eixo representando uma característica e o outro eixo representando a segunda característica. Essa forma de visualização ajuda a identificar correlações entre características.

A escolha entre tabelas e gráficos de linha depende do tipo de análise a ser realizada e dos objetivos da pesquisa. Em alguns casos, a combinação de tabelas e gráficos de linha pode ser mais eficaz para comunicar os resultados de maneira abrangente.

### 3.4.2 Ferramenta para visualização dos resultados

Este trabalho vai além da extração e visualização das características mencionadas. Para a visualização dos paralelismos textuais identificados pelo método, foi desenvolvida uma ferramenta web, denominada CORLI.

A Figura 3.4 representa a interface inicial da CORLI. Nesta figura, é possível observar o menu lateral, onde os módulos podem ser selecionados de acordo com o objetivo da análise.

#### Módulo Paralelismos

Este é o módulo principal para a visualização de resultados. A partir dele, é possível examinar os paralelismos identificados com base na variável limiar.

A Figura 3.5 mostra a organização deste módulo e apresenta as opções de visualização de resultados disponíveis.

Os resultados obtidos de acordo com a seção anterior são apresentados em uma tabela, onde estão listados os pares de características onde foram identificadas correlações com valores iguais ou superiores ao limiar. Também são fornecidos detalhes sobre o método usado para calcular a correlação (Pearson ou Spearman), juntamente com os valores do coeficiente global, e os coeficientes mínimo e máximo da correlação segmentada. A última coluna da tabela indica quantas vezes as correlações com valores iguais ou acima do limiar foram detectadas.

Após isso, é possível selecionar qualquer linha da tabela para visualizar o paralelismo escolhido. Essa visualização é apresentada em três gráficos.

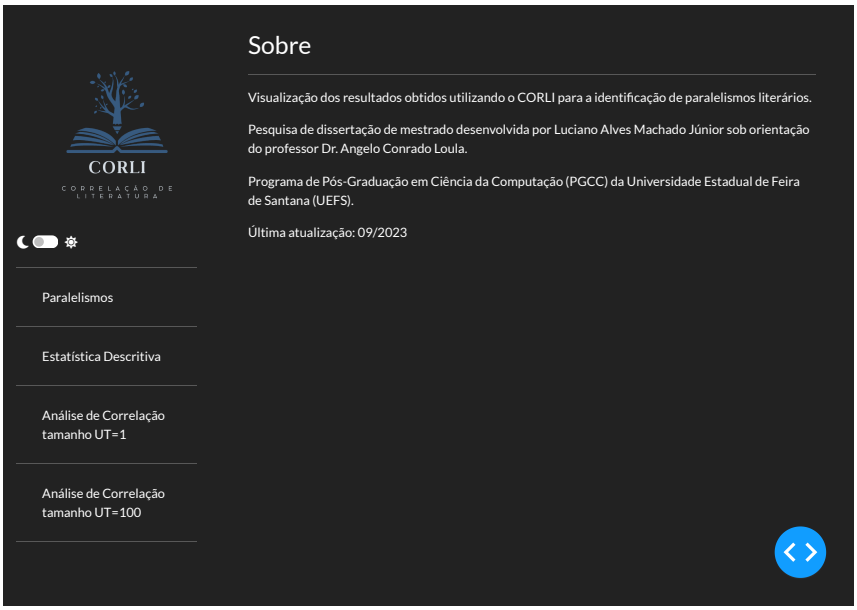


Figura 3.4: Interface inicial para visualização de resultados da CORLI

O primeiro gráfico é um gráfico de linhas que ilustra a correlação deslizante (segmentada) ao longo de todo o livro. Ele mostra marcadores em todos os pontos onde foram identificados as correlações acima do limiar, ou seja, os paralelismos encontrados.

Os outros dois gráficos são gerados a partir da escolha de uma das janelas onde o paralelismo ocorreu através de uma lista.

O segundo gráfico é um gráfico de dispersão da janela selecionada. Nele, cada ponto representa os valores do par de características em cada UT que compõe a janela.

O terceiro gráfico exibe a variação dos valores de cada uma das características que compõem o par selecionado ao longo da amostra de UTs da janela selecionada.

Por fim, são apresentados os textos de cada uma das UTs da amostra da janela, permitindo uma leitura aproximada dos trechos correspondentes.

### Módulo Estatística Descritiva

O módulo denominado Estatística Descritiva oferece a flexibilidade de selecionar qualquer característica extraída neste trabalho.

A Figura 3.6 ilustra a visualização de resultados a partir deste módulo.

Com base na escolha do usuário, o sistema exibe até dois gráficos de linhas, demonstrando a variação da característica selecionada ao longo do livro. Um gráfico considera o tamanho da UT=1, enquanto o outro leva em consideração o tamanho da UT escolhido pelo usuário, este último é exibido apenas quando o tamanho da

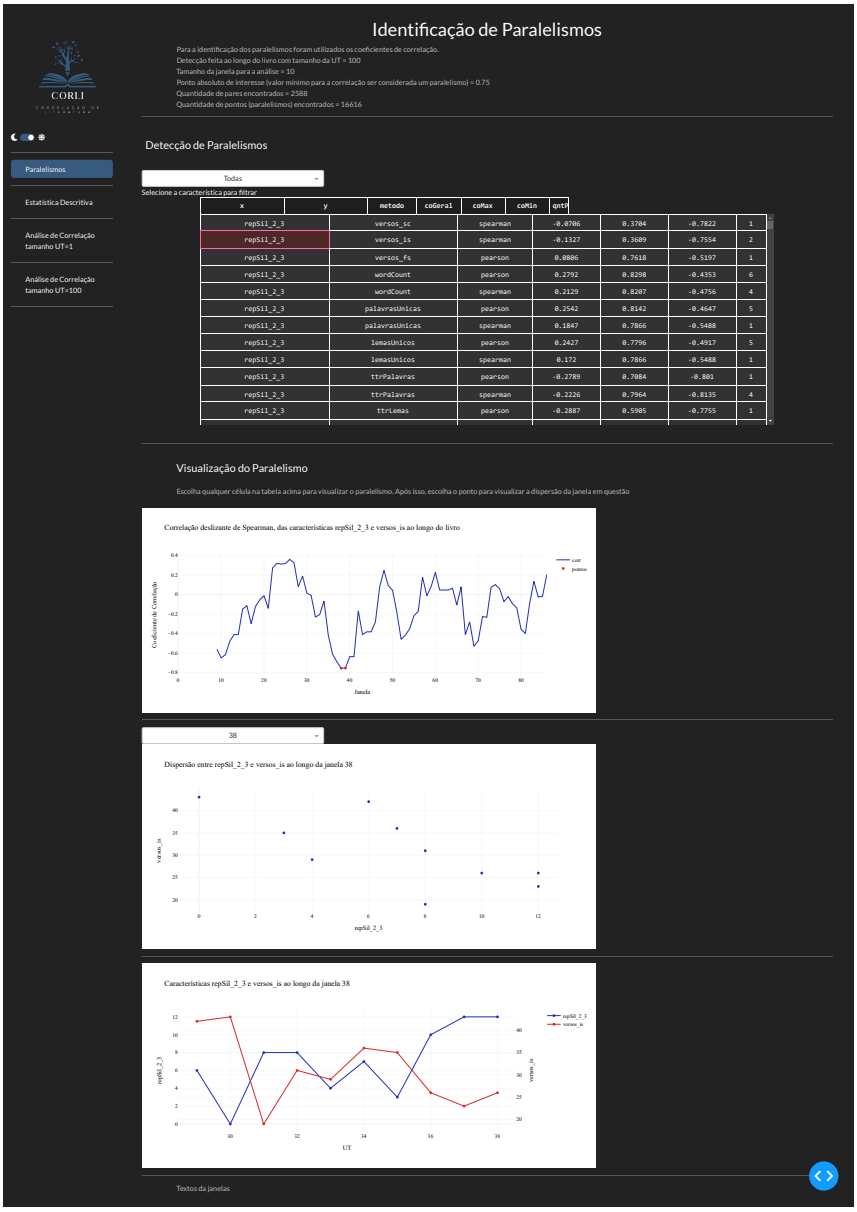


Figura 3.5: Módulo da CORLI para visualização de paralelismos detectados

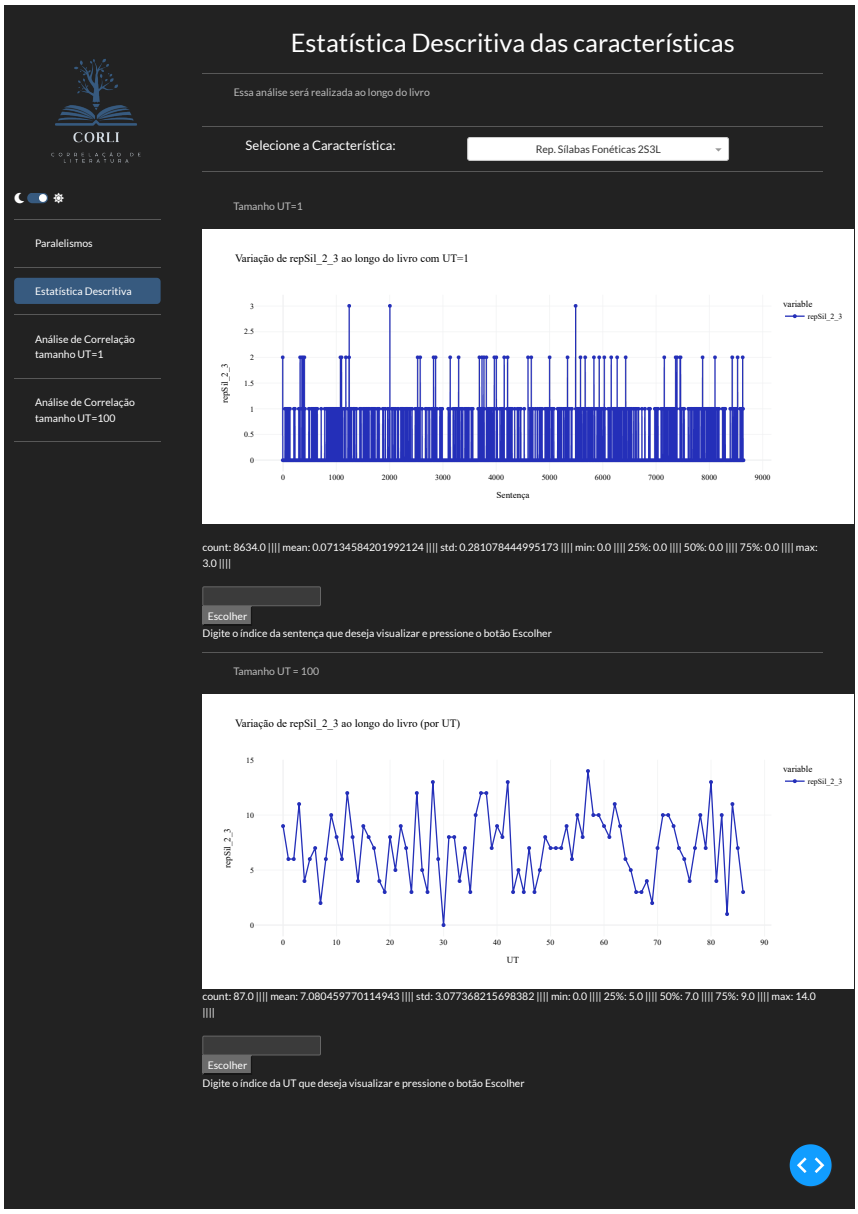


Figura 3.6: Módulo para visualização de resultados usando estatística descritiva

UT escolhido difere de 1. Além disso, abaixo de cada gráfico é possível visualizar medidas estatísticas de contagem, média, desvio padrão, mínimo, quartis e máximo.

Similarmente ao módulo anterior, é possível escolher uma sentença para visualizar o seu conteúdo textual. Além disso, é possível visualizar as sentenças que compõem determinada UT, também a partir do seu índice.

**Módulos de Análise de Correlação**

Existem dois módulos de Análise de Correlação disponíveis, um para visualizar correlações considerando o tamanho da UT=1 e outro considerando o tamanho da UT escolhido pelo usuário. O segundo módulo é exibido somente quando o usuário escolhe um tamanho de UT diferente de 1.

A Figura 3.7 representa a organização desses módulos de análise de correlação e exibe as diferentes maneiras de visualizar os resultados.

Independentemente da escolha, esses módulos compartilham funcionalidades semelhantes. Ambos permitem a seleção de dois níveis para gerar uma matriz de correlação entre eles e escolher o método de correlação (Pearson ou Spearman).

Além da matriz de correlação, é possível visualizar um gráfico de dispersão entre as duas características selecionadas, um gráfico de linhas que mostra os valores das características em cada UT e outro gráfico que apresenta a correlação deslizante das características.

A única distinção entre os módulos de correlação se torna visível somente quando o tamanho da UT escolhida pelo usuário é maior que 1. Isso ocorre porque, com esse valor, também é possível utilizar gráficos e correlações provenientes da modelagem de tópicos.

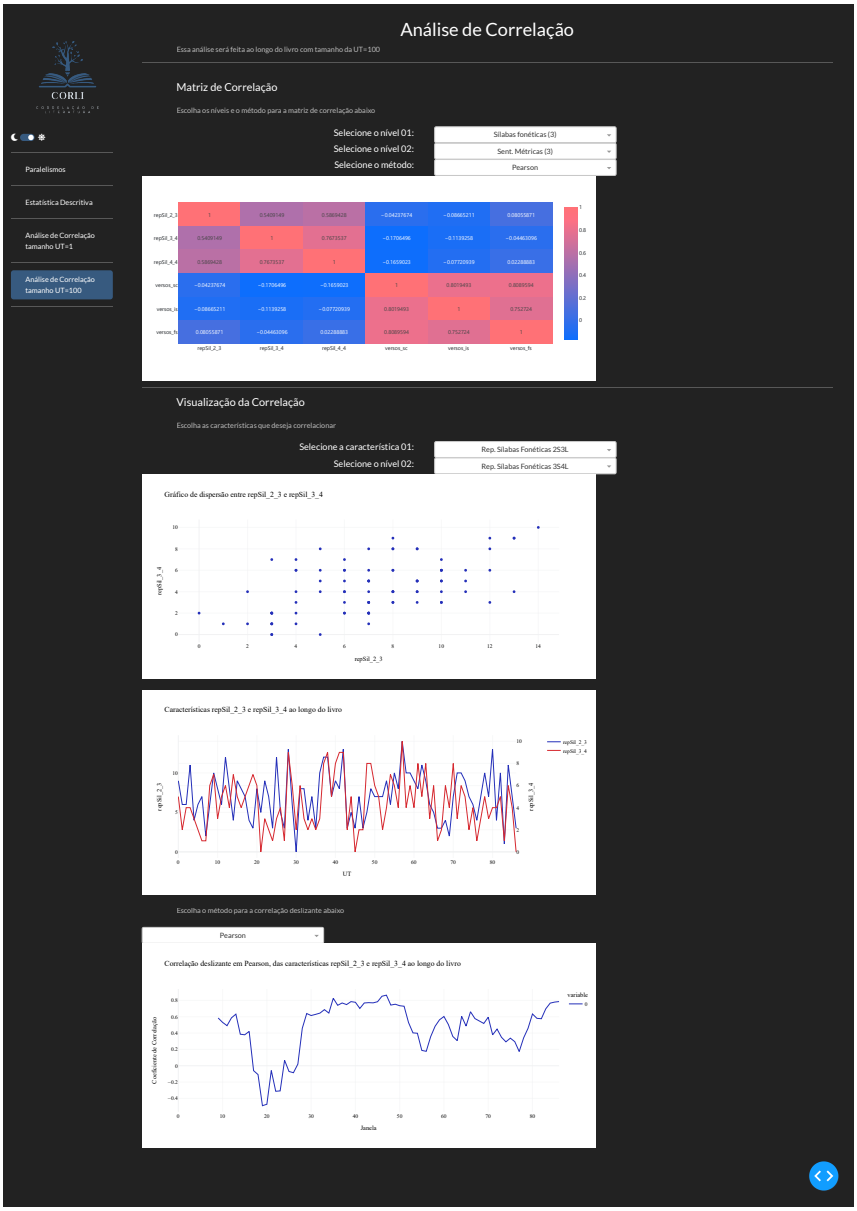


Figura 3.7: Módulo para visualização de correlações extraídas

# Capítulo 4

# Resultados

Este capítulo tem como propósito exemplificar o funcionamento do método computacional proposto na identificação de paralelismos interníveis em textos literários da literatura brasileira. Para isso, exploramos como a variação dos parâmetros do método impacta os resultados obtidos e a variabilidade dos padrões de paralelismo identificados, com foco na demonstração dos resultados da aplicação do método, destacando os padrões de paralelismo identificados, suas variações e impactos. Apresentamos e discutimos os resultados obtidos, investigando sua contribuição para uma compreensão quantitativa dos padrões textuais presentes nas obras utilizadas, bem como as implicações da utilização do método computacional para a pesquisa em linguística computacional e estudos literários.

Para a obtenção dos resultados aqui expostos, foram utilizados os seguintes livros: *Macunaíma*, de Mário de Andrade; *Dom Casmurro*, de Machado de Assis; e *Os Sertões*, de Euclides da Cunha. Esses livros são objeto de análise em outras pesquisas, como as de Carvalho et al. (2020) e Lima et al. (2021) e a relevância e contribuições para a literatura brasileira justificam a sua utilização nos resultados dessa pesquisa. Dado que *Os Sertões* é o livro mais extenso e foco principal das pesquisas mencionadas, os resultados aqui apresentados se concentram principalmente nele.

## 4.1 Impacto das diferentes formas de identificação de paralelismos

Esta seção tem como objetivo investigar e compreender como a escolha de diferentes medidas de correlação e formas de extração de características afetam os resultados da identificação de paralelismos pelo método computacional proposto.

