TEXTO DIGITAL
Revista de Literatura, Linguística, Educação e Artes

Identificação computacional de padrões interníveis em textos da Literatura Brasileira

Computational identification of interlevel patterns in Brazilian Literature texts

Angelo Loula[(a)]; Luciano Alves Machado Júnior[(b)]

a Afiliação – angelocl@uefs.br

b Afiliação – lucianoamjunior@gmail.com

Resumo: Em textos literários, como a poesia e a 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 deste trabalho 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 interníveis entre elas. Os resultados obtidos com o intuito de demonstrar esse método computacional foram extraídos da obra literária brasileira *Os Sertões* de Euclides da Cunha. 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 alguns dos 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.

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 paper 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 interlevel patterns between them. The results obtained to demonstrate this computational method were extracted from the Brazilian literary works *Os Sertões* by Euclides da Cunha. 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.

Introdução

A linguagem, sistema complexo de comunicação humana, configura-se como um artefato cultural, social, histórico e ideológico fundamental para a apreensão das nuances de uma dada cultura. Sua intrínseca riqueza manifesta-se através de distintos níveis de organização e análise, que abrangem desde a estrutura fonológica e morfológica até as dimensões sintática, semântica e pragmática (Abaurre; Abaurre; Pontara, 2006), e a consideração desses níveis é basilar para uma investigação aprofundada da constituição textual.

Adicionalmente, a linguagem opera mediante um conjunto de funções que modulam a interação comunicativa. A taxonomia proposta por Roman Jakobson (Jakobson; Pomorska, 1985) destaca seis funções diferentes: referencial, emotiva, poética, fática, conativa e metalinguística. A função poética, em particular, com sua ênfase na mensagem em si e no emprego estratégico de múltiplos recursos linguísticos, revela-se particularmente pertinente ao estudo de textos literários. Esses recursos, denominados dispositivos poéticos (Goodrich, 1997), incluem padrões gramaticais, rítmicos, métricos, verbais e visuais, evidenciados em figuras de linguagem como repetições, rimas, aliterações e anáforas.

Para Jakobson (Jakobson; Pomorska, 1985), a estrutura da poesia é um paralelismo contínuo, um fenômeno que se caracteriza pela presença de padrões literários de repetição, sendo definido por Jakobson como um artifício poético de retornos recorrentes, e que não estão presentes apenas na linguagem poética, sendo encontrados também na prosa literária. Segundo Jakobson, 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; Pomorska, 1985).

A análise da ocorrência e da interação entre os diversos níveis linguísticos e funções da linguagem pode, portanto, fornecer conhecimentos valiosos sobre a organização e a expressividade de obras em prosa da literatura brasileira. Longe de reduzir o texto a uma quantificação de recursos, essa abordagem quantitativa complementa a leitura tradicional, pois, ao quantificar fenômenos literários, ela permite revelar padrões que a análise qualitativa convencional poderia não identificar.

Neste contexto, o presente artigo propõe e demonstra um método computacional para a identificação de padrões textuais interníveis em textos de prosa da literatura brasileira. A abordagem aqui apresentada permite localizar quantitativamente, ao longo de uma obra, trechos em que esses padrões ocorrem, oferecendo uma ferramenta para a observação de aspectos linguísticos que, tradicionalmente, dependiam apenas de análise qualitativa. Para isso, o método utiliza a extração de características textuais em diferentes níveis linguísticos, seguida pela análise de correlação para identificar a coocorrência de fenômenos distintos.

A computação literária, um subcampo das Humanidades Digitais, pode ser utilizada para a realização de análises textuais automatizadas,

identificando padrões e semelhanças em diferentes vertentes, localizando ao longo das obras e quantificando suas ocorrências (Rommel, 2004). Conforme apontam Jockers (2013) e Underwood (2019), essa abordagem amplia as possibilidades de investigação, permitindo, por exemplo, a análise de grandes volumes de textos e observação de tendências estilísticas, temáticas e estruturais. Moretti (2005), ao propor a leitura distante (*distant reading*), reforça a ideia de que a visualização e quantificação de dados textuais podem revelar fenômenos literários que escapam à análise qualitativa convencional.

Nesse sentido, a abordagem quantitativa e computacional atua como um ponto de partida, evidenciando padrões de frequência, distribuição e coocorrência de recursos linguísticos que seriam difíceis de discernir na leitura manual. Ela não substitui as análises qualitativas – como a análise estrutural, teórica, crítica e contextual, que se concentram em interpretar as nuances e os significados da obra –, mas as complementa, direcionando o pesquisador a trechos específicos para investigações mais aprofundadas. O método computacional, portanto, oferece uma nova lente para a 'leitura próxima', fornecendo evidências para a formulação de novas perguntas e para o aprofundamento das análises já existentes.

A extração de padrões estilométricos e características linguísticas em textos foi realizada em trabalhos anteriores em diversas línguas, com destaque para diferentes níveis de análise. Eder, Piasecki e Walkowiak (2017) desenvolveram o sistema WebSty para atribuição de autoria com base em características morfológicas, gramaticais e semânticas, demonstrando a utilização de técnicas de processamento de linguagem natural e aprendizado de máquina para a identificação de autoria e a capacidade de realizar análises em múltiplos níveis linguísticos. Já Dell'orletta, Montemagni e Venturi (2013) analisaram traços distintivos da prosa italiana, utilizando características

linguísticas variadas para classificação por gênero e legibilidade, com resultados promissores em tarefas de categorização textual.

De Roc Boronat e Wanner (2017) destacaram a importância de incorporar características de dependência sintática e discurso na identificação de autoria, enquanto Lagutina *et al*. (2020) compararam características estilométricas de níveis básicos e avançados, evidenciando a relevância da análise integrada de diferentes tipos de características para melhorar o desempenho em classificação textual. Ambos os estudos reforçam a importância de técnicas inter-níveis para a análise estilométrica.