Devido às diferentes formas de avaliação de correlação,os resultados obtidos possuem uma variabilidade que depende das correlações utilizadas. Além disso, a escolha de utilizar as características em suas frequências absolutas ou relativas também implica em variações nos resultados obtidos. Dessa forma, para demonstrar o impacto dessas

diferentes possibilidades, esta seção está dividida em subseções que exemplificam a variabilidade nos resultados, sendo elas:

- Correlação Global e Correlação Segmentada
- Correlação de Pearson e Correlação de Spearman
- Correlação utilizando a frequência absoluta e Correlação utilizando a frequência relativa

### 4.1.1 Correlação Global e Correlação Segmentada

O objetivo desta subseção é analisar e comparar os resultados obtidos ao aplicar a correlação global, que considera o texto como um todo, com a correlação segmentada, que avalia padrões em segmentos menores do texto. Isso permite compreender como a escolha entre essas duas abordagens impacta a identificação de paralelismos em diferentes níveis do texto e, assim, contribuir para uma compreensão do funcionamento do método proposto.

Nesta seção o livro utilizado é *Os Sertões* e para calcular os coeficientes de correlação utiliza-se Pearson.

Para avaliar as correlações globais e, consequentemente, observar se essas correlações se mantêm consistentes ao longo do livro ou se variam ao longo do mesmo através de correlações segmentadas através de janelas deslizantes, é possível realizar uma análise dos resultados das correlações entre os pares de características através de uma matriz de correlação entre diferentes níveis. Essa matriz apresenta os valores dos coeficientes gerais de correlação para cada par de características presentes nos níveis escolhidos.

A Figura 4.1 apresenta a matriz de correlação global de Pearson entre as características[1] dos níveis "Métrico" e "Sentimental", este último utilizando as frequências relativas (considerando a quantidade de palavras da UT) das características, com UT de tamanho 100, permitindo a observação dos coeficientes de correlação entre todos os pares que compõem esses dois níveis. A correlação mais alta entre esses níveis é encontrada entre o par de características "versos de início de sentença" (*versos_is*) e "frequência de neutros" (*neuFreq*), com um valor aproximado de 0.55.

Para visualizar a relação entre as características e comparar a dispersão conjunta dos valores obtidos é possível utilizar um gráfico de dispersão, como ilustrado na Figura 4.2, que representa a dispersão entre "versos de início de sentença"(*versos_is*) e "frequência de neutros"(*neuFreq*). A partir deste gráfico, é possível notar uma relação linear moderadamente positiva, justificando assim o coeficiente de correlação global de 0.55.

Entretanto, o paralelismo é um fenômeno que pode ocorrer em alguns trechos do livro, não necessariamente caracterizando um fenômeno que ocorre durante toda a

[1] As abreviações utilizadas na figura estão disponíveis no Anexo C

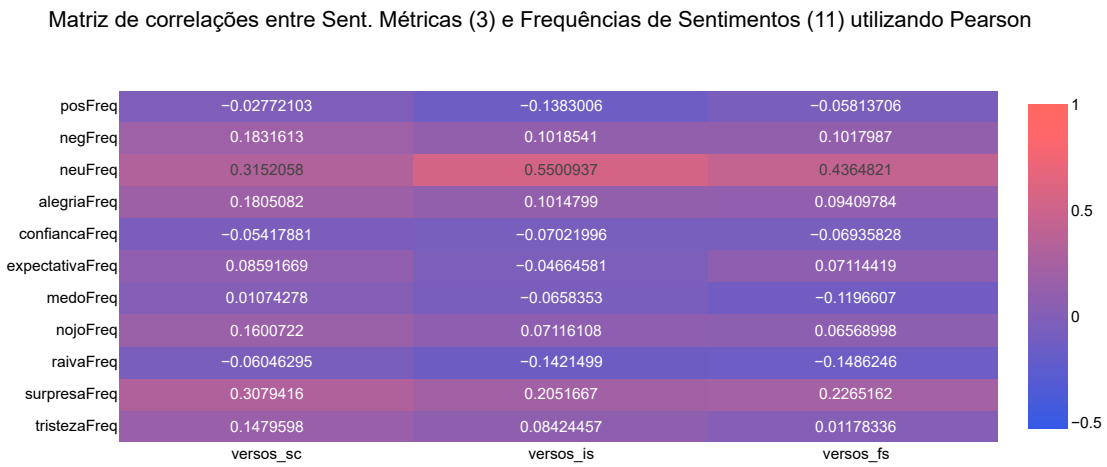


Figura 4.1: Matriz de correlação global de Pearson entre as características dos níveis "Métrico" e "Sentimental" com UT de tamanho 100.

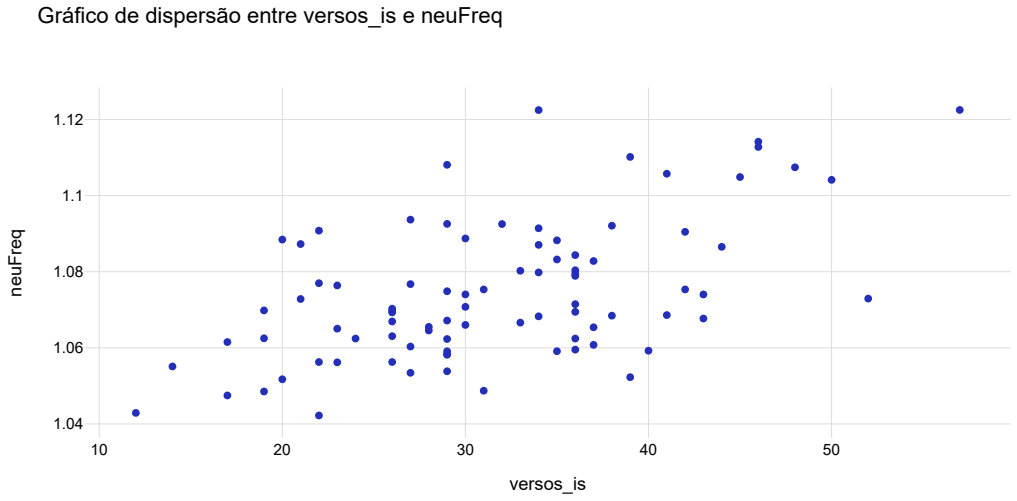


Figura 4.2: Dispersão entre o par de características "versos de início de sentença" e "frequência de neutros" com UT de tamanho 100

narrativa. Dessa forma, a utilização das janelas deslizantes para a realização de um cálculo de correlação segmentada permite a identificação de paralelismos que podem ocorrer em determinados trechos do livro.

Para que uma correlação seja identificada como um paralelismo ela precisa atingir um limiar mínimo absoluto. Levando em consideração o valor padrão de 0.75 para esse limiar, estabelecido com o intuito de identificar como paralelismos apenas as correlações altas ou fortes, excluindo relações mais fracas que podem ser consideradas irrelevantes, o par de características composto por "versos de início de sentença"(*versos_is*) e "frequência de neutros"(*neuFreq*) não é considerado um paralelismo, de acordo com o limiar proposto.

Para exemplificar a análise de correlação segmentada, calculada a partir de trechos do livro utilizando o método de janelas deslizantes, utilizamos o mesmo par de características, composto por "versos de início de sentença"(*versos_is*) e "frequência de neutros"(*neuFreq*). Com uma janela deslizante de tamanho 10 (composta por 10 UTs), sendo cada uma das UTs de tamanho 100 (que contém 100 sentenças cada, totalizando 1000 sentenças por janela) torna-se possível extrair a correlação segmentada. Essa análise segmentada permite visualizar que o coeficiente de correlação global com valor de 0.55 não se mantém consistente ao longo de toda a obra, como ilustrado na Figura 4.3[2].

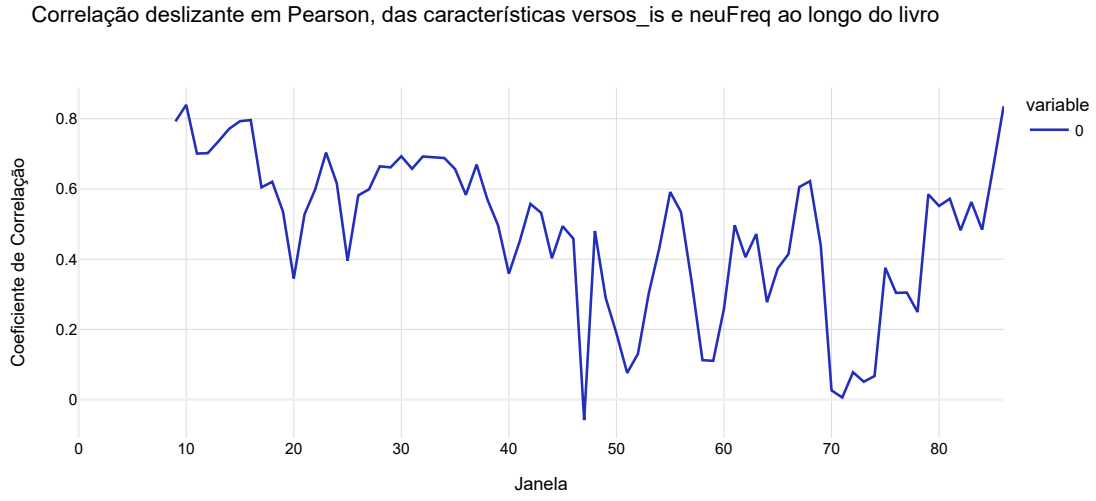


Figura 4.3: Correlação segmentada entre o par de características "versos de início de sentença" e "frequência de neutros" com UT de tamanho 100 e janela deslizante de tamanho 10

[2]A correlação segmentada começa a partir da janela 9 pois este valor está relacionado ao tamanho da janela, neste caso de tamanho 10. Sendo assim, o índice de início da correlação=tamanho da janela - 1

A Figura 4.3 exibe a correlação segmentada entre o par de características "versos de início de sentença" e "frequência de neutros" considerando uma UT de tamanho 100 e uma janela de tamanho 10. A partir dessa representação gráfica, é possível observar que, mesmo com uma correlação global de 0.55, a correlação segmentada revela que em determinados trechos o coeficiente atinge aproximadamente 0.86, sugerindo que a partir de uma correlação segmentada é possível encontrar trechos com paralelismos, considerando o limiar estabelecido em 0.75. Por outro lado, há também momentos em que o valor da correlação se aproxima de 0, indicando a existência de trechos sem nenhuma relação entre as características .

Esse tipo de análise permite a identificação de paralelismos em diferentes partes da obra e em pares de características que não necessariamente apresentam uma forte correlação global.

Dentre as correlações globais apresentadas na Figura 4.1 e que foram identificadas utilizando o método proposto existe a correlação entre as características "versos no final de sentença" e "frequência da emoção nojo". Ao aplicar a escala do método de correlação de Pearson, obteve-se um valor de coeficiente global equivalente a 0.0657, que se aproxima de 0. Esse valor reflete uma correlação quase nula, significando que não há uma relação entre as duas características quando considerada a narrativa completa, além disso, esse par não atende aos critérios de identificação de paralelismo, de acordo com o limiar estabelecido e as características de uma correlação forte.

É possível visualizar esse resultado em um gráfico de linhas, destacando as janelas nas quais foram identificadas correlações acima do limiar. Esse resultado está apresentado na Figura 4.4, onde são identificados quatro paralelismos, representados pelos pontos vermelhos e que correspondem às janelas nas quais foram identificadas correlações acima do limiar. Esses pontos estão localizados nas janelas de número 53[3], 57[4], 58[5] e 59[6], que correspondem aos trechos do livro em que esses paralelismos foram encontrados.

Ao analisar a correlação utilizando janelas deslizantes, é identificado o valor de 0.8656 na janela 58, representando o coeficiente de correlação máximo nesse contexto. Por outro lado, o coeficiente mínimo registrado para esse par de características foi de -0.6361 (janela 21), indicando que esta correlação varia consideravelmente em diferentes partes do texto, o que demonstra que a utilização do método computacional proposto pode ser utilizado para encontrar correlações em trechos do livro e indica que nem sempre as correlações (e consequentemente os paralelismos) podem ser identificadas utilizando a série completa (correlação global).

É relevante destacar que as janelas 57, 58 e 59 exibem paralelismos consecutivos, o que pode sugerir a presença de um trecho na obra onde o paralelismo é mais evidente

[3]Correspondente às sentenças de 4400 a 5399.
[4]Correspondente às sentenças de 4800 a 5799.
[5]Correspondente às sentenças de 4900 a 5899.
[6]Correspondente às sentenças de 5000 a 5999.

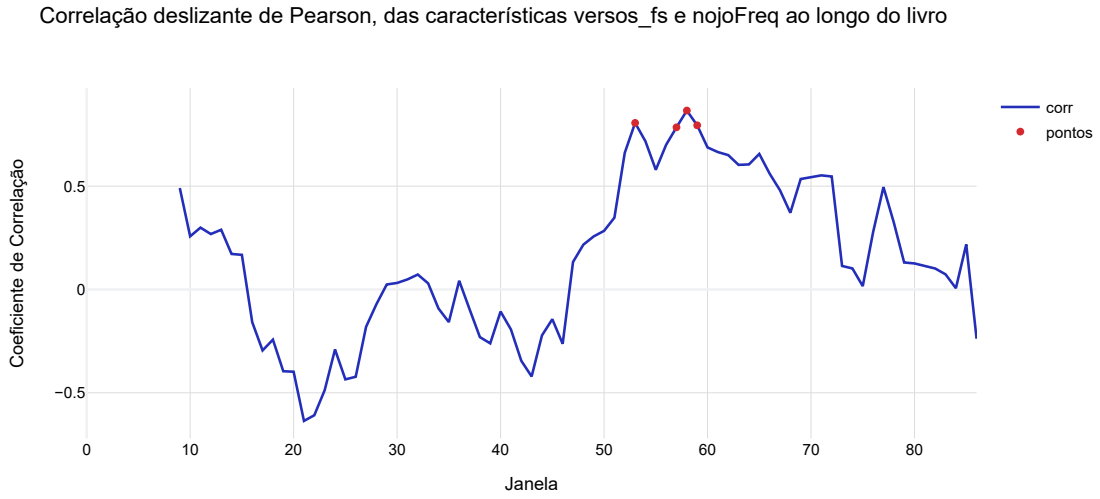


Figura 4.4: Correlação segmentada entre as características "versos de final de sentença" e "frequência de nojo" utilizando a correlação de Pearson com UT tamanho 100 e janela tamanho 10

e, consequentemente, mais facilmente identificável durante a leitura.

Ao analisar a janela com o valor máximo de correlação, que neste caso é a janela de número 58, pode-se constatar, por meio de um gráfico de dispersão, a presença de uma leve inclinação positiva entre os pares de características. Isso é evidenciado na Figura 4.5.

Além disso, é possível visualizar os valores dessas características nas UTs que compõem a janela 58 por meio de um gráfico de linhas, conforme representado na Figura 4.6. Na linha azul, encontram-se os valores correspondentes à característica "versos no final de sentença"(*versos_fs*) enquanto na linha vermelha estão os valores da característica "frequência da emoção nojo"(*nojoFreq*). A partir dessa figura, torna-se evidente a presença de uma correlação entre essas duas características, uma vez que os valores apresentam variações semelhantes, aumentando e diminuindo de maneira paralela.

Os resultados desta subseção destacam a capacidade do método computacional em identificar paralelismos em trechos específicos dos livros por meio da análise de correlação segmentada com o uso de janelas deslizantes. Além disso, fica evidente que a correlação global pode apontar potenciais pares de correlação para posterior análise, com a função do método computacional de identificá-los para facilitar a interpretação e análise dos resultados. Essa análise sublinha a importância de considerar tanto a correlação global quanto a correlação segmentada ao investigar paralelismos em textos literários, enriquecendo a compreensão das variações nos padrões textuais e ressaltando a utilidade do método computacional na identificação de paralelismos

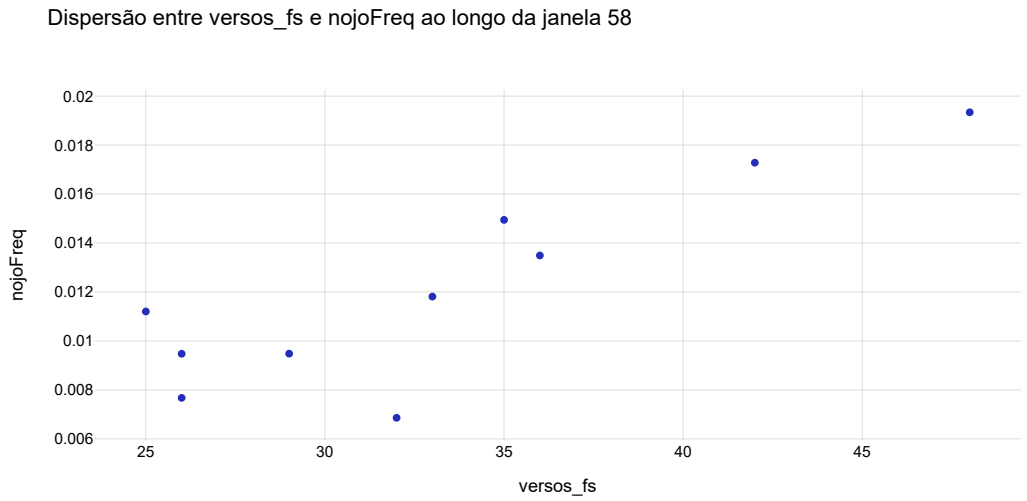


Figura 4.5: Gráfico de dispersão entre as características "versos de final de sentença"(*versos_fs*) e "frequência de nojo"(*nojoFreq*) das UTs componentes da janela 58

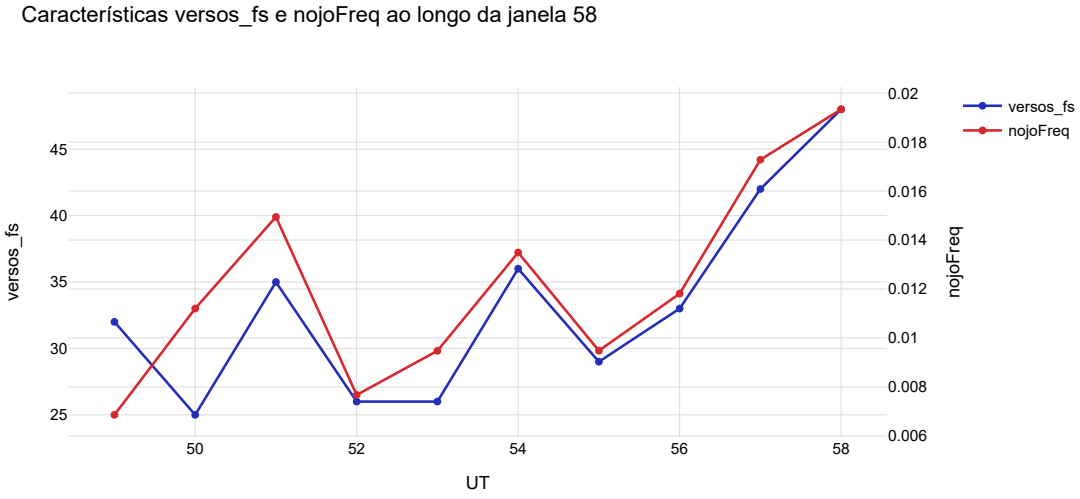


Figura 4.6: Gráfico de linhas entre as características "versos de final de sentença"(*versos_fs*) e "frequência de nojo"(*nojoFreq*) das UTs componentes da janela 58

que podem não ser prontamente evidentes em uma análise global do texto.