No campo da Análise de Sentimentos, Jacobs (2019) introduziu a ferramenta SentiArt, que calcula perfis emocionais e de personalidade de personagens literários, enquanto Min e Park (2019) analisaram as variações de sentimentos em narrativas literárias usando o software LIWC, conectando interações entre personagens às flutuações emocionais ao longo das tramas. Esses trabalhos destacam como ferramentas computacionais podem revelar aspectos emocionais e estilísticos em textos.

Waumans, Nicod`Eme e Bersini (2015) exploraram redes sociais de personagens em romances, extraindo padrões estilométricos a partir da análise topológica das interações. Esse estudo, apesar de focado na estrutura narrativa, contribui para a análise textual por meio da integração de técnicas quantitativas e qualitativas.

Para obras literárias em língua portuguesa, há estudos que exploram a identificação de padrões estilométricos e múltiplos níveis linguísticos na língua portuguesa. A exemplo disso temos o trabalho de Galina, Flores e Komati (2019) cujo objetivo é identificar, por meio dos atributos léxicos de seus textos, os livros de Guimarães Rosa e Clarice Lispector, demonstrando em seus resultados que o uso do conjunto de palavras com

similaridade de Jaccard torna possível a separação dos livros por sua autoria. Outro exemplo é o estudo de Corso, Fossa e Oliveira (2005) que utiliza a teoria de redes para uma análise quantitativa do estilo de textos, comparando as obras de Machado de Assis e Rui Barbosa, além de fragmentos de um jornal local, para caracterizar a riqueza estilística.

Existem ainda algumas iniciativas voltadas a construção de corpora de obras literárias em língua portuguesa. Um exemplo é a Literateca como parte do projeto Linguateca, um centro de recursos distribuído para o processamento computacional da língua portuguesa (Santos, 2009). A partir da Literateca já foi estudado, por exemplo, a classificação computacional de obras em períodos literários.

Há trabalhos também que propõe métodos computacionais para identificação de paralelismos em obras literárias da língua portuguesa. Um exemplo é o trabalho de Carvalho, Loula e Queiroz (2020), que identifica padrões métricos de versificação em prosa literária brasileira através do *MIVES (Mining Verse Structure)*, um sistema computacional. Outro exemplo é o método computacional desenvolvido por Lima, Loula e Queiroz (2021), que realiza a identificação de padrões fonéticos em textos de prosa brasileira através do *ALLPRO (Mining Alliteration in Prose)*.

Embora estes estudos anteriores tenham avançado na identificação de padrões em diferentes níveis linguísticos, ainda há poucos trabalhos que exploram as relações entre esses níveis, especialmente na prosa literária. Este artigo propõe e demonstra um método computacional para identificar padrões textuais interníveis, definidos aqui como concorrência de fenômenos de interesse de níveis linguísticos distintos em um mesmo trecho de texto literário em prosa. A abordagem aqui apresentada permite localizar quantitativamente, ao longo de uma obra, trechos em que esses padrões ocorrem, oferecendo uma ferramenta sistemática para

a observação de aspectos linguísticos que, tradicionalmente, dependiam apenas de análise qualitativa. Assim, espera-se que esse método computacional possa trazer novas perspectivas para estudos literários, promovendo a elaboração de novas perguntas e possibilitando análises quantificáveis em estudos literários.

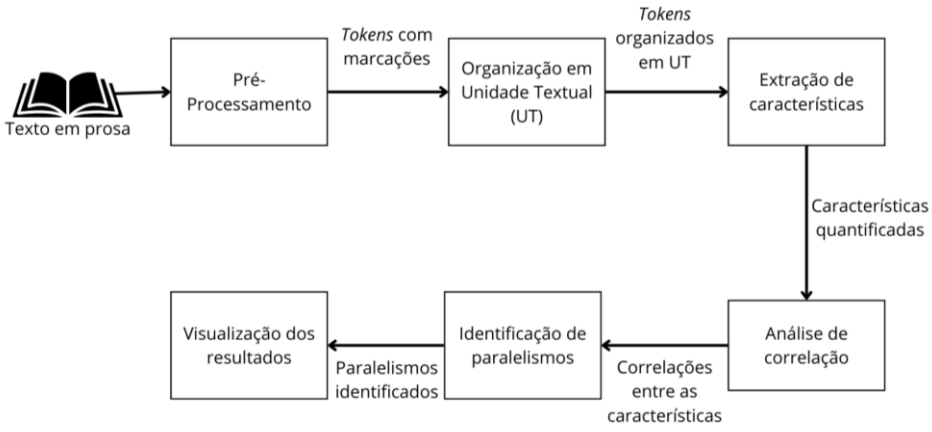
Identificação computacional de padrões interníveis

A identificação de um padrão textual internível envolve o processamento computacional do texto da prosa literária para extrair características linguísticas de diferentes níveis e analisar se estes se relacionam de alguma forma. A aplicação de métodos computacionais com a capacidade de quantificar e correlacionar características extraídas de diferentes níveis linguísticos possibilita evidenciar estes padrões.

Assim, propomos um método computacional baseado em análise de correlações entre características extraídas de diferentes níveis linguísticos, oferecendo uma abordagem quantitativa que evidencia padrões interníveis por meio da quantificação dessas correlações. Com o uso de ferramentas computacionais, torna-se possível quantificar elementos textuais variados, como sílabas fonéticas, entidades nomeadas, polaridades emocionais e padrões métricos.

O método computacional proposto segue um fluxo estruturado, desde a entrada do texto em prosa até a identificação de padrões literários e a visualização dos resultados, conforme a Figura 1 que apresenta esse fluxo com uma visão geral do processo metodológico utilizado.

Figura 1 – Fluxograma do funcionamento do método computacional



Fonte: elaborada pelos autores (2025).

Conforme demonstrado na Figura 1, com a entrada do texto, inicia-se o pré-processamento, que consiste na *tokenização* de palavras e pontuações. Cada *token* recebe marcações associadas à sua classe gramatical, entidade nomeada, lema, polaridade e emoção. A polaridade refere-se à orientação sentimental de uma palavra ou trecho, classificando-a como positiva, negativa ou neutra. Já as emoções são classificações mais específicas de estados afetivos, como alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza, que são atribuídas aos *tokens*. Essas marcações são realizadas com o auxílio de dicionários léxicos, permitindo a quantificação dessas características.