### 4.1.2 Correlação de Pearson e Correlação de Spearman

A variação entre as medidas de Pearson e Spearman, que são empregadas para identificar paralelismos, oferece a capacidade de capturar tanto relações lineares quanto monotônicas entre as características analisadas. Essa abordagem possibilita uma avaliação do método proposto quanto à variabilidade no cálculo da correlação.

O método de correlação de Pearson é utilizado para medir relações lineares entre variáveis. Ele é sensível a mudanças na magnitude e na direção das relações, o que o torna uma escolha adequada quando se deseja avaliar o grau de associação linear entre características textuais. Por outro lado, o método de correlação de Spearman é mais apropriado quando se busca capturar relações monotônicas, ou seja, relações que podem não ser estritamente lineares, mas ainda seguem um padrão consistente de aumento ou diminuição das características em questão.

Ao analisar essa variabilidade nos resultados, é possível visualizar a diferença entre as medidas para a análise de características textuais em diferentes contextos. Isso não apenas aprimora a compreensão das relações entre as características, mas também contribui para a validade e confiabilidade do método.

Para visualizar essa divergência entre as medidas é possível comparar a matriz de correlação representada na Figura 4.1 (que utiliza Pearson para calcular a correlação global entre os níveis Métrico e Sentimental em uma UT de tamanho 100) com a matriz de mesma configuração para a medida de Spearman, representada na Figura 4.7, onde é possível observar que em comparação com sua equivalente em Pearson alguns valores estão maiores e outros menores, representando a diferença entre as formas de cálculo de coeficiente.

O paralelismo da subseção anterior, identificado entre “versos de final de sentença” e “frequência de nojo”, representado na Figura 4.4, foi estabelecido com base no método de correlação de Pearson. No entanto, ao empregar o método de correlação de Spearman, os valores diferem. O coeficiente global para esse mesmo par de características utilizando Spearman é de 0.1755 enquanto com Pearson obteve-se 0.0657. O coeficiente mínimo de correlação segmentada utilizando Spearman é de -0.6991, enquanto com Pearson obteve-se -0.6361. Já o coeficiente máximo de correlação segmentada utilizando Spearman atinge o valor de 0.7903, enquanto com Pearson temos o valor de 0.8656.

O impacto na mudança de medida para o cálculo da correlação também pode ser observado na correlação segmentada. Ao utilizar a correlação de Spearman no par de características de “versos de final de sentença” e de “frequência de nojo” para a analise de correlação, o método computacional identificou apenas um trecho com ocorrência de paralelismo, considerando o limiar de 0.75. Esse paralelismo foi encontrado na janela de número 58, que é a mesma janela na qual é encontrado o paralelismo com o coeficiente mais alto usando Pearson.

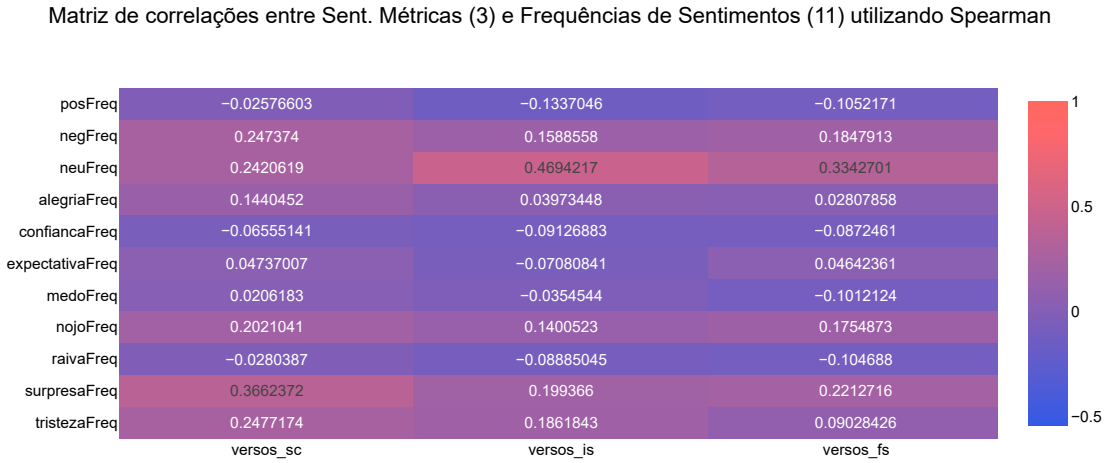


Figura 4.7: Matriz de correlação global de Spearman entre as características dos níveis Métrico e Sentimental com UT de tamanho 100

Com o intuito de visualizar a correlação segmentada desses pares a partir de janelas deslizantes e utilizando o método de correlação de Spearman, é possível observar, na Figura 4.8, o único coeficiente de correlação suficiente para a identificação como um paralelismo, de acordo com o limiar estabelecido.

Ao comparar as Figuras 4.4 e 4.8, torna-se evidente a diferença no uso de diferentes medidas de correlação. Utilizando o método de correlação de Pearson, o gráfico parece mais suavizado, e mais paralelismos foram encontrados em comparação com o método de correlação de Spearman.

É possível notar que os gráficos ainda exibem certa similaridade, especialmente entre as janelas 21 e 22, onde são observados baixos valores de correlação. Tanto com Pearson quanto com Spearman, os coeficientes variam entre -0.6 e -0.7 nessas janelas, evidenciando que mesmo com a utilização de medidas de correlação diferentes é possível a identificação de valores semelhantes para a correlação.

Com isso, a utilização de mais de uma medida de correlação pode evidenciar que alguns dos paralelismos identificados independem da medida utilizada e que podem ser paralelismos mais evidentes na obra. Enquanto outros podem ser identificados apenas em uma medida, essa possibilidade pode ser investigada utilizando o método computacional proposto.

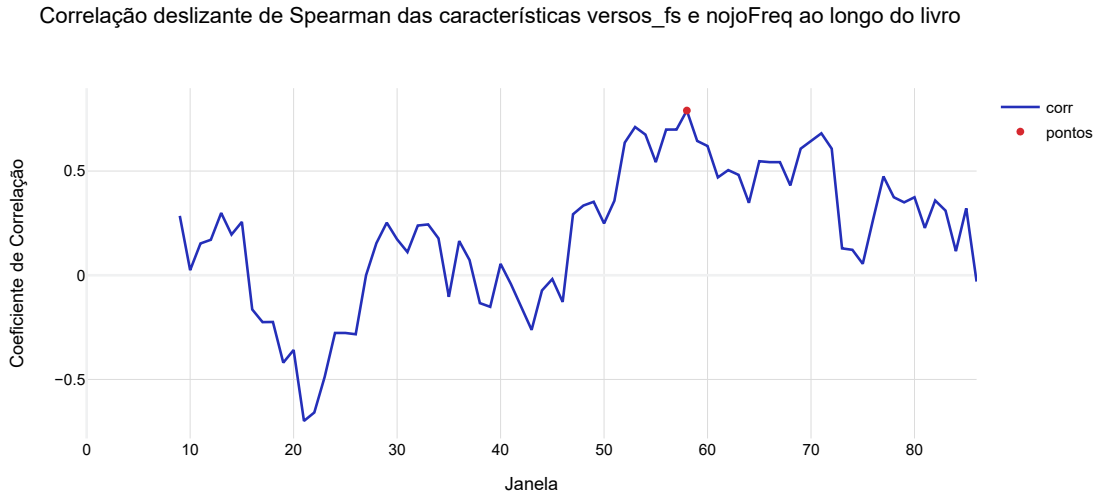


Figura 4.8: Correlação segmentada entre as características "versos de final de sentença" e "frequência de nojo" utilizando a correlação de Spearman com UT tamanho 100 e janela tamanho 10

### 4.1.3 Correlação utilizando a frequência absoluta e Correlação utilizando a frequência relativa

Frequentemente, nas pesquisas de análise textual, as características são quantificadas de forma absoluta, o que pode levar a distorções e desafios na interpretação dos resultados. Ao analisar, por exemplo, a quantidade de artigos (classe gramatical) em um texto literário, pode-se perceber que uma UT com 100 palavras tende a apresentar mais artigos do que uma UT com apenas 50 palavras.

Além disso, pode ser possível identificar uma alta correlação positiva entre o número de palavras neutras e o número de artigos em uma UT. No entanto, a associação entre o número de palavras neutras e o número de artigos pode ser uma associação espúria se não for considerado o número de palavras da UT como variável de confusão. UTs com mais palavras podem conter tanto mais palavras neutras quanto mais artigos, simplesmente porque há mais palavras no texto. Isso não significa necessariamente que há uma relação direta entre palavras neutras e artigos, mas sim que essa relação é influenciada pelo número de palavras da UT.

Diante desse cenário, essa seção aborda a utilização de frequências relativas em comparação com frequências absolutas na quantificação de características textuais utilizando o método computacional proposto. A adoção de frequências relativas permite uma abordagem mais equitativa, na qual as características são consideradas em relação à quantidade de palavras das UTs, possibilitando análises mais consistentes e comparáveis entre diferentes textos e evitando uma relação espúria causada pela influência do número de palavras na UT.

Para ilustrar a diferença entre frequências, é possível comparar os paralelismos identificados entre as características “artigos” e “palavras neutras” (sendo essas as frequências absolutas das características) com os paralelismos identificados entre as características “frequência de artigos” e “frequência de palavras neutras” (sendo essas as frequências relativas das características, com base na quantidade de palavras da UT). Utilizando o método de correlação de Pearson no livro *Os Sertões* e com UT de tamanho 100, obteve-se um coeficiente global de aproximadamente 0.9 para seus valores de frequência absoluta enquanto para seus valores de frequência relativa obteve-se um coeficiente global de aproximadamente 0.066. Isso representa um aumento de cerca de 1264% no coeficiente global de Pearson entre as frequências absolutas e relativas para esse par de características. Esse resultado demonstra como uma relação espúria devido à quantidade de palavras em uma UT pode ocorrer na identificação de paralelismos e influenciar os resultados obtidos através do método computacional proposto.

Essa diferença entre a utilização dos pares em frequências absolutas e a utilização dos pares em frequências relativas pode ser visualizada através do gráfico de dispersão representando essas correlações globais. A Figura 4.9 exibe a utilização do par de características “artigos”(*art*) e “palavras neutras”(*neutros*), como as características de frequência absoluta, enquanto a Figura 4.10 refere-se às frequências relativas das características “artigos”(*artFreq*) e “palavras neutras”(*neuFreq*).

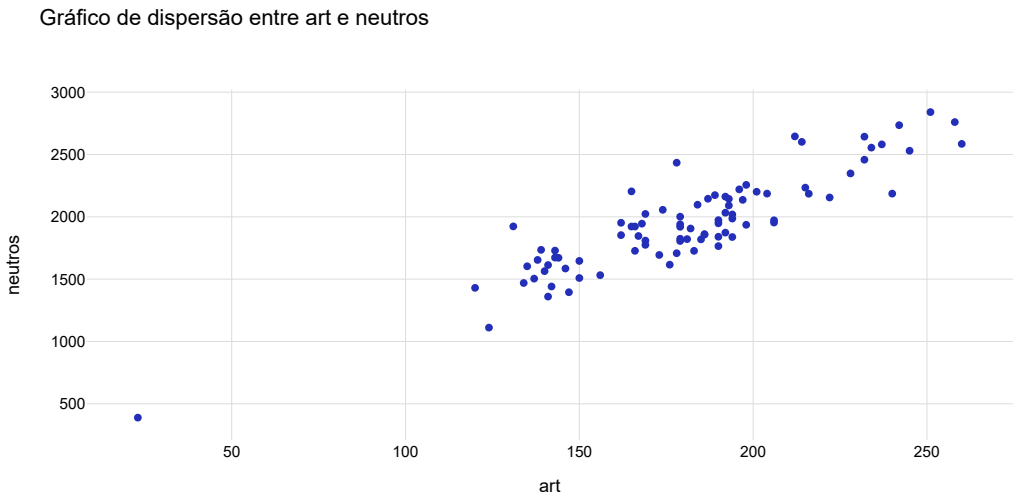


Figura 4.9: Gráfico de dispersão entre as características de frequência absoluta “artigos”(*art*) e “palavras neutras”(*neutros*)

A partir das Figuras 4.9 e 4.10 é possível observar que a correlação entre o par de características de frequência absoluta “artigos” e “palavras neutras” possuem uma correlação forte, enquanto com o par de características de frequência relativa “frequência de artigos” e “frequência de palavras neutras” não há uma correlação visível. Isso

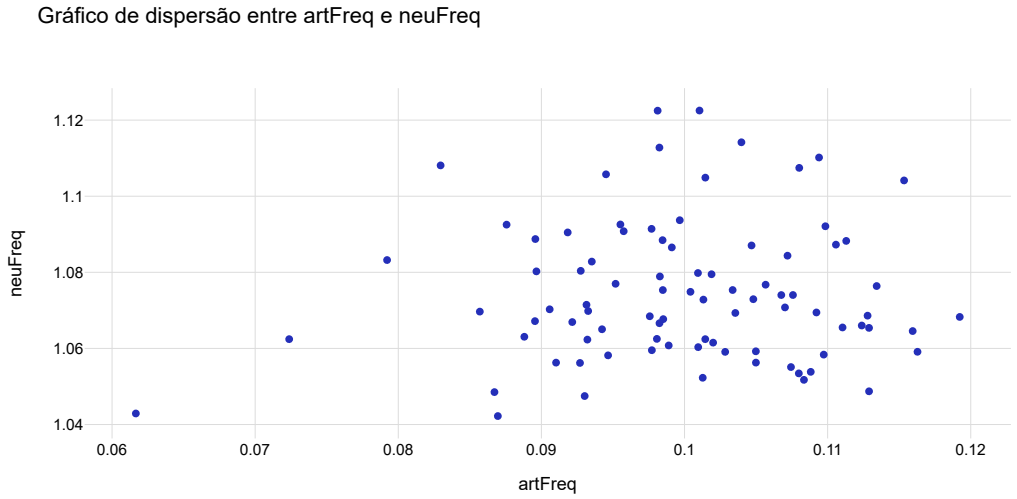


Figura 4.10: Gráfico de dispersão entre as características de frequência relativa "frequência de artigos"(*artFreq*) e "frequência de palavras neutras"(*neuFreq*)

reforça a ideia de uma relação espúria entre as características, causada por uma variável de confusão, no caso a quantidade de palavras da UT.

Essa relação pode ser melhor demonstrada ao analisar a identificação de paralelismos a partir da correlação segmentada a partir de janelas de tamanho 10 (com UTs com tamanho de 100 sentenças o que corresponde a um máximo de 1000 sentenças por janela), conforme as Figuras 4.11 e 4.12 que representam, a partir do método de correlação de Pearson, respectivamente, a correlação segmentada utilizando a frequência absoluta e a frequência relativa das características.

Como observado na Figura 4.12, utilizando o método de correlação de Pearson não são encontrados paralelismos entre o par de características "frequência de artigos" e "frequência de palavras neutras", que representam a frequência relativa das características, pois não há nenhum valor absoluto de correlação segmentada que atenda o valor de coeficiente de correlação mínimo de 0.75, estabelecido a partir do limiar. Já utilizando o método de correlação de Pearson para o par de características de frequência absoluta "artigos" e "palavras neutras" são identificados 62 paralelismos, conforme a Figura 4.11.