Posteriormente, os *tokens* são organizados em Unidades Textuais (UTs), representando trechos do texto a serem analisados de forma individual como uma unidade textual. Essas UTs podem variar em tamanho, abrangendo desde sentenças ou parágrafos até seções maiores, como capítulos. A escolha do tamanho da UT (em número de sentenças) depende dos objetivos da análise e das características específicas do estudo.

Na etapa de extração de características, propriedades textuais de diferentes níveis são transformadas em valores quantitativos associados a

cada UT. Esses dados quantitativos alimentam a etapa de análise de correlação, na qual as relações entre as características textuais são avaliadas. A identificação de padrões interníveis é então realizada com base nos valores de correlação, permitindo a análise global (texto completo) ou segmentada (trechos localmente delimitados). Por fim, os resultados são visualizados por meio de tabelas e gráficos, destacando as correlações e os padrões textuais identificados.

a) Divisão das Características em Níveis

O método computacional proposto visa identificar padrões textuais em diferentes níveis, o que requer a organização das características extraídas em categorias distintas.

Ao todo foram utilizadas 78 características organizadas em oito níveis:

- Nível Fonético (3 características) - Inclui características relacionadas à fonética do texto: Repetições de sílabas fonéticas;
- Nível Métrico (3 características) - Concentra-se na estrutura de versificação do texto: Padrões métricos em sentenças completas, início e final de sentenças;
- Nível Lexical (5 características) - Engloba características relacionadas ao vocabulário textual: contagem de palavras, contagem de lemas únicos, contagem de palavras únicas e *Type-Token Ratio* (TTR) das palavras e TTR dos lemas;
- Nível Gramatical (22 características) - Inclui características ligadas às categorias gramaticais das palavras: contagem de: adjetivos, advérbios, artigos, conjunções, interjeições, numerais, pontuações, pronomes, preposições, substantivos e verbos; frequência relativa (em relação à quantidade de *tokens* total): de adjetivos, de advérbios, de artigos, de conjunções, de interjeições,

de numerais, de pontuações, de pronomes, de preposições, de substantivos e de verbos;

- Nível Sentimental (24 características) - Considera características relacionadas à análise de sentimentos. A análise de sentimentos é o processo computacional de identificação, extração e classificação de informações afetivas expressas em textos. Nesse contexto, a polaridade refere-se à orientação sentimental de uma palavra ou trecho, classificando-a como positiva, negativa ou neutra. Já os rótulos de emoção, como alegria, confiança, expectativa, medo, nojo, raiva, surpresa e tristeza, são classificações mais específicas de estados afetivos, atribuídas aos tokens. Essas marcações, tanto de polaridade quanto de emoção, são realizadas com o auxílio de dicionários léxicos. A carga emocional da Unidade Textual (UT), por sua vez, é uma medida que quantifica a intensidade da emoção presente em um trecho, somando os valores de emoção de cada token na UT, permitindo uma análise mais detalhada da expressividade do texto. Dessa forma, incluímos as características: contagem de *tokens*: positivos, negativos, neutros, de alegria, de confiança, de expectativa, de medo, de nojo, de raiva, de surpresa e de tristeza; polaridade da UT, carga emocional da UT, frequência relativa (em relação à quantidade de *tokens* total): de positivos, de negativos, de neutros, de alegria, de confiança, de expectativa, de medo, de nojo, de raiva, de surpresa e de tristeza;
- Nível Gramatical e Sentimental (8 características) - Agrupa características que combinam aspectos gramaticais e sentimentais: contagem de adjetivos positivos, de adjetivos negativos, de advérbios positivos, de advérbios negativos, polaridade de adjetivos, polaridade de advérbios, carga emocional de adjetivos e carga emocional de advérbios;

- Nível de Entidades Nomeadas (8 características) - Inclui características relacionadas ao reconhecimento de entidades nomeadas. O reconhecimento de entidades nomeadas (REN) é uma subárea do processamento de linguagem natural que se dedica à identificação e classificação de entidades do mundo real em um texto. No contexto deste trabalho, o REN categoriza palavras e expressões em classes predefinidas, como pessoa, local ou outras entidades nomeadas gerais como organizações, por exemplo. A partir desse processo, o método computacional pode quantificar as ocorrências dessas entidades, gerando as características: contagem de pessoas, de locais, de pessoas e locais e de entidades gerais; frequência relativa (em relação à quantidade de *tokens* total): de pessoas, de locais, de pessoas e locais e de entidades gerais;
- Nível de Tópicos (5 características) - Aborda características relacionadas à modelagem de tópicos no texto. A modelagem de tópicos é uma técnica que permite identificar temas recorrentes e abstratos em um conjunto de textos. Para transformar os tópicos identificados em características, a quantificação é a relevância de cada tópico na UT. O método foi configurado para encontrar os 5 principais tópicos, portanto esse nível possui 5 características, cada uma representando a relevância do tópico em cada uma das UTs.

Essa categorização permite a análise de diferentes aspectos do texto e como eles podem se relacionar, estabelecendo padrões textuais entre esses diferentes níveis. A escolha de características assim como a organização das características em níveis podem ser modificadas, conforme as necessidades específicas de cada pesquisa, trazendo flexibilidade ao método computacional proposto.

b) Extração de Características

A extração de características tem sido aplicada em diversas análises textuais. Este procedimento envolve a transformação de um texto em uma representação numérica, para isso se faz necessária a implementação de uma etapa de pré-processamento. Importante ressaltar que ao enviar um texto para o método computacional proposto, todo ele passa por essa etapa de pré-processamento. Sendo assim, ao submeter uma obra literária inteira, ela será processada e serão extraídas características em todo o texto.

No presente estudo, a etapa de pré-processamento, que inclui a *tokenização*, a marcação de classes gramaticais e entidades nomeadas, além dos lemas das palavras, foi realizada seguindo e implementando a abordagem proposta por Ferreira, Oliveira e Rodrigues (2019). O processo envolve:

- *Tokenização*: Divisão de um texto em unidades menores chamadas *tokens* (palavras e pontuações).
- Marcação de Classes Gramaticais (*PoS Tagging*): Identificação dos tokens como substantivo, adjetivo, advérbio etc.
- Marcação de Entidades Nomeadas: Categorização de entidades nomeadas (ex.: pessoa, lugar, organização).
- Lematização: Redução de palavras à sua forma básica.