Conforme exposto no início dessa subseção, as características quantificadas em valores absolutos podem causar a identificação de paralelismos enganosos ao estabelecer relações espúrias causadas pela quantidade de palavras da UT. Entretanto, essas características não foram removidas do método computacional pois podem ser utilizadas na identificação de paralelismos entre outros níveis. Dessa forma, a utilização das características em sua frequência relativa podem servir para validar a identificação de paralelismos ao estabelecer correlações mais confiáveis para o método

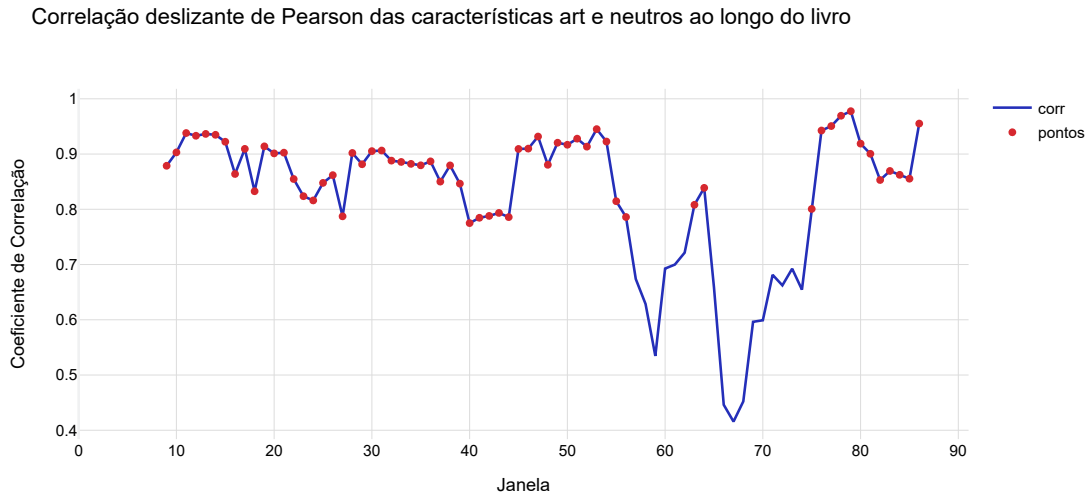


Figura 4.11: Correlação segmentada entre as características de frequência absoluta "artigos" e "palavras neutras" utilizando o coeficiente de Pearson com tamanho da UT=100 e tamanho da janela=10

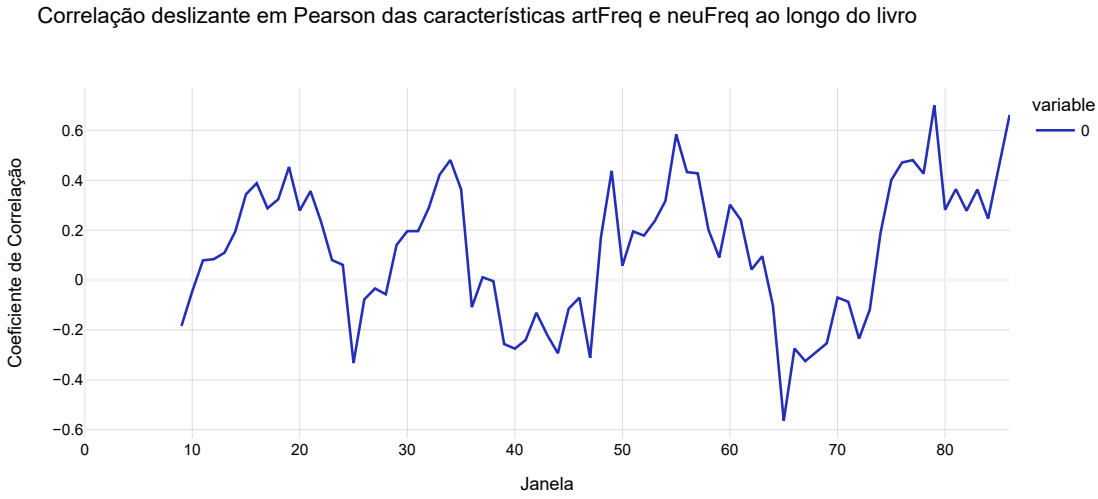


Figura 4.12: Correlação segmentada entre as características de frequência relativa "frequência de artigos" e "frequência de palavras neutras" utilizando o coeficiente de Pearson com tamanho da UT=100 e tamanho da janela=10

computacional proposto.

## 4.2 Influência da variação dos parâmetros nos resultados

A função principal do método computacional proposto é a identificação de paralelismos, a qual possui três parâmetros configuráveis: o tamanho da Unidade Textual (UT), o limiar para a identificação de paralelismos com base nos coeficientes de correlação e o tamanho da janela deslizante.

Unidades Textuais (UTs) representam segmentos do texto em sentenças. Uma UT de tamanho 1 corresponde a uma única sentença, enquanto uma UT de tamanho 100 abrange um conjunto de 100 sentenças contíguas. As janelas deslizantes consistem em agrupamentos de UTs maiores que percorrem o texto, permitindo uma análise das variações das características textuais em diferentes partes do texto. Quando configuramos o tamanho da janela para 10 UTs e o tamanho das UTs para 100 sentenças, cada janela abrangerá 1000 sentenças consecutivas.

A mudança em qualquer um desses tamanhos resulta em conjuntos de resultados distintos com base nas escolhas de valores feitas para esses parâmetros.

Para avaliar as variações dos parâmetros e seus impactos na pesquisa proposta, a Tabela 4.1 demonstra numericamente como a variação do tamanho da UT em conjunto com o tamanho da janela afeta a quantidade de pares de características que possuem valores iguais ou acima do limiar, obtidos a partir do livro *Os Sertões*. Nesse caso, o limiar estabelecido foi o limiar padrão de 0.75 utilizando o coeficiente de Pearson e a correlação segmentada.

Tabela 4.1: Quantidade de pares de características onde foram identificados paralelismos de acordo com diferentes valores dos parâmetros.

| **Tamanho da UT** | | **1** | **50** | **100** |
|---|---|---|---|---|
| **Tamanho da Janela** | 10 | 3216 | 3371 | 2588 |
| | 12 | 3288 | 2604 | 1751 |
| | 15 | 3226 | 1569 | 913 |

Através da análise da Tabela 4.1, é evidente que o aumento no tamanho da UT leva a uma redução na quantidade de resultados obtidos. Além disso, o aumento no tamanho da janela deslizante também ocasiona uma menor quantidade de resultados.

Outra maneira de visualizar a variação dos resultados pode ser observada na Tabela 4.2, que representa a mesma combinação de parâmetros, mas desta vez exibindo a soma da quantidade de trechos (janelas) onde foram identificados paralelismos no livro *Os Sertões* e encontrados nas janelas considerando todos os pares de características. Esta variação também ocorre devido à mudança do tamanho da janela e do tamanho das UTs.

Tabela 4.2: Quantidade de trechos(janelas) onde foram identificados paralelismos de acordo com diferentes valores dos parâmetros.

| **Tamanho da UT** | | **1** | **50** | **100** |
|---|---|---|---|---|
| **Tamanho da Janela** | **10** | 1474278 | 36704 | 16616 |
| | **12** | 1169761 | 27119 | 11939 |
| | **15** | 924116 | 19718 | 8279 |

A partir da Tabela 4.2, é possível observar que assim como na Tabela 4.1, o aumento do tamanho das UTs e o aumento do tamanho da janela diminuem a quantidade de paralelismos identificados.

Além disso, a variação do parâmetro correspondente ao limiar absoluto também exerce influência na quantidade de paralelismos identificados, como pode ser observado na Tabela 4.3. Nessa tabela, fica evidente que o aumento do limiar leva a uma diminuição na quantidade de paralelismos encontrados. Essa redução ocorre devido à maior exigência imposta para que uma correlação seja reconhecida como um paralelismo. Vale ressaltar que, nos resultados apresentados, as variáveis tamanho das UTs e tamanho da janela foram mantidas fixas em 100 e 10, respectivamente.

Tabela 4.3: Quantidade de pares de características onde foram encontrados paralelismo e quantidade de trechos com paralelismos identificados de acordo com diferentes valores para o limiar absoluto, utilizando o tamanho das UTs em 100 e o tamanho das janelas em 10.

| | **0.75** | **0.85** | **0.95** |
|---|---|---|---|
| **Pares de características** | 2588 | 1123 | 146 |
| **Trechos com paralelismos** | 16616 | 5759 | 1144 |

Conforme as Tabelas 4.1 e 4.2, ao utilizar o tamanho da UT de 100 e o tamanho da janela de 10, para o livro *Os Sertões*, é possível identificar um total de 2588 pares de características que, em algum momento, atenderam ao limiar estabelecido em 0.75 na escala de coeficientes de correlação, sendo que a quantidade de paralelismos que foram encontrados nesses pares foi de 16616.

Dessa forma, é evidente que a escolha dos parâmetros, como o tamanho da Unidade Textual (UT) e o tamanho da janela, tem um impacto significativo na identificação de paralelismos por meio do método computacional proposto. Para exemplificar e detalhar como essas variações influenciam os resultados, as próximas seções abordarão dois aspectos específicos: a variação no tamanho da UT e a variação no tamanho da janela. Isso permitirá uma análise das implicações das escolhas de parâmetros na identificação de paralelismos.

### 4.2.1 Variação no tamanho da UT

A variação do tamanho da UT é um aspecto a ser considerado na análise das características extraídas dos textos literários. Isso ocorre porque o tamanho da UT pode impactar a análise de características textuais, pois a quantificação das características pode variar com base no tamanho da UT.

Portanto, ao explorar a variação do tamanho da UT, o objetivo é entender como essa variação influencia as correlações e os paralelismos identificados nos pares de características textuais.

Essa abordagem permite evitar conclusões inadequadas e validar as correlações identificadas em uma análise internível dos textos literários.

Ao reduzir o tamanho da UT de 100 para 50, mantendo o tamanho da janela em 10 e o limiar em 0.75, é possível identificar 3371 pares de características nos quais foram encontrados paralelismos. A quantidade total de paralelismos identificados nesses pares é de 36704.

Com a configuração de parâmetros mencionada, também são identificados paralelismos entre as características "versos de final de sentença" e "frequência de nojo". No entanto, os valores e quantidades encontrados são diferentes em comparação com a configuração anterior. Ao utilizar a UT de tamanho 100 com o método de correlação de Pearson, os valores do coeficiente de correlação global e os coeficientes mínimo e máximo de correlação segmentada são 0.0657, -0.6361 e 0.8656, respectivamente. Por outro lado, ao utilizar a UT de tamanho 50, esses valores são de 0.1479, -0.6114 e 0.7559, respectivamente. Com a UT de tamanho 100, são identificados 4 trechos com paralelismos entre estas características, localizados nas janelas de número 53, 57, 58 e 59, como demonstrado na Figura 4.4. Já com a UT de tamanho 50, são encontrados 3 trechos com paralelismos, conforme ilustrado na Figura 4.13, que correspondem às janelas de número 113, 114 e 117.

Utilizando a UT de tamanho 50, os trechos com paralelismos identificados encontram-se nas janelas de número 113, 114 e 117, que englobam as sentenças de número 5200 a 5699 na janela de número 113, de 5250 a 5749 na janela de número 114, e de 5400 a 5899 na janela de número 117. Por outro lado, ao utilizar a UT de tamanho 100, os trechos com paralelismo identificado são encontrados nas janelas de número 53, 57, 58 e 59, com as sentenças correspondentes aos intervalos de 4400 a 5399 na janela 53, de 4800 a 5799 na janela de número 57, de 4900 a 5899 na janela de número 58, e de 5000 a 5999 na janela de número 59. Com essa observação, é possível notar a semelhança na localização dos trechos onde esses paralelismos são identificados, tanto com uma UT de tamanho 50 quanto com uma UT de tamanho 100, demonstrando que, nesse caso, os paralelismos ocorrem em locais correspondentes, indicando uma consistência na identificação desses padrões, independentemente da variação desse parâmetro.

Para averiguar se essa semelhança de localização entre os trechos com paralelismo identificado se mantém, utilizando diferentes tamanhos de UT, independentemente

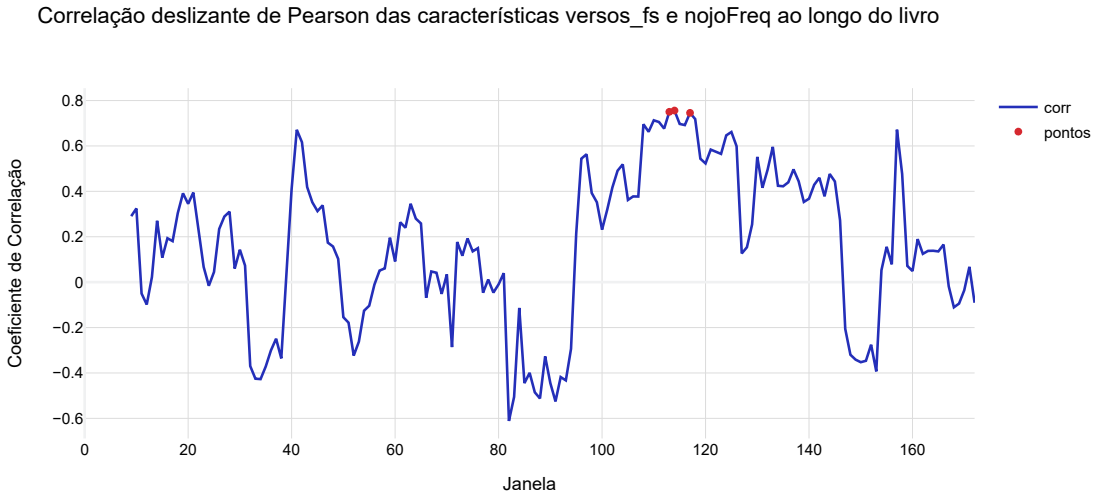


Figura 4.13: Correlação segmentada entre as características “versos de final de sentença” e “frequência de nojo” utilizando a correlação de Pearson com UT de tamanho 50 e janela de tamanho 10

da mudança no método de correlação utilizado, exploraremos a análise de correlação de Spearman nas mesmas características de “versos de final de sentença” e “frequência de nojo”, variando os tamanhos das UTs em 100 e 50.

O paralelismo entre as características “versos de final de sentença” e “frequência de nojo” também foi identificado utilizando o método de correlação de Spearman (conforme exposto na Figura 4.8). Ao utilizar UTs de tamanho 100 com Spearman, o valor do coeficiente global e os valores dos coeficientes mínimo e máximo da correlação segmentada foram 0.1755, -0.6991 e 0.7903, respectivamente. Já com UTs de tamanho 50 com Spearman, esses valores foram de 0.2044, -0.5775 e 0.7683, respectivamente, indicando uma proximidade entre esses coeficientes utilizando diferentes tamanhos de UT.

Com UTs de tamanho 100, é identificado apenas 1 paralelismo, como evidenciado na Figura 4.8 e localizado na janela 58. Por outro lado, com UTs de tamanho 50, são encontrados 2 paralelismos, localizados nas janelas de número 115 e 145, como ilustrado na Figura 4.14.

Para UT com tamanho 50, os paralelismos entre as características “versos de final de sentença” e “frequência de nojo” são identificados nos trechos 115 e 145, onde as sentenças da janela 115 correspondem aos números de sentença de 5300 a 5799, enquanto as sentenças da janela 145 vão de 6800 a 7299.

Esse mesmo paralelismo ocorre na janela de número 58 para UT de tamanho 100, que corresponde às sentenças de número 4900 a 5899. Com isso, há uma semelhança entre as duas variações de tamanho de UT, em que as sentenças de 5300 a 5799 são

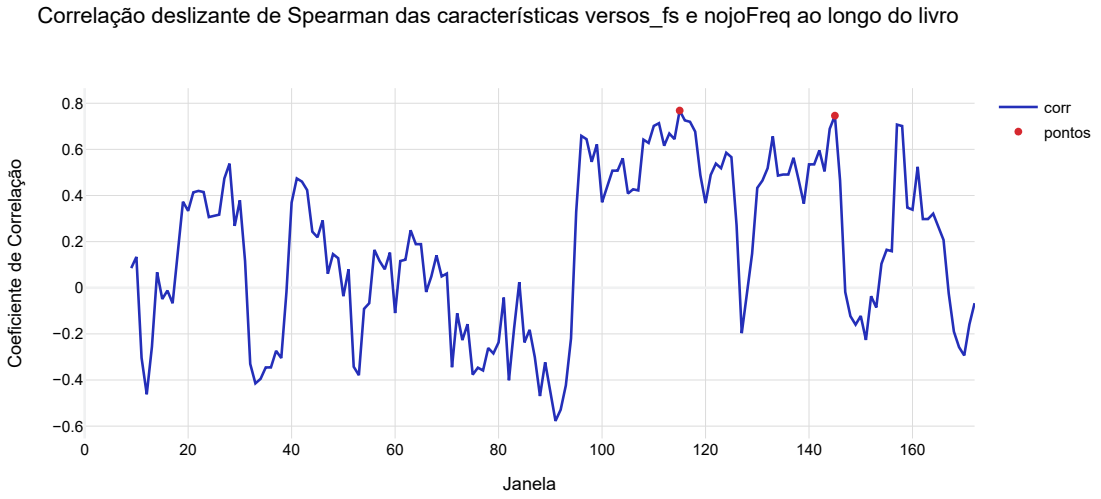


Figura 4.14: Correlação segmentada entre as características "versos de final de sentença" e "frequência de nojo" utilizando a correlação de Spearman com UT de tamanho 50 e janela de tamanho 10

comuns a ambas as configurações. Esse intervalo em comum é exatamente a janela de número 115 para UT de tamanho 50.

Entretanto, embora observado resultados semelhantes entre as variações no tamanho da UT nesses casos, não é possível afirmar que esse padrão sempre ocorrerá, visto que alguns trechos com paralelismos que foram identificados utilizando UT de tamanho 100 não foram encontrados com UT de tamanho 50 e vice-versa, portanto, mudanças no tamanho das UTs podem levar a variações nos resultados.

### 4.2.2 Variação no tamanho da janela

A variação do tamanho da janela pode influenciar nos resultados, afetando a detecção de paralelismos e as correlações entre características textuais. Ao explorar essa variação, é possível avaliar a sensibilidade do método a diferentes configurações e identificar diferenças nos resultados ao variar o tamanho da janela. Além disso, ao compreender como o tamanho da janela impacta os resultados, é possível tomar decisões sobre a configuração do método para análises textuais específicas, tornando-o mais flexível e adaptável.

Ao utilizar a janela de tamanho 15, por exemplo, também é possível identificar o paralelismo entre "versos de final de sentença" e "frequência de nojo" e visualizar sua variação causada pelo aumento no tamanho da janela.

Utilizando o tamanho da janela em 15 com as UTs de tamanho 100, são identificados trechos com paralelismo entre essas características utilizando tanto a correlação de

Pearson quanto a de Spearman, com algumas diferenças encontradas em relação ao tamanho da janela em 10 com as UTs de tamanho 100.

Enquanto com tamanho da janela em 10 são encontrados 4 paralelismos com Pearson, localizados nas janelas 53, 57, 58 e 59 (vide Figura 4.4) e 1 paralelismo com Spearman, localizado na janela 58 (vide Figura 4.8), utilizando o tamanho da janela em 15 é encontrado 1 paralelismo com Pearson e 1 com Spearman, conforme Figuras 4.15 e 4.16, ambos localizados na janela 58.