Um exemplo de marcação dessas características pode ser observado no Quadro 1, que apresenta a identificação de classes gramaticais, lemas e entidades nomeadas em uma sentença do livro Os Sertões.

Quadro 1 – Marcação de características estilométricas em uma sentença do livro Os Sertões.

Sentença	*Token*	Classe	Lema	Entidade
Em todos prevaleciam os mesmos elementos, que eram o desespero de Diogo Coelho.	Em	prp	em	-
	todos	pron-det	todo	
	prevaleciam	v-fin	prevalecer	
	os	art	o	
	mesmos	pron-det	mesmo	
	elementos	n	elemento	
	,	punc	,	
	que	pron-indp	que	
	eram	v-fin	ser	
	o	art	o	
	desespero	n	desespero	
	de	prp	de	
	Diogo	N	diogo	Pessoa
	Coelho	prop	coelho	
	.	punc	.	-

Fonte: elaborado pelos autores (2025).

A identificação das classes gramaticais atribui rótulos específicos a cada *token* como: prp (preposição), n (substantivo comum), N (substantivo próprio), v-fin (verbo finito), punc (pontuação) e outros. Além disso, o lema é registrado para cada *token*, e as entidades nomeadas são marcadas.

Após a marcação dos *tokens* é possível agrupá-los em sentenças a partir do agrupamento desses *tokens* em trechos com base nos sinais de pontuação de final de sentença (!, ?, ., ...). Essa abordagem permite a

contagem de sentenças no texto, organizando a extração de características por sentença e originando um parâmetro ajustável chamado Unidade Textual (UT). Uma UT de tamanho 1 equivale a uma sentença, ao ajustar esse parâmetro para 100, por exemplo, o texto é dividido em blocos de até 100 sentenças. Cada UT carrega uma quantidade de *tokens* e características quantificáveis, como contagem de artigos, quantidade de *tokens*, quantidade de entidades nomeadas, frequência de sentimentos, frequência de emoções etc.

Inicialmente são extraídas as características de cada sentença, agrupadas em três níveis distintos. No nível Lexical são consideradas a contagem total de palavras na sentença, a contagem de lemas únicos, de palavras únicas, o TTR das palavras e o TTR dos lemas. No nível Gramatical são computadas a contagem e a frequência relativa (em relação ao total de *tokens* da sentença) das seguintes classes gramaticais: adjetivos, advérbios, artigos, conjunções, interjeições, numerais, pronomes, preposições, sinais de pontuação, substantivos e verbos. Por fim, no nível de Entidade Nomeada, são extraídas a contagem e a frequência relativa (também em relação ao total de *tokens* da sentença) de menções a pessoas, locais, pessoas e locais e de outras entidades nomeadas em geral.

A extração de características de nível Sentimental e de nível Gramatical e Sentimental, foi realizada a partir da Análise de Sentimentos das sentenças, entendida como o processo computacional de identificação, extração e classificação de informações afetivas expressas em textos (Pang; Lee, 2008). Para isso, utilizamos os dicionários léxicos LIWC (*Linguistic Inquiry and Word Count*) (Pennebaker *et al*., 2001), OpLexicon (*OpinionLexicon*) (Souza; Vieira, 2012), SentiLex-PT 02 (Silva *et al*., 2010) e NrcEmolex (Mohammad; Turney, 2013), que identificam a polaridade (positiva, negativa, neutra) e emoções específicas (alegria, confiança, expectativa, nojo, raiva e surpresa) dos *tokens* ou lemas, para a marcação

e associação de cada um desses *tokens* e lemas nas polaridades e emoções aos quais esses *tokens* ou lemas estão associados. A contagem de nojo', por exemplo, é computada com base na ocorrência de *tokens* ou lemas que são previamente associados a essa emoção no dicionário NrcEmolex (Mohammad; Turney, 2013). As palavras aberração, fedor e repugnância, caso presentes no texto, são *tokens* que seriam classificados como 'nojo' a partir do dicionário utilizado.

Em seguida, são extraídas características de nível fonético e de nível métrico. Para o nível métrico, utiliza-se o sistema *MIVES* (Carvalho; Loula; Queiroz, 2020), que reconhece estruturas métricas poéticas na prosa a partir da escansão silábica e da aplicação de fenômenos fonológicos inter e intravocabulares, como sinalefa e diérese. A partir disso, são extraídas três características: contagem de versos localizados no início da sentença, contagem de versos localizados no final da sentença e contagem de versos localizados em uma sentença completa. Para o nível fonético, utiliza-se o sistema *ALLPRO* (Lima; Loula; Queiroz, 2021), responsável por identificar repetições sonoras (como aliterações), baseando-se na conversão grafema-fonema e na segmentação silábica do texto, identificando repetições silábicas entre palavras separadas por 2, 3 ou 4 posições. Com isso são extraídas três características distintas: contagem de repetições silábicas com distância de duas palavras, contagem de repetições silábicas com distância de três palavras e contagem de repetições silábicas com distância de quatro palavras.

O Quadro 2 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 identificadas 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 (#).

Quadro 2 - Exemplo de sentença escandida e classificada 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)

Fonte: elaborado pelos autores (2025).

Já o Quadro 3 ilustra as repetições identificadas na sentença de número 1247 do livro "Os Sertões", indicando a sílaba repetida entre colchetes ([ ]).

Quadro 3 – 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. Silábica [a]	a conquistar um di[a]. [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.

Fonte: elaborado pelos autores (2025).

Nesse processo, uma mesma sentença pode ser contada várias vezes caso apresente mais de uma repetição silábica, como na sentença nº 1247, que exibe três tipos de repetições distintas.