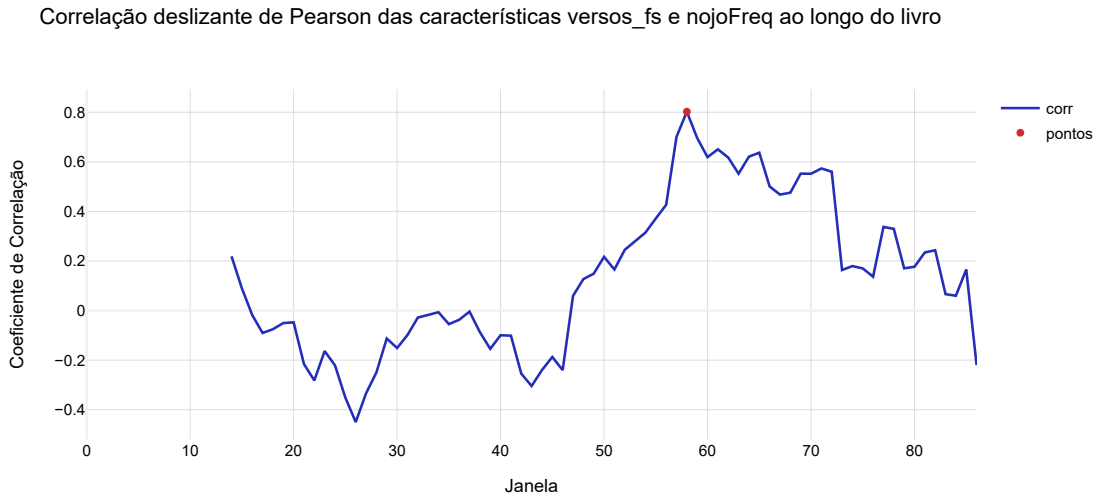


Figura 4.15: Correlação segmentada entre as características “versos de final de sentença” e “frequência de nojo” utilizando a correlação de Pearson com UT de tamanho 100 e janela de tamanho 15

Conforme observado nas Figuras 4.15 e 4.16, tanto com a correlação de Pearson quanto com a de Spearman, o trecho com paralelismo entre as características “versos de final de sentença” e “frequência de nojo” é a janela de número 58 que compreende as UTs de 44 a 58 com suas sentenças de 4400 a 5899.

É possível observar que a janela de número 58 também apresenta paralelismos identificados utilizando o valor de 10 para o tamanho da janela. No entanto, mesmo se tratando de valores de parâmetro diferentes, as janelas compreendem algumas sentenças iguais. Conforme mencionado anteriormente, para o tamanho 15, a janela compreende as sentenças de 4400 a 5899, enquanto para o tamanho 10, essa mesma janela compreende as UTs com as sentenças de 4900 a 5899. Isso demonstra, mais uma vez, que apesar de influenciar nos resultados, a variação de parâmetros também pode ser utilizada para comprovar a correlação e consequentemente o paralelismo encontrado.

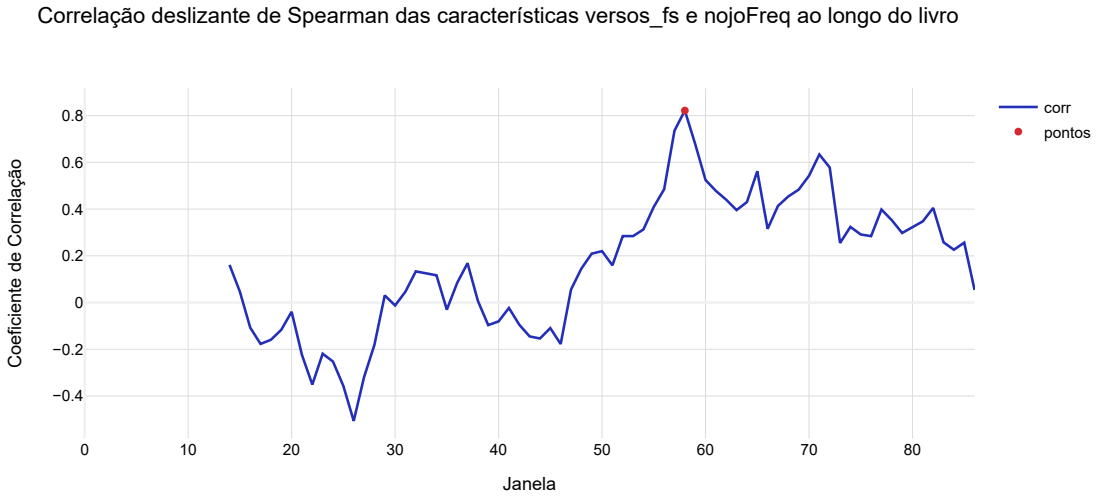


Figura 4.16: Correlação segmentada entre as características "versos de final de sentença" e "frequência de nojo" utilizando a correlação de Spearman com UT de tamanho 100 e janela de tamanho 15

## 4.3 Diversidade de paralelismos interníveis identificados

A variação dos parâmetros resulta em uma ampla gama de resultados diferentes, o que significa que os paralelismos encontrados podem variar significativamente entre diferentes tamanhos de UTs e tamanhos de janelas.

Dessa forma, para fins de verificação dos trechos com paralelismos identificados, os parâmetros serão fixados em 100 para o tamanho da UT, 10 para o tamanho da janela e 0.75 para o limiar. Nesta seção, o livro utilizado continuará sendo *Os Sertões*.

O objetivo dessa seção é demonstrar e exemplificar a amplitude de paralelismos interníveis que podem ser identificados pelo método proposto visando ilustrar a diversidade de paralelismos presentes no livro *Os Sertões* e como o método computacional pode ser utilizado para identificá-los.

Um dos paralelismos identificados entre o nível "Fonético" e o nível "Métrico" ocorre entre características de "repetições silábicas 3S4L"(*repSil_3_4*) e "versos em sentenças completas"(*versos_sc*). Esse paralelismo chama atenção, pois, ao utilizar o método de correlação de Spearman, o coeficiente global obtido foi de -0.1834, uma correlação quase nula, enquanto com a utilização da correlação segmentada foram identificados 4 trechos com paralelismos que possuem coeficientes acima do limiar de 0.75, todos eles correlações negativas, conforme a Figura 4.17.

Na Figura 4.17, é possível identificar que os trechos com paralelismo ocorrem nas

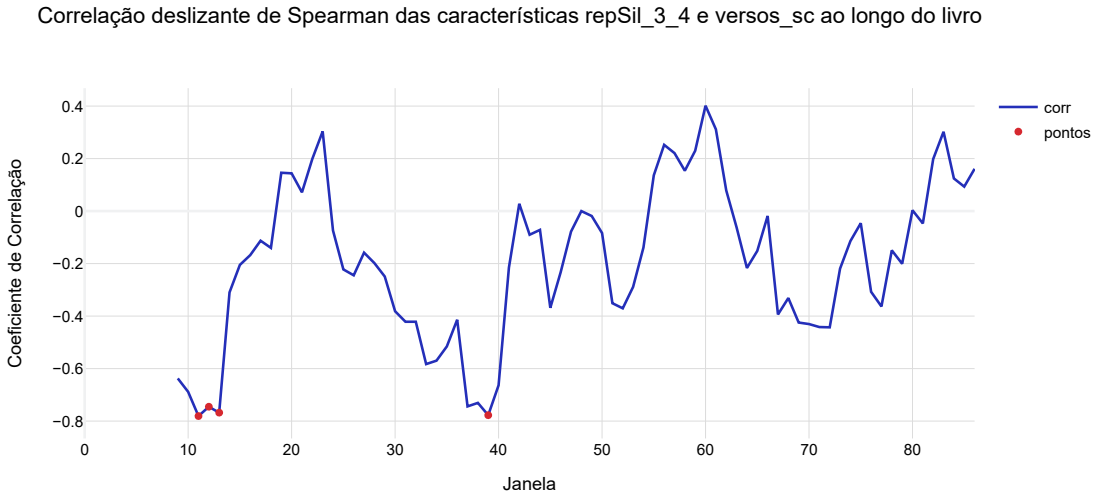


Figura 4.17: Correlação segmentada entre as características “repetições silábicas 3S4L” e “versos de sentença completa” utilizando a correlação de Spearman com UT de tamanho 100 e janela de tamanho 10

janelas de número 11, 12, 13 e 39, que ao serem considerados seus valores absolutos possuem um valor de coeficiente de correlação igual ou acima do limiar estabelecido em 0.75.

Como os coeficientes dessas janelas apresentam correlações negativas eles indicam que ao acontecer um aumento na quantidade de “repetições silábicas 3S4L” há uma diminuição na quantidade de “versos de sentença completa” e quando há uma diminuição na quantidade de “repetições silábicas 3S4L” há um aumento na quantidade de “versos de sentença completa”. Esse fenômeno pode ser observado na Figura 4.18, que demonstra os valores das características nas UTs ao longo da janela 11.

A identificação dessa correlação permite reconhecer que existem paralelismos entre as características identificadas nos trabalhos de Carvalho et al. (2020) e (Lima et al., 2021) e pode indicar que as características desses níveis têm alguma relação ou paralelismo.Essa indicação possibilita estudos futuros acerca desse fenômeno utilizando essas características e o método computacional aqui proposto para investigar essa hipótese.

Outro paralelismo internível que em sua correlação global apresenta um coeficiente próximo de 0 foi identificado entre as características “TTR de palavras” e “frequência de alegria” dos níveis “Léxico” e “Sentimental”, respectivamente.

Utilizando o método de correlação de Spearman, o coeficiente de correlação global dessas características é de 0.03, caracterizando uma correlação nula. Entretanto, ao utilizar a correlação segmentada são encontradas correlações fortes positivas no início

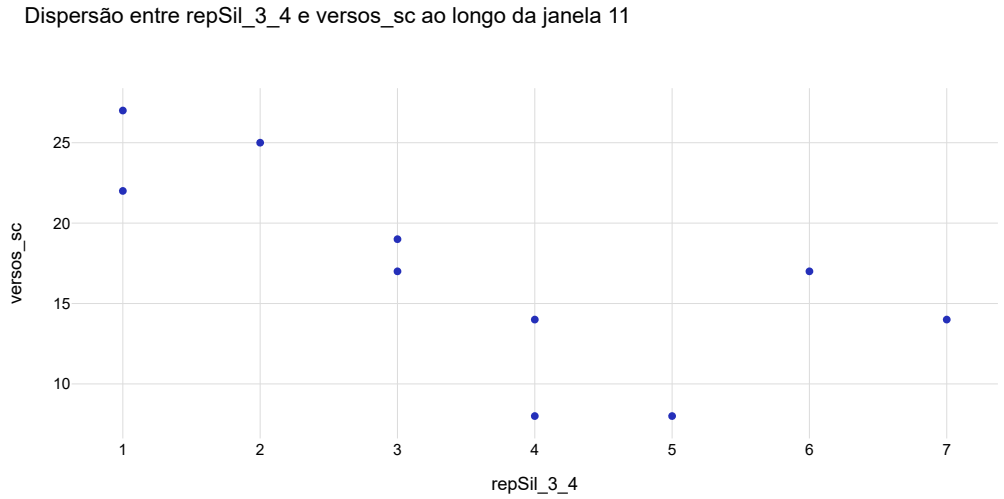


Figura 4.18: Gráfico de dispersão entre as características "repetições silábicas 3S4L"(*repSil_3_4*) e "versos em sentenças completas"(*versos_sc*) ao longo da janela 11

do livro e correlações fortes negativas no final, apesar da variação desses valores ao longo do livro, conforme exposto na Figura 4.19.

A Figura 4.19 demonstra que existem paralelismos em trechos diferentes do livro e em trechos seguidos, o que pode ser algo intencional ou não, já que a correlação não indica causalidade.

Nesse caso, a dimensão dada a partir da utilização das frequências relativas é contrária à ideia apresentada pela utilização de suas frequências absolutas, já que ao utilizar esses valores as fortes correlações negativas encontradas no final do livro se tornam fortes correlações positivas a partir de sua frequência absoluta, conforme a Figura 4.20.

Este exemplo reafirma um paralelismo no trecho em comum, já que mesmo a mudança nos valores de correlação e a utilização da frequência absoluta, em vez da frequência relativa, não afetaram a identificação desse fenômeno nos trechos finais do livro. Sendo assim, a utilização da frequência relativa pode ser aplicada para verificar um paralelismo encontrado com a sua frequência absoluta, possibilitando uma visão diferente dos resultados em alguns casos.

A utilização da modelagem de tópicos possibilita uma visão diferente dos resultados que envolvem as características nesse nível de extração. Dentre os paralelismos identificados pelo método computacional utilizando o nível "Tópicos", temos o presente entre as características "versos de sentença completa" (do nível "Métrico") e "Tópico 2".

Esse paralelismo encontrado exemplifica que mesmo com as altas contagens de 0 para

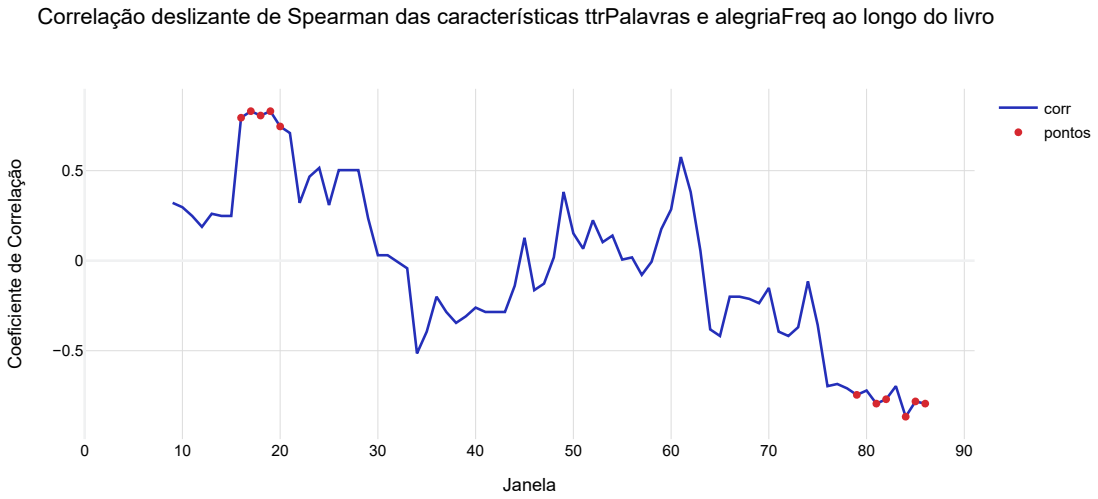


Figura 4.19: Correlação segmentada entre as características "TTR de palavras" e "frequência de alegria" utilizando o método de correlação de Spearman com UT de tamanho 100 e janela de tamanho 10

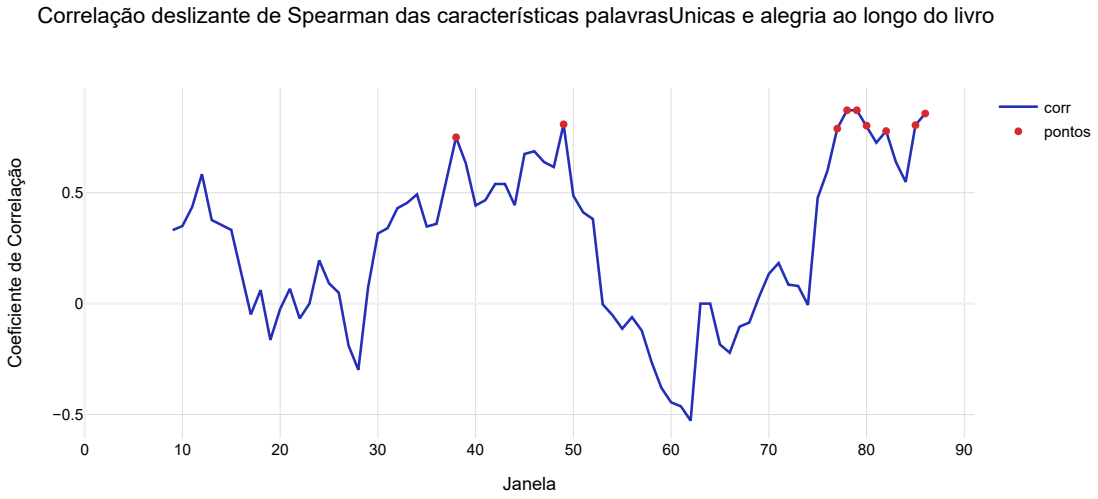


Figura 4.20: Correlação segmentada entre as características "palavras únicas" e "alegria" utilizando o método de correlação de Spearman com UT de tamanho 100 e janela de tamanho 10

as características tópicos (causada pela relevância do tópico em determinadas UTs) foi possível encontrar associações entre os níveis "Métrico" e "Tópicos" a partir das características "versos de sentença completa" e "Tópico 2", conforme demonstrado na Figura 4.21.

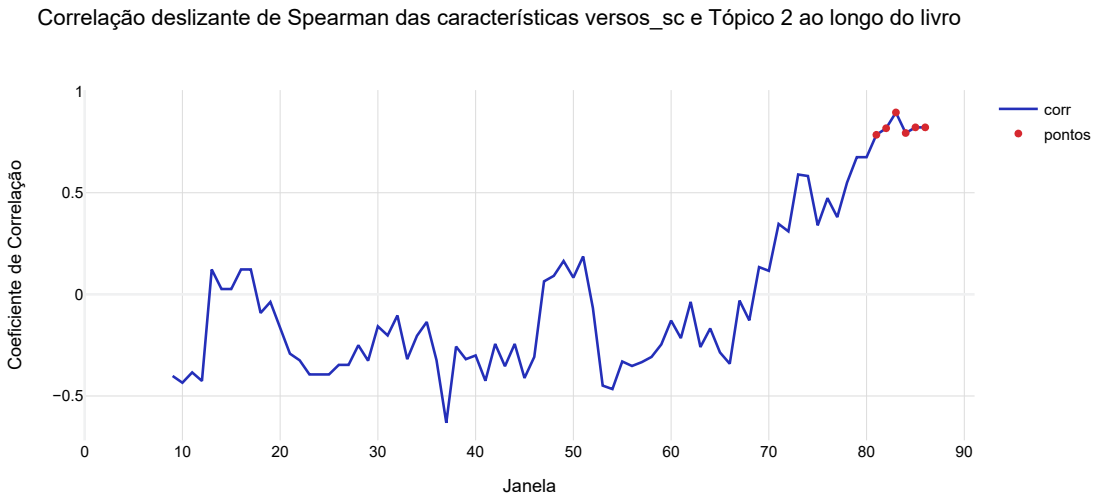


Figura 4.21: Correlação segmentada entre as características "versos de sentença completa" e "Tópico 2" com UT de tamanho 100 e janela de tamanho 10

Essa extração pode ser visualizada tanto pelo gráfico de dispersão da Figura 4.22 quanto pelo gráfico de linhas da Figura 4.23, que demonstram, por meio de gráficos distintos, os valores presentes na janela de número 83, onde foi encontrado o ponto máximo de correlação. Isso sugere que, em alguns casos, a presença de valores iguais a 0 para a relevância dos tópicos não impacta nos resultados obtidos a partir do método computacional proposto.