Ao reunir as sentenças em UTs maiores, é possível incluir nas análises características de nível semântico por meio da modelagem de tópicos. Essa técnica permite identificar temas recorrentes no texto e avaliar a distribuição e a relevância de cada um ao longo da obra. A modelagem de tópicos, como proposto por Blei, Ng e Jordan (2003) no algoritmo *Latent Dirichlet Allocation* (LDA), opera a partir da análise estatística das ocorrências de termos, agrupando-os em tópicos latentes. Assim, os

tópicos se tornam características textuais quantificáveis e integráveis às análises interníveis realizadas.

c) Análise de Correlação e Identificação de Padrões Interníveis

O método computacional proposto baseia-se na análise de correlação entre as ocorrências das características extraídas. Após a extração e quantificação das características linguísticas para cada Unidade Textual (UT) , a análise de correlação é realizada para avaliar a relação de coocorrência entre essas características. Para isso, é utilizado o coeficiente de correlação de Pearson, aplicado aos conjuntos de dados gerados pelas UTs. A correlação pode ser calculada de forma global, considerando todo o texto, ou de forma segmentada, através de janelas deslizantes que percorrem o texto e calculam a correlação em trechos sucessivos. Essa técnica de janelas deslizantes divide o texto em blocos de tamanho fixo que se movem de forma incremental, permitindo o cálculo da correlação em cada um desses blocos. Sendo assim, a abordagem segmentada permite identificar variações locais na relação entre as características, revelando padrões que se manifestam em passagens específicas da obra.

A correlação representa uma relação ou associação entre dois fenômenos. No contexto da estatística, a correlação é uma medida que descreve a força e a direção da relação linear entre duas variáveis quantitativas. Em outras palavras, ela indica o quão bem os valores de uma variável se movem em relação aos valores de outra variável (Larson; Farber 2015). A análise de correlação fornece uma medida objetiva da conexão entre variáveis, evitando interpretações subjetivas e oferecendo uma avaliação numérica confiável da relação entre elas.

O coeficiente de correlação varia entre -1 e 1, refletindo a intensidade e a direção da relação: valores próximos de 1 indicam uma forte correlação linear positiva, enquanto valores próximos de -1 sugerem

uma forte correlação negativa. Na ausência de uma correlação linear, o valor tende a zero, mas é importante destacar que um valor próximo de zero indica apenas a inexistência de uma relação linear, não excluindo outras formas de associação.

As correlações de Pearson e Spearman calculam um valor único que representa a relação entre duas variáveis, considerando o texto como um todo. Contudo, em análises literárias, pode ser necessário observar como essa relação varia ao longo do texto. Para isso, utiliza-se o cálculo de correlação por janelas deslizantes, que segmenta o texto em trechos sucessivos e calcula a correlação entre as características em cada trecho. Essa técnica permite identificar variações locais na relação entre as características, possibilitando observar, por exemplo, se uma correlação forte se mantém ao longo da narrativa ou se ocorre apenas em passagens específicas do texto.

Apesar de uma correlação forte entre duas variáveis indicar uma relação consistente, ela não implica causalidade. Portanto, uma análise mais aprofundada é essencial para determinar se existe uma relação de causa e efeito.

No método proposto, a definição de padrões interníveis, baseia-se em um limiar configurável, que representa o valor mínimo absoluto do coeficiente de correlação necessário para considerar esta correlação entre as características como um padrão internível. O valor padrão de 0,75 foi adotado, pois valores iguais ou acima deste representam correlações fortes (Larson; Farber, 2015). Importante ressaltar que o método foca nas correlações entre características de diferentes níveis, desconsiderando relações entre características do mesmo nível.

A aplicação do método envolve a análise combinatória de pares de características, permitindo a identificação de padrões textuais em segmentos distintos de um texto literário. Assim, é possível diversificar o

tamanho das janelas deslizantes e o número de unidades de texto analisadas, o que facilita uma análise detalhada e adaptável ao conteúdo.

Essa abordagem permite não apenas identificar padrões textuais complexos, mas também investigar como aspectos linguísticos distintos interagem na obra estudada, revelando interdependências e padrões que podem enriquecer a interpretação literária.

Resultados

Para exemplificar o método computacional proposto na identificação de padrões interníveis em textos literários, aplicamos o método computacional ao livro *Os Sertões* de Euclides da Cunha, obra de grande relevância para a literatura brasileira e frequentemente utilizada em estudos literários. Para esta exemplificação foram utilizados como parâmetros o tamanho da UT igual a 10, significando que cada UT possui 10 sentenças e o tamanho da janela igual a 10, ou seja, cada janela possui ao todo 100 sentenças. Além disso, para uma exemplificação mais concisa foram utilizados apenas os coeficientes de correlação de Pearson. A seguir, são apresentados os principais achados, evidenciando as correlações entre diferentes níveis linguísticos.

Uma das formas de identificação de padrões interníveis pode ser feita a partir da correlação global entre as características. Essa técnica calcula a correlação a partir da extração das características considerando todo o livro.

Entretanto, esses padrões textuais podem ocorrer apenas em alguns trechos do livro, não caracterizando necessariamente uma regularidade presente ao longo de toda a narrativa. Dessa forma, a utilização das janelas deslizantes para o cálculo de correlação segmentada permite a identificação de padrões interníveis localizados, que se manifestam em determinados trechos do livro.

Para que uma correlação segmentada seja identificada como um padrão internível, ela precisa atingir um limiar mínimo absoluto. Para essa análise, o valor definido foi de 0,75, a fim de identificar apenas padrões textuais oriundos de correlações fortes, que podem estar acima de +0,75 ou abaixo de -0,75.

Figura 2 – Correlação segmentada entre as características “Contagem de Palavras” e “Contagem de Substantivos”

Fonte: elaborada pelos autores (2025).