Com isso, as correlações e, consequentemente, os paralelismos encontrados utilizando as características do nível "Tópicos" também podem ser objetos de estudos futuros, mesmo com as limitações encontradas nesse nível de característica.

As características extraídas do nível "Entidades Nomeadas" também podem ter paralelismos associados a outros níveis extraídos, a exemplo do paralelismo encontrado a partir das características "frequência de Entidades Nomeadas" e "frequência de raiva", sendo esta última extraída do nível "Sentimental".

A Figura 4.24 demonstra os paralelismos encontrados a partir das características "frequência de Entidades Nomeadas" e "frequência de raiva", onde é possível observar a existência de 8 trechos com paralelismo. Esses trechos são marcados por correlações negativas nas janelas de 32 a 37 e correlações positivas nas janelas 65 e 66.

Sendo assim, esse resultado identifica uma relação variável entre essas características

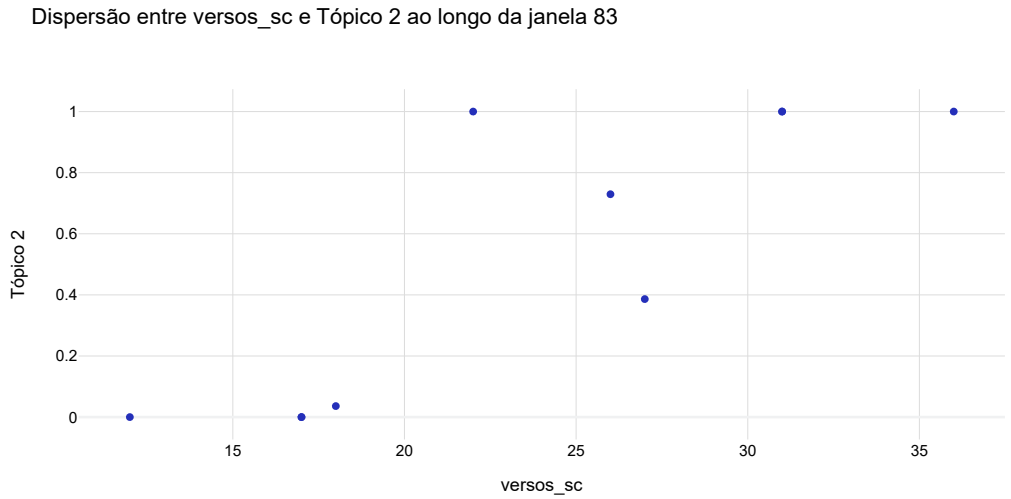


Figura 4.22: Gráfico de dispersão entre as características "versos de sentença completa"(*versos_sc*) e "Tópico 2" na janela de índice 83

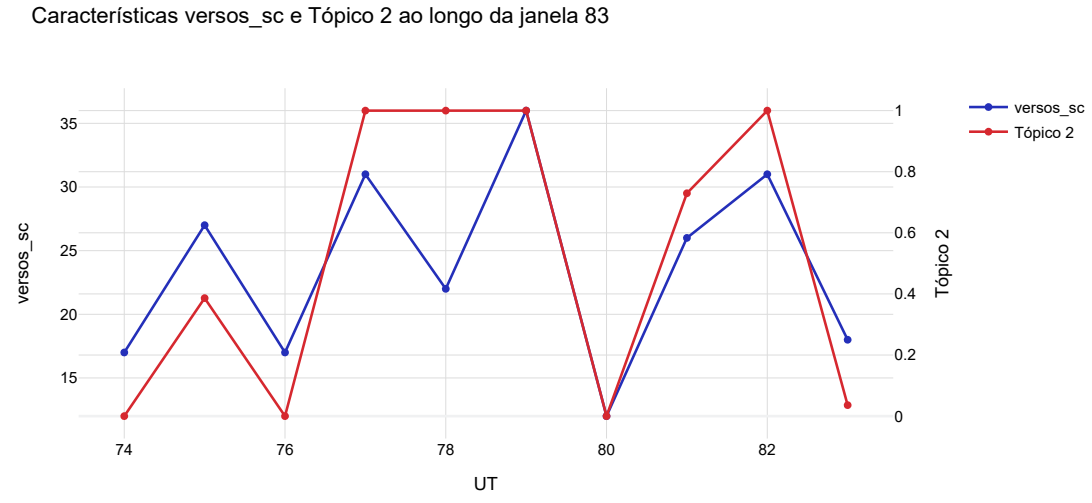


Figura 4.23: Gráfico de linhas entre as características "versos de sentença completa"(*versos_sc*) e "Tópico 2" na janela de índice 83

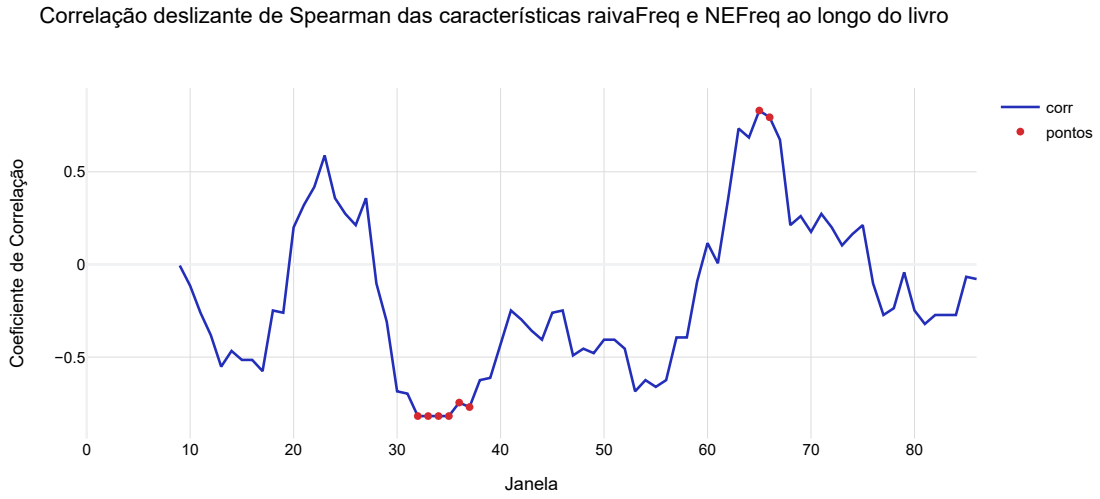


Figura 4.24: Correlação segmentada entre as características "frequência de Entidades Nomeadas" e "frequência de raiva" com UT de tamanho 100 e janela de tamanho 10

e que pode ser averiguada a partir da visualização dos trechos em que o paralelismo ocorre através do método computacional proposto.

Esse resultado, somado aos demais paralelismos apresentados nessa seção servem como exemplos dos resultados que podem ser obtidos através da aplicação do método computacional proposto, demonstrando certa diversidade nos resultados que podem ser extraídos e estudados futuramente a partir desse método computacional. Isso contribui para uma compreensão mais abrangente e aprofundada da aplicabilidade do método na análise de textos literários, destacando sua capacidade de revelar padrões interdimensionais que podem não ser prontamente perceptíveis em uma análise convencional. Essa seção também ajuda a consolidar a utilidade do método na pesquisa em linguística computacional e estudos literários, ao mostrar como ele pode enriquecer a análise de textos literários de maneira quantitativa.

## 4.4 Paralelismos em diferentes livros de prosa brasileira

Com o intuito de para ilustrar como o método pode ser utilizado em diversos livros da literatura brasileira, além do livro central de análise (*Os Sertões*), para identificar e analisar relações interníveis entre obras, para essa seção, o método computacional proposto foi aplicado em outros livros da literatura brasileira, especificamente nos livros *Dom Casmurro*, de Machado de Assis e *Macunaíma*, de Mário de Andrade.

Para garantir a comparabilidade dos resultados entre as diferentes obras, os mesmos valores de parâmetros foram empregados: o tamanho da UT em 100, a janela de tamanho 10 e o limiar em 0.75.

A Tabela 4.4 apresenta uma comparação dos resultados obtidos nos diferentes livros ao utilizar a correlação segmentada para identificar trechos com paralelismos.

Tabela 4.4: Comparação dos resultados entre diferentes livros considerando o tamanho da UT=100 e o tamanho da janela=10.

| **Livro** | ***Dom Casmurro*** | ***Macunaíma*** | ***Os Sertões*** |
|---|---|---|---|
| **Sentenças** | 4354 | 3426 | 8634 |
| **UTs** | 44 | 35 | 87 |
| Pares em que foram encontrados paralelismos | 1354 | 1527 | 2588 |
| Trechos com paralelismos | 6287 | 7785 | 16616 |

A análise da Tabela 4.4 revela que os livros utilizados na pesquisa apresentam uma variação nas quantidades de paralelismos identificados. Vale destacar que essa variação não parece estar diretamente relacionada à extensão dos livros, uma vez que *Macunaíma* possui mais paralelismos do que *Dom Casmurro*, apesar deste último ter mais sentenças.

A aplicação do método computacional em outras obras literárias permite identificar diferentes paralelismos entre elas, entretanto, alguns pares de características apresentam paralelismos nas três obras utilizadas.

Um dos pares de características em que são encontrados paralelismos nas três obras é o par "frequência de Entidades Nomeadas" e "frequência de raiva", como demonstrado na Figura 4.24 para o livro *Os Sertões*. Nas próximas seções, será explorado esse paralelismo entre as características nos livros *Dom Casmurro*, de Machado de Assis, e *Macunaíma*, de Mário de Andrade.

### 4.4.1 Dom Casmurro

Conforme a Tabela 4.4, a aplicação do método computacional no livro *Dom Casmurro* resultou na identificação de 6287 paralelismos distribuídos em 1354 pares de características.

Dentre os pares onde foram identificados trechos com paralelismos já expostos na seção anterior para o livro *Os Sertões*, o único em comum com o livro *Dom Casmurro* é o paralelismo entre o par de características "frequência de Entidades Nomeadas" e "frequência de raiva" (Figura 4.24 para o livro *Os Sertões*).

Como os paralelismos para o livro *Dom Casmurro* são encontrados nos mesmos trechos tanto para a correlação de Pearson quanto para a correlação de Spearman, e a Figura 4.24 apresenta a utilização da correlação de Spearman para o livro *Os Sertões*, para fins de comparação, essa medida será utilizada para exemplificação do resultado.

Em *Dom Casmurro* o par de características "frequência de Entidades Nomeadas" e "frequência de raiva" apresenta 5 trechos com paralelismos, enquanto para o livro *Os Sertões* são encontrados 8 paralelismos para esse mesmo par de características.

A Figura 4.25 apresenta a visualização da correlação segmentada utilizando Spearman com os trechos com paralelismo identificados pelo método computacional proposto entre estas características para o livro *Dom Casmurro*.

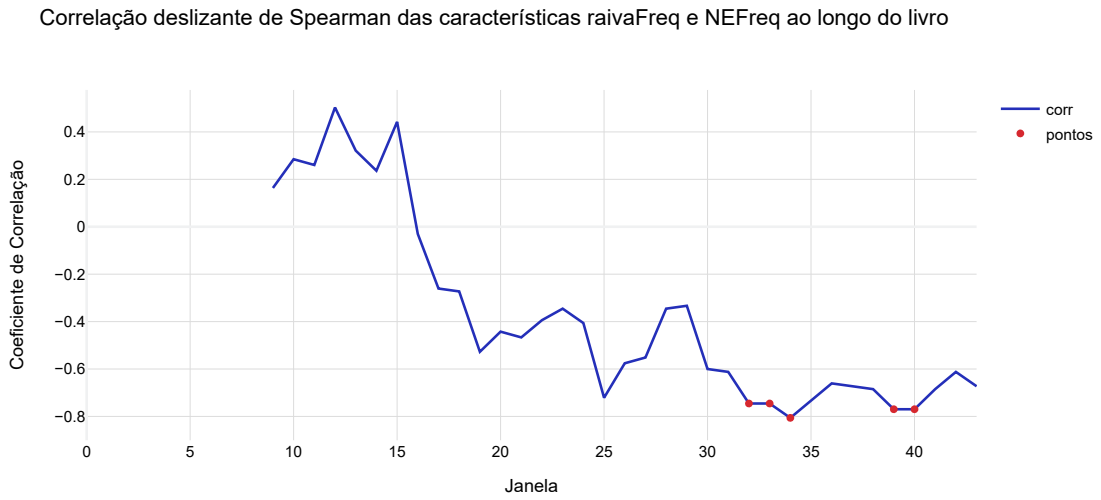


Figura 4.25: Correlação segmentada entre as características "frequência de Entidades Nomeadas" e "frequência de raiva" no livro *Dom Casmurro* utilizando a correlação de Spearman com UT de tamanho 100 e janela de tamanho 10

A partir da Figura 4.25 é possível observar a existência de correlações negativas acima do limiar de 0.75, o que, para o método computacional proposto, caracteriza a existência de paralelismo nestes trechos. Neste caso, os paralelismos foram encontrados nas janelas de índice 32, 33, 34, 39 e 40, sendo o índice 34 o que possui a correlação mais forte, podendo ser observada a partir do gráfico de dispersão da Figura 4.26.

Este resultado pode dar origem a investigações baseadas na extração de paralelismos comuns entre diferentes obras literárias, proporcionando novas perspectivas e oportunidades para estudos mais aprofundados sobre elas. Além disso, é possível investigar pares de características com trechos com paralelismos que não foram identificados no livro *Os Sertões*.

### 4.4.2 Macunaíma

De acordo com a Tabela 4.4, a aplicação do método computacional proposto no livro *Macunaíma*, de Mário de Andrade resultou na identificação de 7785 paralelismos distribuídos em 1527 pares de características.

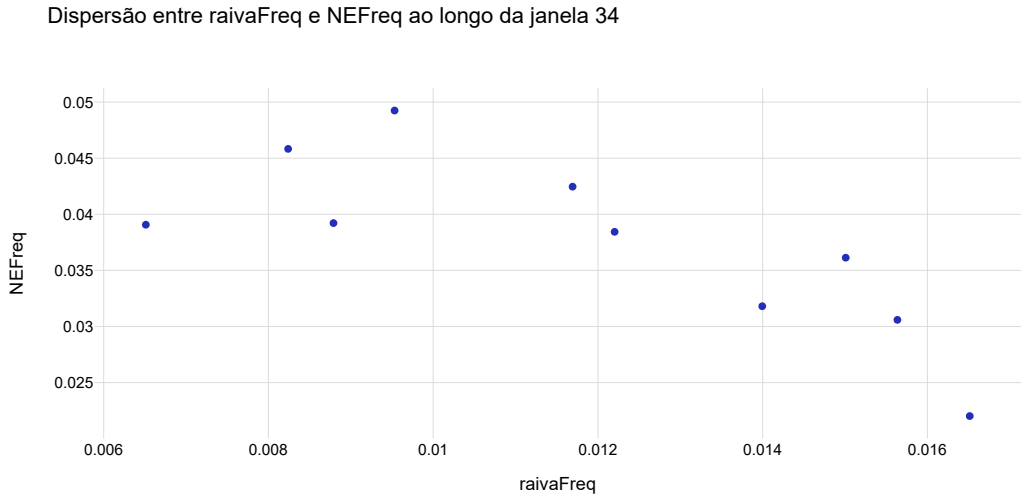


Figura 4.26: Gráfico de dispersão da janela 34 entre as características "frequência de Entidades Nomeadas" e "frequência de raiva" no livro *Dom Casmurro*

Assim como debatido na subseção anterior, o único paralelismo em comum entre os livros *Os Sertões* e *Dom Casmurro* é o paralelismo entre o par de características "frequência de Entidades Nomeadas" e "frequência de raiva" (Figura 4.24 para o livro *Os Sertões* e Figura 4.25 para o livro *Dom Casmurro*), que também é encontrado no livro *Macunaíma*.

Enquanto para o livro *Dom Casmurro* este par de características apresenta 5 paralelismos e para o livro *Os Sertões* apresenta 8 paralelismos, utilizando o livro *Macunaíma* são encontrados 3 paralelismos, nas janelas de índice 22, 24 e 25, conforme exposto na Figura 4.27.

Os paralelismos expostos na Figura 4.27, assim como todos os paralelismos encontrados em *Dom Casmurro* para o mesmo par de características (Figura 4.25) e a maioria dos paralelismos encontrados em *Os Sertões* (Figura 4.24), foram obtidos a partir de correlações negativas, o que pode indicar algum padrão entre as obras que pode ser investigado com a utilização do método computacional proposto.

Com isso, a aplicação do método computacional proposto em outras obras da literatura brasileira pode trazer novas visões sobre a forma de um autor ou de um período literário. Além disso, podem ser realizadas análises textuais de forma comparativa entre uma obra e outra, abrindo um leque de possibilidades no âmbito das pesquisas nas áreas de linguística computacional e análise textual.

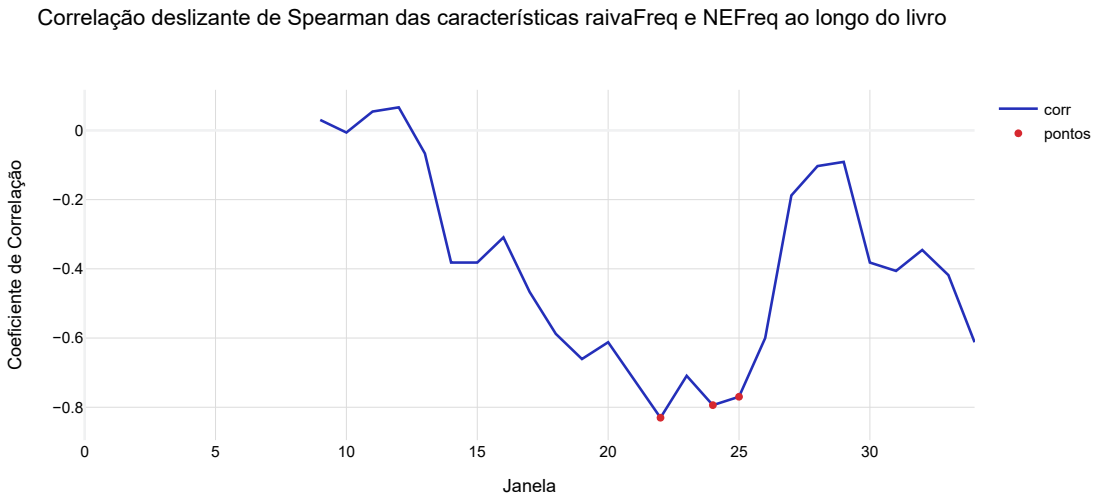


Figura 4.27: Correlação segmentada entre as características "frequência de Entidades Nomeadas" e "frequência de raiva" no livro *Macunaíma* utilizando a correlação de Spearman com UT de tamanho 100 e janela de tamanho 10

# Capítulo 5

# Conclusões

As pesquisas e experimentos realizados ao longo da construção desta dissertação culminaram no desenvolvimento de um método computacional para a identificação de padrões textuais em diferentes níveis linguísticos. Os resultados apresentados aqui não apenas demonstram a eficácia do método proposto na detecção de padrões textuais na forma de paralelismos, mas também estabelecem as bases para uma nova abordagem na análise textual.

Ao analisar o livro *Os Sertões* de Euclides da Cunha, pudemos identificar uma variedade de paralelismos, evidenciando a diversidade dos resultados em diferentes partes da obra. Além disso, estendemos nossa pesquisa para incluir outras obras literárias como *Dom Casmurro* e *Macunaíma*. Surpreendentemente, mesmo diante das diferenças estilísticas entre autores e obras, constatamos que algumas características apresentaram padrões de correlação semelhantes quando comparadas entre essas diferentes obras.

Apesar de termos apresentado apenas alguns exemplos de trechos com paralelismos identificados, é relevante destacar que nosso método computacional identificou mais de um milhão deles em análises mais abrangentes. Isso não apenas enriquece nosso entendimento das relações textuais, mas também destaca a complexidade dos paralelismos presentes em textos literários.

Os exemplos selecionados ao longo desta dissertação foram escolhidos para proporcionar uma visão abrangente das possibilidades oferecidas pelo nosso método computacional.

Um dos destaques deste trabalho é a introdução de uma abordagem quantitativa em um domínio de pesquisa que historicamente se baseia em análises qualitativas. Isso não só abre novas perspectivas de estudo, mas também oferece novas investigações a questões de pesquisa preexistentes. A análise quantitativa oferece uma visão objetiva e mensurável dos resultados, tornando-se uma ferramenta valiosa tanto para a validação quanto para a contestação de hipóteses estabelecidas anteriormente.

Essa abordagem significa que nosso método computacional pode oferecer assistência significativa a pesquisadores da área de Linguística em estudos literários mais aprofundados, permitindo a formulação de novas questões na área.

Mesmo não encontrando uma pesquisa ou estudo de uma ferramenta ou método computacional que proponha a aplicação da correlação para fins de identificação de paralelismos, a fim de realizar uma comparação com o método proposto nesta pesquisa, é possível realizar o ajuste de características e a extração e posterior inclusão de mais características para diversificar, avaliar e validar os resultados expostos.

Algumas das melhorias possíveis para esta pesquisa incluem a utilização de grupos de controle, a aplicação do método proposto em obras de poesia, a utilização de outros métodos de extração de sentimentos e posterior marcação dos *tokens* além da utilização de dicionários (como proposto aqui), a utilização de outras ferramentas mais precisas para a identificação de entidades nomeadas, de forma a melhorar a marcação e análise dessas estruturas no método proposto e a utilização e implementação de outros níveis para a análise.

Dessa forma, abrimos as portas para futuras investigações e melhorias, expandindo o leque de possibilidades disponibilizadas pelo nosso método computacional.

# Referências


Abaurre, M. L. M. e Pontara, M. (2006). Gramática: texto: análise e construção de sentido. *São Paulo: Moderna*.

Almuhareb, A., Alkharashi, I., Saud, L. A., e Altuwaijri, H. (2013). Recognition of classical arabic poems. In *Proceedings of the Workshop on Computational Linguistics for Literature*, páginas 9–16.

Amancio, D. R. (2015). A complex network approach to stylometry. *PloS one*, 10(8):e0136076.

Andrade, C. D. d. (2004). *Sentimento do mundo*. Record.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., e Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, páginas 166–170, New York City. Association for Computational Linguistics.

Balage Filho, P., Pardo, T. A. S., e Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Balint, M., Dascalu, M., e Trausan-Matu, S. (2016). Classifying written texts through rhythmic features. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, páginas 121–129. Springer.

Bandeira, M. (2019). *Lira dos cinquent'anos*. Global Editora e Distribuidora Ltda.

Bechara, E. (2012). *Moderna gramática portuguesa*. Nova Fronteira.

Bird, S., Klein, E., e Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., e Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bolt, T. J., Flynt, J. H., Chaudhuri, P., e Dexter, J. P. (2019). A stylometry toolkit for latin literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, páginas 205–210.

Burrows, J. (2004). Textual analysis.". *A companion to digital humanities*, páginas 323–347.

Busa, R. (1980). The annals of humanities computing: The index thomisticus. *Computers and the Humanities*, páginas 83–90.

Busa, R. (1992). *Thomae Aquinatis Opera omnia cum hypertextibus in CD-ROM*. Editoria Elettroniea Editel Milan.

Busa, R. A. (2008). Foreword: perspectives on the digital humanities. *A Companion to Digital Humanities, Hoboken, Wiley-Blackwell. Accès: http://www. digitalhumanities. org/companion/view*.

Cardoso, M., Loula, A., e Pires, M. G. (2016). Automated fuzzy system based on feature extraction and selection for opinion classification across different domains. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24:93–122.

Carvalho, R., Loula, A., e Queiroz, J. (2020). Identificação computacional de estruturas métricas de versificação na prosa de euclides da cunha. *Revista de Estudos da Linguagem*, 28(1).

Cavanagh, D., Gillis, A., Keown, M., Loxley, J., e Stevenson, R. (2014). *The Edinburgh Introduction to Studying English Literature*. Edinburgh University Press.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Costa, M. e Duarte, D. (2019). Avaliação de abordagens probabilísticas de extração de tópicos em documentos curtos. In *Anais da XV Escola Regional de Banco de Dados*, páginas 51–60. SBC.

de Camões, L. (1998). *Sonetos de Camões*, volume 4. Atelie Editorial.

Dell'Orletta, F., Montemagni, S., e Venturi, G. (2013). Linguistic profiling of texts across textual genres and readability levels. an exploratory study on italian fictional prose. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, páginas 189–197.

Drucker, J. (2021). *The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship*. Routledge.

Eder, M., Piasecki, M., e Walkowiak, T. (2017). An open stylometric system based on multilevel text analysis. *Cognitive Studies| Études cognitives*, (17).

Ek, A., Wirén, M., Östling, R., Björkenstam, K. N., Grigonytė, G., e Gustafson-Capková, S. (2018). Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ellegård, A. (1962). *A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772*. Acta Universitatis Gothoburgensis.

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 634–644.

Elson, D. K., McKeown, K., e Dames, N. J. (2010). Extracting social networks from literary fiction.

Eve, M. P. (2022). *The digital humanities and literary studies*. Oxford University Press.

Ferreira, J., Gonçalo Oliveira, H., e Rodrigues, R. (2019). Improving NLTK for processing Portuguese. In *Symposium on Languages, Applications and Technologies (SLATE 2019)*. In press.

Frog, E. (2014). A preface to parallelism. In *Parallelism in Verbal Art and Performance*, páginas 7–28. University of Helsinki, Folklore Studies, Department of Philosophy, History ....

Gianitsos, E., Bolt, T., Chaudhuri, P., e Dexter, J. P. (2019). Stylometric classification of ancient greek literary texts by genre. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, páginas 52–60.

Hawthorn, J. (2000). A glossary of contemporary literary theory. london: Arnold.

Hockey, S. (2004). The history of humanities computing. *A companion to digital humanities*, páginas 3–19.

Indurkhya, N. e Damerau, F. J. (2010). *Handbook of natural language processing*, volume 2. CRC Press.

J. Fox, J. (2014). *Explorations in semantic parallelism*. Anu Press.

Jacobs, A. M. (2019). Sentiment analysis for words and fiction characters from the perspective of computational (neuro-) poetics. *Frontiers in Robotics and AI*, 6:53.

Jakobson, R. (2008). *Lingüística e comunicação*. Editora Cultrix.

Jakobson, R. e Pomorska, K. (1985). Diálogos. tradução elisa a. kossovitch.

Jänicke, S., Franzini, G., Cheema, M. F., e Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In *EuroVis (STARs)*, páginas 83–103.

Jänicke, S., Franzini, G., Cheema, M. F., e Scheuermann, G. (2017). Visual text analysis in digital humanities. *Comput. Graph. Forum*, 36(6):226–250.

Jockers, M. L. e Underwood, T. (2015). *Text-Mining the Humanities*, chapter 20, páginas 291–306. John Wiley & Sons, Ltd.

Juola, P. (2013). How a computer program helped show jk rowling write a cuckoo's calling. *Scientific American, August 20th*.

Kakkonen, T. e Kakkonen, G. G. (2011). Sentiprofiler: creating comparable visual profiles of sentimental content in texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, páginas 62–69.

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., e Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16.

Kao, J. e Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, páginas 8–17.

Kaplan, D. M. e Blei, D. M. (2007). A computational approach to style in american poetry. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, páginas 553–558. IEEE.

Kehoe, A. e Gee, M. (2013). emargin: A collaborative textual annotation tool. *Ariadne*, (71).

Lagutina, K., Lagutina, N., Boychuk, E., Larionov, V., e Paramonov, I. (2021). Authorship verification of literary texts with rhythm features. In *2021 28th Conference of Open Innovations Association (FRUCT)*, páginas 240–251. IEEE.

Lagutina, K., Lagutina, N., Boychuk, E., e Paramonov, I. (2020). The influence of different stylometric features on the classification of prose by centuries. In *2020 27th Conference of Open Innovations Association (FRUCT)*, páginas 108–115. IEEE.

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., e Demidov, P. (2019). A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, páginas 184–195. IEEE.

Lahiri, S. e Mihalcea, R. (2013). Authorship attribution using word network features. *arXiv preprint arXiv:1311.2978*.

Larson, R. e Farber, B. (2015). Estatística aplicada [applied statistics].

Lee, J. S. e Yeung, C. Y. (2012). Extracting networks of people and places from literary texts. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, páginas 209–218.

Lima, L. O. d. S. et al. (2021). Computational identification of phonological parallelisms in brazilian literary prose.

Lisboa, H. (2018). *O menino poeta: obra completa*. Editora Peirópolis LTDA.

Liu, B. e Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, páginas 415–463. Springer.

Manning, C., Raghavan, P., e Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Meireles, C. (1960). *Metal rosicler*, volume 9. Livros de Portugal.

Melka, T. S. e Místeckỳ, M. (2020). On stylometric features of h. beam piper's omnilingual. *Journal of Quantitative Linguistics*, 27(3):204–243.

Min, S. e Park, J. (2019). Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, 14(12):e0226025.

Mohammad, S. M. e Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Mohsin, M. A. e Beltiukov, A. (2019/05). Summarizing emotions from text using plutchik's wheel of emotions. In *Proceedings of the 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)*, páginas 291–294. Atlantis Press.

Montemurro, M. A. e Zanette, D. H. (2013). Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. *PloS one*, 8(6):e66344.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. Verso.

Mosteller, F. e Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Nadeau, D. e Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Oberländer, L. A. M. e Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, páginas 2104–2119.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Pennebaker, J. W., Francis, M. E., e Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Plutchik, R. (1982). A psychoevolutionary theory of emotions.

Presner, T. (2010). Digital humanities 2.0: a report on knowledge.

Ramnial, H., Panchoo, S., e Pudaruth, S. (2016). Authorship attribution using stylometry and machine learning techniques. In *Intelligent Systems Technologies and Applications*, páginas 113–125. Springer.

Ramsay, S. (2013). *Algorithmic Criticism*, chapter 26, páginas 477–491. John Wiley & Sons, Ltd.

Rauscher, J., Swiezinski, L., Riedl, M., e Biemann, C. (2013). Exploring cities in crime: significant concordance and co-occurrence in quantitative literary analysis. In *Proceedings of the Workshop on Computational Linguistics for Literature*, páginas 61–71.

Řehůřek, R. e Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, páginas 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Richardson, J. E. (2004). Sourcebook on rhetoric: Key concepts in contemporary rhetorical studies.

Romero, S. (1954). *Folclore brasileiro: Cantos populares do Brasil, tomo 1-2*, volume 75. José Olympio.

Rommel, T. (2004). Literary studies. *A companion to digital humanities*, páginas 88–96.

Rustici, C. M. (1997). Sonnet writing and experiential learning. *College Teaching*, 45(1):16–18.

Saito, S., Ohno, S., e Inaba, M. (2010). A platform for cultural information visualization using schematic expressions of cube. páginas 365–367.

Schreibman, S., Siemens, R., e Unsworth, J. (2004). The digital humanities and humanities computing: An introduction. *A companion to digital humanities*, páginas 288–290.

Shao, M. e Qin, L. (2014). Text similarity computing based on lda topic model and word co-occurrence. In *2014 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014)*, páginas 199–203. Atlantis Press.

Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*, páginas 1–27.

Silva, M. J., Carvalho, P., e Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language*, páginas 218–228. Springer.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., e Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Sousa, C. (1900). Violões que choram. *Faróis. Rio de Janeiro: Tip. do Instituto Nacional*, páginas 58–64.

Souza, M. e Vieira, R. (2011). Construction of a portuguese opinion lexicon from multiple resources. *Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2011, Brasil.*

Ströbel, M., Kerz, E., Wiechmann, D., e Qiao, Y. (2018). Text genre classification based on linguistic complexity contours using a recurrent neural network. In *MRC@ IJCAI.*

Sun, Z., Sarma, P. K., Sethares, W. A., e Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis.

Tang, K.-M., Huang, C.-K., e Lee, C.-m. (2012). Selection of discriminative features for translation texts. In *Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology*, páginas 49–64.

van Cranenburgh, A. (2018). Cliche expressions in literary and genre novels. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, páginas 34–43.

Vieira, R. e Lima, V. L. S. (2001). Lingüística computacional: princípios e aplicações.

Wanner, L. et al. (2017). On the relevance of syntactic and discourse features for author profiling and identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, páginas 681–687.

Waumans, M. C., Nicodème, T., e Bersini, H. (2015). Topology analysis of social networks extracted from literature. *PloS one*, 10(6):e0126470.

Xiao, W. e Sun, S. (2020). Dynamic lexical features of phd theses across disciplines: A text mining approach. *Journal of Quantitative Linguistics*, 27(2):114–133.

Xu, Z., Liu, L., Song, W., e Du, C. (2017). Text genre classification research. In *2017 International Conference on Computer, Information and Telecommunication Systems (CITS)*, páginas 175–178. IEEE.

Yeruva, V. K., ChandraShekar, M., Lee, Y., Rydberg-Cox, J., Blanton, V., e Oyler, N. A. (2020). Interpretation of sentiment analysis in aeschylus's greek tragedy. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, páginas 138–146.

Zhao, C., Song, W., Liu, L., Du, C., e Zhao, X. (2017). Research on author identification based on deep syntactic features. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, páginas 276–279. IEEE.