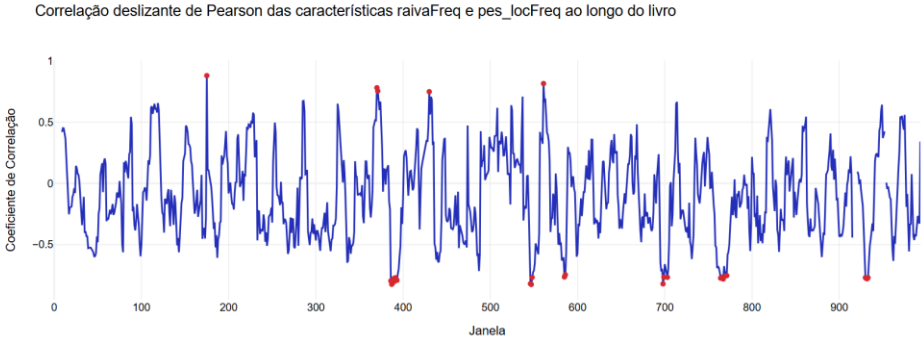
A correlação segmentada entre as características “Contagem de Palavras” e “Contagem de Substantivos”, exposta na Figura 2, demonstra, por meio de pontos vermelhos, os trechos do livro Os Sertões onde a correlação entre essas características ultrapassa o limiar absoluto de 0,75. Isso revela um padrão de correlação em grande parte do livro, o que pode é corroborado pela correlação global entre essas características, que possui o valor de 0,9577. Esta identificação, contudo, pode ter resultado de uma correlação espúria, considerando que trechos com maior número de palavras tendem, naturalmente, a apresentar também um maior número de substantivos.

Para oferecer uma outra perspectiva, o método computacional proposto também considera a frequência relativa à quantidade de palavras em cada trecho para o cálculo da correlação, fazendo com que esse cálculo seja menos influenciado pelo comprimento do trecho.

A aplicação do método computacional revelou padrões interníveis significativos entre diferentes níveis linguísticos, evidenciados por correlações específicas entre as características extraídas. Utilizando o método proposto, foi possível identificar uma correlação entre as características “frequência da emoção raiva” (nível “Sentimental”) e “frequência de entidades nomeadas de pessoas e/ou locais” (nível “Entidades nomeadas”). Por outro lado, aplicando o coeficiente de correlação de Pearson, obteve-se um valor global de -0,1224. Esse valor representa uma correlação negativa e fraca, indicando a inexistência de uma relação consistente entre as duas características quando considerada a narrativa completa. Sendo assim, esse par de características não atende aos critérios de identificação de padrões interníveis, conforme o limiar estabelecido e as características de uma correlação forte.

Entretanto, ao se observar esse resultado em um gráfico de linhas realizado a partir de correlações segmentadas (Figura 3), é possível identificar alguns padrões textuais, representados pelos pontos vermelhos que indicam as janelas nas quais a correlação ultrapassa o limiar estabelecido.

Figura 3 – Correlação segmentada entre as características “frequência da emoção raiva” e “frequência de entidades nomeadas de pessoas e/ou locais” utilizando a correlação de Pearson



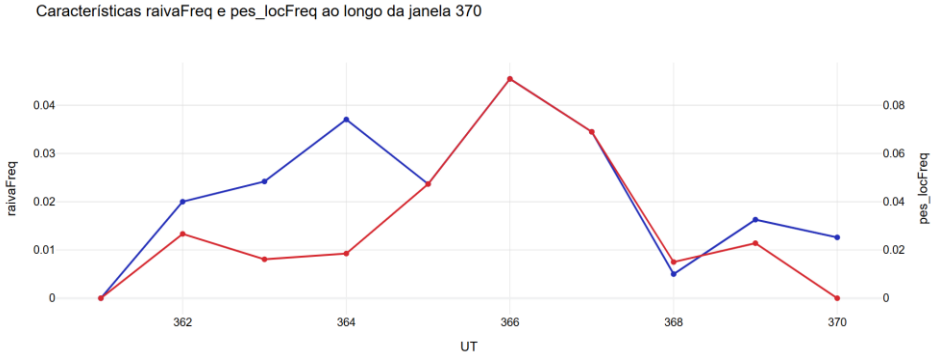
Fonte: elaborada pelos autores (2025).

Ao analisar a correlação utilizando janelas deslizantes, identificou-se o valor de 0,8809 na janela 175, representando o coeficiente de correlação máximo para esse par de características. Por outro lado, o coeficiente mínimo registrado foi de -0.8454 (janela 547), indicando que essa correlação varia consideravelmente em diferentes partes do texto. Isso demonstra que o método computacional proposto permite encontrar correlações específicas em determinados trechos do livro, evidenciando que nem sempre os padrões interníveis podem ser identificados com base apenas na correlação global.

Além disso, as janelas 370 e 371 apresentam padrões textuais em trechos consecutivos, com correlações positivas e fortes (0,7819 e 0,7551, respectivamente), sugerindo que o aumento da frequência de raiva pode estar relacionado com o aumento da frequência de entidades nomeadas de pessoas e/ou locais. Isso indica a presença de um trecho na obra onde o padrão textual é mais evidente e, consequentemente, mais facilmente percebido durante a leitura.

É possível visualizar os valores dessas características nas UTs que compõem as janelas por meio de um gráfico de linhas, conforme representado na Figura 4. Na linha azul estão os valores da característica “frequência da emoção raiva” (*raivaFreq*) e, na linha vermelha, os valores da característica “frequência de entidades nomeadas de pessoas e/ou locais” (*pes_locFreq*). A partir da figura, observa-se uma correlação entre essas duas características, uma vez que os valores apresentam trechos com variações semelhantes, aumentando e diminuindo de maneira paralela.

Figura 4 – Gráfico de linhas entre as características “frequência da emoção raiva” e “frequência de entidades nomeadas de pessoas e/ou locais” das UTs componentes da janela 370



Fonte: elaborada pelos autores (2025).