# Apêndice A

# Levantamento de características

Tabela A.1: Características mais comuns encontradas durante o levantamento bibliográfico.

| **Nº** | **Característica** | **Grupo Chave** | **Nível Textual** | **Artigos** |
|---|---|---|---|---|
| 1 | O número médio de palavras por frase | Comprimento da Sentença | Sintático | Dell'Orletta et al. (2013); Ströbel et al. (2018); Gianitsos et al. (2019); van Cranenburgh (2018); Bolt et al. (2019); Tang et al. (2012); Wanner et al. (2017); Zhao et al. (2017); Balint et al. (2016); Lagutina et al. (2020); Yeruva et al. (2020) |
| 2 | O desvio padrão de palavras por frase | | | |
| 3 | A diferença entre o número máximo e mínimo de palavras por frase em um texto | | | |
| 4 | Frequência de letras | Frequência de caracteres | Sintático | Oberländer e Klinger (2018); Tang et al. (2012); Wanner et al. (2017); Zhao et al. (2017); Almuhareb et al. (2013); Lagutina et al. (2021); Balint et al. (2016); Xu et al. (2017); Lagutina et al. (2020); Amancio (2015) |
| 5 | Frequência de sinais de pontuação | | | |
| 6 | O número médio de caracteres por palavra | Comprimento da palavra | Sintático | Dell'Orletta et al. (2013); Ströbel et al. (2018); Tang et al. (2012); Wanner et al. (2017); Zhao et al. (2017); Lagutina et al. (2021); Balint et al. (2016); Kaplan e Blei (2007); Lagutina et al. (2020); Melka e Místeckỳ (2020) |
| 7 | O desvio padrão do comprimento da palavra | | | |

| | | | | |
|---|---|---|---|---|
| 8 | A diferença entre o comprimento máximo da palavra e o comprimento mínimo da palavra | | | |
| 9 | Frequência das palavras mais frequentes no top k | Frequência da palavra | Léxico | Dell'Orletta et al. (2013); Kao e Jurafsky (2012); van Cranenburgh (2018); Wanner et al. (2017); Zhao et al. (2017); Almuhareb et al. (2013); Lagutina et al. (2021); Balint et al. (2016); Xu et al. (2017); Kaplan e Blei (2007); Amancio (2015); Melka e Místeckỳ (2020); Lahiri e Mihalcea (2013); Montemurro e Zanette (2013) |
| 10 | Número de palavras distintas (riqueza lexical) | | | |
| 11 | Frequência das palavras irrelevantes | | | |
| 12 | Frequência dos n-gramas | | | |
| 13 | TTR= números de palavras diferentes/total de tokens | Relação Tipo/ Token (Type/ Token Ratio - TTR) | Léxico | Dell'Orletta et al. (2013); Ströbel et al. (2018); Kao e Jurafsky (2012); van Cranenburgh (2018); Melka e Místeckỳ (2020); Xiao e Sun (2020) |
| 14 | Frequência relativa de substantivos, verbos, pronomes, adjetivos, advérbios, conjunções | Distribuição da parte do discurso | Morfossintático | Dell'Orletta et al. (2013); Tang et al. (2012); Wanner et al. (2017); Zhao et al. (2017); Rauscher et al. (2013); Lee e Yeung (2012); Xu et al. (2017); Kaplan e Blei (2007); Lagutina et al. (2020); Melka e Místeckỳ (2020); Kakkonen e Kakkonen (2011,?); Cardoso et al. (2016) |
| 15 | Relação de palavras de conteúdo (verbos, substantivos , adjetivos e advérbios) ao número total de tokens lexicais em um texto | Densidade lexical | Morfossintático | Dell'Orletta et al. (2013); Ströbel et al. (2018); Gianitsos et al. (2019); Ek et al. (2018); Bolt et al. (2019); Zhao et al. (2017); Rauscher et al. (2013); Elson et al. (2010); Lee e Yeung (2012); Kaplan e Blei (2007); Lagutina et al. (2020); Melka e Místeckỳ (2020); Elsner (2012); Cardoso et al. (2016) |

| | | | | |
|---|---|---|---|---|
| 16 | Distribuição dos verbos de acordo com seu modo, tempo e pessoa | Modo, tempo e pessoa dos verbos | Morfossintático | Dell'Orletta et al. (2013); Ek et al. (2018) |
| 17 | Distribuição de diferentes tipos de dependências sintáticas (por exemplo, sujeito, objeto direto, modificador, etc.) | Distribuição de tipos de dependências | Sintático | Dell'Orletta et al. (2013) |
| 18 | O número máximo de irmãos em qualquer um dos níveis da árvore | Análise de profundidade de árvores de palavras | Sintático | Dell'Orletta et al. (2013); Wanner et al. (2017); Zhao et al. (2017); Elsner (2012) |
| 19 | O fator de ramificação: o número médio de filhos por nível. | | | |
| 20 | Número de raízes verbais | Predicados verbais | Sintático | Dell'Orletta et al. (2013); Wanner et al. (2017) |
| 21 | Aridade calculada como o número de ligações de dependência instanciadas compartilhando o mesmo núcleo verbal | | | |
| 22 | Distribuição de predicados verbais por aridade | | | |
| 23 | Porcentagem de predicados verbais com sujeito elíptico | | | |
| 24 | Contagem das conjunções comuns | Conjunções | Sintático | Gianitsos et al. (2019); Bolt et al. (2019); Wanner et al. (2017); Argamon et al. (2007); Lee e Yeung (2012); Kaplan e Blei (2007) |
| 25 | Distribuição de orações subordinadas e orações principais | Subordinação | Sintático | Dell'Orletta et al. (2013); Bolt et al. (2019) |
| 26 | Ordenação relativa das subordinadas em relação à oração principal | | | |

| | | | | |
|---|---|---|---|---|
| 27 | Profundidade média de 'cadeias' de orações subordinadas incorporadas | | | |
| 28 | A distribuição de 'cadeias' de orações subordinadas incorporadas por profundidade | | | |
| 29 | Distribuição de orações subordinadas pós-verbais | | | |
| 30 | Medida o comprimento da dependência em termos das palavras que ocorrem entre o núcleo e o dependente | Comprimento das ligações de dependência | Sintático | Dell'Orletta et al. (2013) |
| 31 | Contagem de rimas finais perfeitas em um poema normalizado pelo número total de palavras | Rima final perfeita e inclinada | Fonético | Kao e Jurafsky (2012); Almuhareb et al. (2013); Balint et al. (2016); Kaplan e Blei (2007); Lagutina et al. (2020) |
| 32 | Contagem de rimas finais inclinadas em um poema normalizado pelo número total de palavras | | | |
| 33 | Contagem de aliteração | Aliteração, consonância e assonância | Fonético | Kao e Jurafsky (2012); Balint et al. (2016); Kaplan e Blei (2007) |
| 34 | Contagem de assonâncias | | | |
| 35 | Contagem de consonâncias | | | |
| 36 | Contagem total de aliterações dividida pelo número total de palavras para obter uma pontuação de aliteração | | | |

| | | | | |
|---|---|---|---|---|
| 37 | Contagem total de assonâncias dividida pelo número total de palavras para obter uma pontuação de assonância | | | |
| 38 | Contagem total de consonâncias dividida pelo número total de palavras para obter uma pontuação de consonância | | | |
| 39-50 | Soma de adjetivos ou advérbios ou adjetivos e bigramas ou advérbios e bigramas ou unigramas e bigramas ou unigramas, bigramas e trigramas (positivos ou negativos) | Léxicos de sentimento | Sintático Semântico Léxico | Kao e Jurafsky (2012); Wanner et al. (2017); Yeruva et al. (2020); Kakkonen e Kakkonen (2011); Cardoso et al. (2016) |
| 51-62 | Soma normalizada adjetivos ou advérbios ou adjetivos e bigramas ou advérbios e bigramas ou unigramas e bigramas ou unigramas, bigramas e trigramas (positivos ou negativos) | | | |
| 63-66 | Contagem de n-gramas de adjetivos ou advérbios (positivos ou negativos) | | | |
| 67-70 | Contagem dos n-gramas de adjetivos ou advérbios (positivos ou negativos) e bigramas compostos por advérbios e adjetivo ou apenas por advérbios | | | |

| | | | | |
|---|---|---|---|---|
| 71-72 | Contagem dos n-gramas de unigramas e bigramas (positivos ou negativos) | | | |
| 73-74 | Contagem dos n-gramas de unigramas, bigramas e trigramas (positivos ou negativos) | | | |
| 75-80 | Subtração entre as somas (positivas e negativas) de adjetivos ou advérbios ou adjetivos e bigramas compostos de advérbio e adjetivo ou advérbios e bigramas compostos apenas por advérbios ou unigramas e bigramas ou unigramas, bigramas e trigramas | | | |
| 81-86 | Subtração entre as contagens (positivas e negativas) de adjetivos ou de advérbios ou de adjetivos e bigramas compostos por advérbio e adjetivo ou advérbios e bigramas compostos apenas por advérbios ou de unigramas e bigramas ou de unigramas, bigramas e trigramas | | | |

| | | | | |
|---|---|---|---|---|
| 87-92 | Classe de polaridade máxima entre adjetivos ou advérbios ou adjetivos e bigramas compostos por advérbio e adjetivo ou adjetivos e bigramas compostos apenas por advérbios ou unigramas e bigramas ou unigramas, bigramas e trigramas | | | |
| 93 | Contagem de menções do personagem por linha | Menção de personagens por Linha/ Frequência de personagens na Narrativa | Sintático Semântico | Ek et al. (2018); Elsner (2012) |
| 94 | Contagem de verbos de fala por linha | | | |
| 95 | Contagem de menção e verbo de fala por linha | | | |
| 96 | Contagem de menções em um diálogo | | | |
| 97 | Frequência bruta dos personagens na narrativa | | | |
| 98 | Frequência no contexto global, ordem de menção, pronome. | | | |
| 99 | Contagem de estrutura métrica por tipo | Características métricas | Rítmico Léxico Fonético | Almuhareb et al. (2013); Balint et al. (2016) |
| 100 | Número de todas as figuras léxico-gramaticais dividido pelo número de sentenças e multiplicado por 100— o número médio de figuras por 100 sentenças | Figuras de Linguagem (Anáfora, anadiplose, diácope, epanalepse, epífora, epizeuxe, polissíndeto, símploce, etc.) | Rítmico Léxico Fonético Sintático | Lagutina et al. (2021); Balint et al. (2016); Lagutina et al. (2020) |

| | | | | |
|---|---|---|---|---|
| 101 | Porcentagens de cada figura de linguagem entre todas as figuras de linguagem | | | |
| 102 | Número de ocorrências de uma figura (anáfora, epífora, etc.) em um texto dividido pelo número de sentenças | | | |
| 103 | Fração das palavras únicas—palavras que se repetem apenas uma vez em figuras rítmicas | | | |
| 104 | Frações de substantivos, verbos, advérbios e adjetivos — entre todas as palavras que aparecem entre as figuras rítmicas. | | | |
| 105 | Contagem de Entidades Nomeadas (Pessoa) | Entidades Nomeadas | Semântico Sintático | Elson et al. (2010); Lee e Yeung (2012) |
| 106 | Contagem de Entidades Nomeadas (Local) | | | |
| 107 | Contagem de Entidades Nomeadas (Pessoa+Local) | | | |
| 108 | Contagem de Entidades Nomeadas | | | |
| 109 | Contagem de Entidades Nomeadas únicas (Pessoa) | | | |
| 110 | Contagem de Entidades Nomeadas únicas (Localização) | | | |
| 111 | Contagem de Entidades Nomeadas (Pessoa) por sentença em um texto | | | |

| | | | | |
|---|---|---|---|---|
| 112 | Contagem de Entidades Nomeadas (Localização) por sentença em um texto | | | |
| 113 | Contagem de Entidades Nomeadas (Pessoa+Localização) por sentença em um texto | | | |
| 114 | Contagem de Entidades Nomeadas por sentença em um texto | | | |
| 115 | Média de Entidades Nomeadas (Pessoa) por sentença em um texto | | | |
| 116 | Média de Entidades Nomeadas (Localização) por sentença em um texto | | | |
| 117 | Média de Entidades Nomeadas (Pessoa+Local) por sentença em um texto | | | |
| 118 | Média de Entidades Nomeadas por sentença em um texto | | | |

# Apêndice B

# Lista de *stopwords*

'a', 'à', 'afinal', 'ah', 'aí', 'ainda', 'ainda', 'alguns', 'ali', 'antes', 'ao', 'aos', 'apenas', 'apenas', 'aquela', 'aquelas', 'aquele', 'aqueles', 'aqui', 'aquilo', 'as', 'às', 'assim', 'assim', 'até', 'cada', 'cair', 'capítulo', 'certa', 'certo', 'coisa', 'com', 'como', 'da', 'dali', 'dar', 'das', 'de', 'dela', 'delas', 'dele', 'deles', 'dentro', 'depois', 'des', 'desde', 'deve', 'disse', 'do', 'dos', 'e', 'é', 'ela', 'elas', 'ele', 'eles', 'em', 'enquanto', 'então', 'então', 'entre', 'era', 'eram', 'éramos', 'essa', 'essas', 'esse', 'esses', 'esta', 'está', 'estamos', 'estão', 'estar', 'estas', 'estava', 'estavam', 'estávamos', 'este', 'esteja', 'estejam', 'estejamos', 'estes', 'esteve', 'estive', 'estivemos', 'estiver', 'estivera', 'estiveram', 'estivéramos', 'estiverem', 'estivermos', 'estivesse', 'estivessem', 'estivéssemos', 'estou', 'et', 'eu', 'faz', 'fazer', 'fez', 'ficou', 'fizera', 'foi', 'fomos', 'for', 'fora', 'foram', 'fôramos', 'forem', 'formos', 'fosse', 'fossem', 'fôssemos', 'fui', 'há', 'haja', 'hajam', 'hajamos', 'hão', 'havemos', 'haver', 'havia', 'havia', 'haviam', 'hei', 'houve', 'houvemos', 'houver', 'houvera', 'houverá', 'houveram', 'houvéramos', 'houverão', 'houverei', 'houverem', 'houveremos', 'houveria', 'houveriam', 'houveríamos', 'houvermos', 'houvesse', 'houvessem', 'houvéssemos', 'ia', 'iam', 'ir', 'isso', 'isto', 'já', 'la', 'lá', 'las', 'les', 'lhe', 'lhes', 'lo', 'logo', 'los', 'maior', 'mais', 'mas', 'me', 'melhor', 'menos', 'mesma', 'mesmo', 'meu', 'meus', 'mim', 'minha', 'minhas', 'muita', 'muito', 'muito', 'na', 'nã$^{3}$s', 'nada', 'nada', 'não', 'nas', 'nem', 'nesse', 'neste', 'nisso', 'nisto', 'no', 'nos', 'nós', 'nossa', 'nossas', 'nosso', 'nossos', 'novos', 'num', 'numa', 'o', 'onde', 'onde', 'os', 'ou', 'ou', 'outra', 'outra', 'outras', 'outras', 'outro', 'outro', 'outros', 'outros', 'pag', 'pág', 'para', 'pela', 'pelas', 'pelo', 'pelos', 'pode', 'podem', 'pois', 'por', 'por', 'porque', 'porque', 'porquê', 'porquê', 'pra', 'qual', 'qualquer', 'quando', 'quanto', 'quase', 'quase', 'que', 'que', 'que', 'quê','quem', 'quer', 'são', 'se', 'sei', 'seja', 'sejam', 'sejamos', 'sem', 'ser', 'será', 'serão', 'serei', 'seremos', 'seria', 'seriam', 'seríamos', 'seu', 'seus', 'si', 'sido', 'só', 'sobre', 'somos', 'sou', 'sr', 'sra', 'sua', 'suas', 'tal', 'também', 'tanta', 'tantas', 'tanto', 'tantos', 'tão', 'tão', 'te', 'tem', 'tém', 'temos', 'tenha', 'tenham', 'tenhamos', 'tenho', 'ter', 'ter', 'terá', 'terão', 'terei', 'teremos', 'teria', 'teriam', 'teríamos', 'teu', 'teus', 'teve', 'tinha', 'tinham', 'tínhamos', 'tive', 'tivemos', 'tiver', 'tivera', 'tiveram', 'tivéramos', 'tiverem', 'tivermos', 'tivesse', 'tivessem', 'tivéssemos', 'toda', 'toda', 'todas', 'todas', 'todo', 'todo',

'todos', 'todos', 'tu', 'tua', 'tuas', 'tudo', 'ui', 'um', 'uma', 'uns', 'vai', 'vai', 'vão', 'vêm' 'vem', 'vinha', 'viu', 'você', 'vocês', 'vos'

# Apêndice C

# Lista de características e seu respectivo nível

Tabela C.1: Lista de características extraídas, seu nome no método e seu respectivo nível

| **Característica** | **Nome** | **Nível** |
|---|---|---|
| Rep.Sílabas Fonéticas 2S3L | repSil_2_3 | Fonético |
| Rep.Sílabas Fonéticas 3S4L | repSil_3_4 | |
| Rep.Sílabas Fonéticas 4S4L | repSil_4_4 | |
| Versificação - Sentenças Completas | versos_sc | Métrico |
| Versificação - Início de Sentença | versos_is | |
| Versificação - Final de Sentença | versos_fs | |
| Contagem de Palavras | wordCount | Lexical |
| Palavras únicas | palavrasUnicas | |
| Lemas únicos | lemasUnicos | |
| TTR de palavras | ttrPalavras | |
| TTR de lemas | ttrLemas | |
| Adjetivo | adj | |
| Advérbios | adv | |
| Artigos | art | |
| Conjunções | conjs | |
| Interjeições | in | |
| Numerais | num | |
| Pontuações | punc | |
| Pronomes | prons | |
| Preposições | prp | |
| Substantivo | noun | |
| Verbos | verbs | |
| Frequência relativa de Adjetivos | adjFreq | |

| | | |
|---|---|---|
| Frequência relativa de Advérbios | advFreq | Gramatical |
| Frequência relativa de Artigos | artFreq | |
| Frequência relativa de Conjunções | conjFreq | |
| Frequência relativa de Interjeições | inFreq | |
| Frequência relativa de Numerais | numFreq | |
| Frequência relativa de Pontuações | puncFreq | |
| Frequência relativa de Pronomes | pronFreq | |
| Frequência relativa de Preposições | prpFreq | |
| Frequência relativa de Substantivos | nounFreq | |
| Frequência relativa de Verbos | verbFreq | |
| Positivos | positivos | Sentimental |
| Negativos | negativos | |
| Neutros | neutros | |
| Alegria | alegria | |
| Confiança | confianca | |
| Expectativa | expectativa | |
| Medo | medo | |
| Nojo | nojo | |
| Raiva | raiva | |
| Surpresa | surpresa | |
| Tristeza | tristeza | |
| Polaridade | polaridade | |
| Carga Emocional | emoCharge | |
| Frequência relativa de Positivos | posFreq | |
| Frequência relativa de Negativos | negFreq | |
| Frequência relativa de Neutros | neuFreq | |
| Frequência relativa de Alegria | alegriaFreq | |
| Frequência relativa de Confiança | confiancaFreq | |
| Frequência relativa de Expectativa | expectativaFreq | |
| Frequência relativa de Medo | medoFreq | |
| Frequência relativa de Nojo | nojoFreq | |
| Frequência relativa de Raiva | raivaFreq | |
| Frequência relativa de Surpresa | surpresaFreq | |
| Frequência relativa de Tristeza | tristezaFreq | |
| Adjetivos Positivos | adjpos | Gramatical e Sentimental |
| Adjetivos Negativos | adjneg | |
| Advérbios Positivos | advpos | |
| Advérbios Negativos | advneg | |
| Polaridade de Adjetivos | adjPol | |
| Polaridade de Advérbios | advPol | |
| Carga Emocional de adjetivos | adjCharge | |
| Carga Emocional de advérbios | advCharge | |
| REN - Pessoa | nePessoa | |

| | | |
|---|---|---|
| REN - Local | neLocal | Entidades Nomeadas |
| REN - Pessoa e Local | nePessoaELocal | |
| REN - Geral | neGeral | |
| Frequência relativa de Pessoa | pessoaFreq | |
| Frequência relativa de Local | localFreq | |
| Frequência relativa de Pessoa/Local | pes_locFreq | |
| Frequência relativa de Entidades Nomeadas | NEFreq | |
| Tópico 1 | Tópico 1 | Tópicos* |
| Tópico 2 | Tópico 2 | |
| Tópico 3 | Tópico 3 | |
| Tópico 4 | Tópico 4 | |
| Tópico 5 | Tópico 5 | |

* Características de Modelagem de Tópicos disponíveis apenas quando o tamanho da UT for maior que 1.