Para possibilitar a visualização dos trechos, o método computacional também exibe os textos que correspondem à janela, permitindo ao usuário identificar as marcações nos trechos correspondentes. A exemplo disso, a janela 370 compreende as sentenças 3610 a 3709, iniciando na UT 361 e finalizando na UT 370, incluindo a UT 365, onde podem ser identificadas palavras que expressam raiva (em azul) e entidades nomeadas de pessoas ou locais (em laranja):

> “*Mas sobre as bandeiras vindas de todos os pontos , do extremo norte e do extremo sul , do Rio Grande ao Amazonas , pairou sempre , intangível , miraculosamente erguida pelos exegetas constitucionais , a soberania do Estado ...Para a resguardar melhor foi removido da Bahia o chefe da força militar , que traçara a sua atitude retilineamente pela lei .E somente depois disto a coluna do major Febrônio - até então oscilante entre Monte Santo e Queimadas e objetivando nas contramarchas as vacilações do governo - seguiu reforçada pela tropa policial e adstrita as deliberações do governo baiano .Perdera-se esterilmente o tempo - que o adversário aproveitara , aparelhando-se a um revide enérgico .Num raio de três léguas em roda de Canudos , fizera-se o deserto .Para todos os rumos e por todas as estradas e em todos os lugares , os escombros*

> *carbonizados das fazendas e dos pousos avultavam , insulando o arraial num grande círculo isolador , de ruínas .Estava pronto o cenário para um emocionante drama da nossa história .147 Travessia do Cambaio I.Monte Santo .Triunfos antecipados .”*

Esse resultado permite visualizar um possível padrão de correlação internível neste trecho, onde a característica do nível sentimental “frequência de raiva” e a característica de nível de entidades nomeadas “frequência de entidades nomeadas de pessoas e/ou locais” parecem estar relacionadas conforme a similaridade gráfica apresentada na Figura 4.

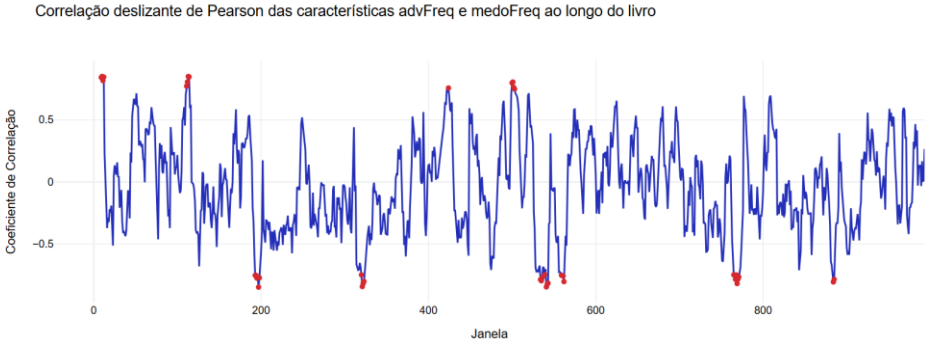
Embora a coocorrência de emoções com a menção de entidades nomeadas possa parecer intuitiva em narrativas, o valor da abordagem computacional reside na sua capacidade de quantificar e localizar com precisão a intensidade e a dinâmica dessa relação ao longo da obra. A correlação global fraca (-0,1224) para este par de características, em contraste com pontos de forte correlação positiva (ex: 0,8809 na janela 175) e negativa (ex: -0,8454 na janela 547) em trechos específicos, demonstra que a relação entre a frequência de raiva e a menção de pessoas/locais não é constante. Pelo contrário, ela se manifesta de forma variável e localizada, evidenciando momentos de intensificação emocional diretamente associados à presença de personagens ou cenários. Essa capacidade de mapear a flutuação desses padrões interníveis oferece um nível de detalhe analítico que possibilita ao pesquisador identificar essas associações e investigar suas implicações literárias nos trechos correspondentes.

Outro padrão textual identificado a partir do nível sentimental foi observado na correlação entre a característica “frequência de medo” e a característica “frequência de advérbios”, pertencente ao nível gramatical. A correlação global entre essas duas características foi de -0,1264, caracterizando uma correlação nula, o que, assim como no exemplo

anterior, indica que, ao considerar a obra como um todo, não há evidência de correlação significativa, afastando a hipótese de um padrão internível recorrente. Entretanto, a partir da correlação segmentada é possível identificar trechos onde há um certo padrão entre essas características.

A Figura 5 apresenta a correlação segmentada entre as características “frequência de medo” (nível Sentimental) e “frequência de advérbios” (nível Gramatical), calculada com o coeficiente de Pearson. O gráfico evidencia janelas (representadas pelos pontos em vermelho) nas quais essas duas características apresentam variação semelhante, indicando padrões textuais localizados em que expressões de medo tendem a coexistir com o uso de advérbios, sugerindo uma possível associação estilística entre aspectos emocionais e marcas gramaticais do texto.

Figura 5 – Correlação segmentada entre as características “frequência de advérbios” e "frequência da emoção medo” utilizando a correlação de Pearson



Fonte: elaborada pelos autores (2025).

Conforme demonstrado na Figura 5, a análise segmentada revela que, em trechos específicos do texto, essas características podem apresentar variações conjuntas. Como exemplo, a janela de número 114 apresentou uma correlação de 0,8459 entre essas duas características, ultrapassando o limiar estabelecido de 0,75 e, portanto, caracterizando um padrão internível localizado. Com base nesse resultado, é possível realizar a leitura do trecho correspondente, destacando em azul os advérbios e, em

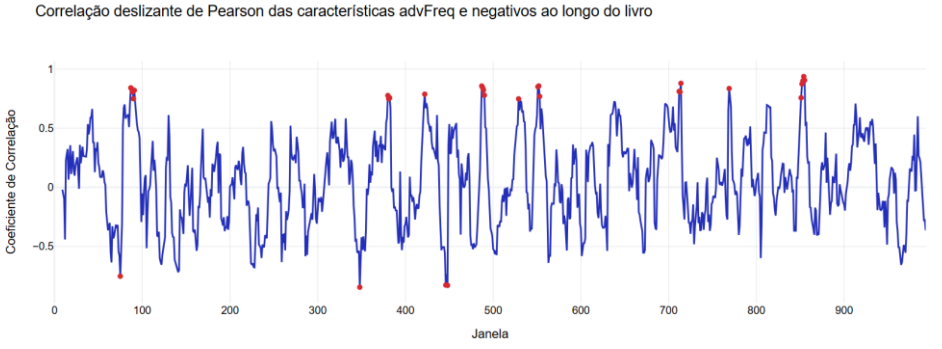
laranja, as palavras associadas à emoção medo presentes nas sentenças em uma das UTs que compõe a janela 114:

> “encanta-o pelo aspecto formosíssimo .arrebata-o , afinal , irresistivelmente na correnteza dos rios .Daí o traçado eloqüentíssimo do Tietê , diretriz preponderante nesse domínio do solo .Enquanto no S.Francisco , no Paraíba , no Amazonas , e em todos os cursos d'água da borda oriental , o acesso para o interior seguia ao arrepio das correntes , ou embatia nas cachoeiras que tombam dos socalcos dos planaltos , ele levava os sertanistas , sem uma remada , para o rio Grande e daí ao Paraná e ao Paranaíba .Era a penetração em Minas , em Goiás , em Santa Catarina , no Rio Grande do Sul , no Mato Grosso , no Brasil inteiro .Segundo estas linhas de menor resistência , que definem os lineamentos mais claros da expansão colonial , não se opunham , como ao norte , renteando o passo às bandeiras , a esterilidade da terra , a barreira intangível dos descampados brutos .Assim é fácil mostrar como esta distinção de ordem física esclarece as anomalias e contrastes entre os sucessos nos dois pontos do país , sobretudo no período agudo da crise colonial , no século 17 .Enquanto o domínio holandês , centralizando-se em Pernambuco , reagia por toda a costa oriental , da Bahia ao Maranhão , e se travavam recontros memoráveis em que , solidárias , enterreiravam o inimigo comum as nossas três raças formadoras , o sulista , absolutamente alheio àquela agitação , revelava , na rebeldia aos decretos da metrópole , completo divórcio com aqueles lutadores .Era quase um inimigo tão perigoso quanto o batavo.”

A análise da característica “frequência de advérbios” também revelou correlações significativas com a característica “contagem de palavras

negativas” em alguns trechos da obra, configurando padrões interníveis entre os níveis gramatical e sentimental.

Figura 6 – Correlação segmentada entre as características “frequência de advérbios” e "contagem de palavras negativas” utilizando a correlação de Pearson



Fonte: elaborada pelos autores (2025).

A Figura 6 exibe a correlação segmentada entre as características “frequência de advérbios” (nível gramatical) e “contagem de palavras negativas” (nível sentimental), utilizando o coeficiente de Pearson. Os pontos em vermelho indicam janelas do texto em que essas duas variáveis apresentam comportamento semelhante, apontando para possíveis padrões textuais em que elementos gramaticais se articulam com conteúdos de sentimentos de polaridade negativa. Esse tipo de associação pode sinalizar trechos nos quais o uso de advérbios contribui para a ênfase ou intensificação de expressões de cunho negativo.

Essa associação sugere uma possível relação recorrente entre o uso de advérbios e expressões de sentimentos de polaridade negativa, que pode ser objeto de investigações futuras a partir de análises mais profundas. Tais correlações reforçam a hipótese de que determinados fenômenos linguísticos tendem a coocorrer em trechos específicos da narrativa. Como exemplo, no trecho correspondente à UT 381, é possível visualizar as marcações em azul, correspondentes aos advérbios, e em laranja, correspondentes às palavras negativas identificadas nas sentenças dessa UT:

> “Às aventuras de um plano temerário , resumindo-se numa investida e num assalto , substituiria operação mais lenta e mais segura .Não fez isto .Fez o inverso : depois de longa inatividade em Monte Santo , a expedição partiu ainda menos aparelhada do que quando ali chegara quinze dias antes , abandonando , ainda uma vez , parte dos restos de um trem de guerra já muitíssimo reduzido .Entretanto , contravindo ao modo de ver dos propagandistas de uma vitoria fácil , chegavam constantes informações sobre o numero e recursos dos fanáticos .E no disparatado das opiniões - entre as que elevavam aquele , no máximo , a quinhentos .e as que o firmavam , decuplicando-o , no mínimo , em cinco mil , cumpria inferir-se uma média razoável .Além disto , de envolta num sussurrar de cautelosas denuncias e mal boquejados avisos , esboçava-se a hipótese de uma traição .Apontavam-se influentes mandões locais , cujas velhas relações com o Conselheiro sugeriam , veemente , a presunção de que o estivessem auxiliando a socapa , fornecendo-lhe recursos e 153 instruindo-o dos menores movimentos da investida .Ainda mais , sabia-se que a tropa , quando mesmo o maior sigilo rodeasse as deliberações , seria , no avançar , precedida e ladeada pelos espias espertos do inimigo , muitos dos quais , verificou-se depois , dentro da própria vila acotovelavam os expedicionários .Uma surpresa , depois de tantos dias perdidos e em tais circunstâncias , era inadmissível.”

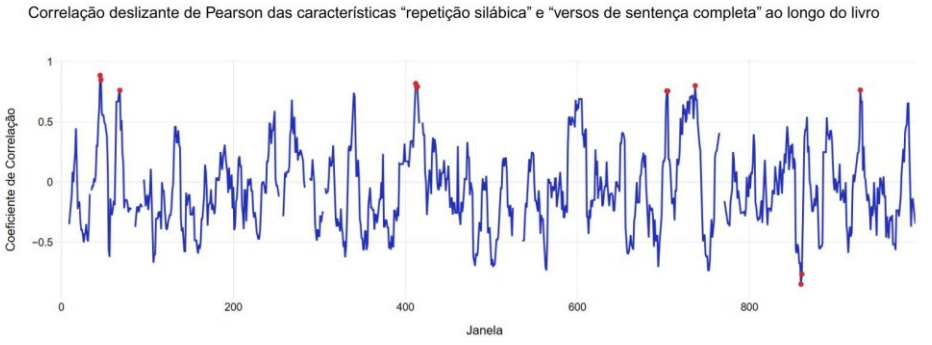
Após essa análise textual aproximada entre a “frequência de advérbios” e a “contagem de palavras negativas” é possível observar que embora a coocorrência dessas características possa ser, em parte, influenciada pela sobreposição gramatical dos advérbios de negação com as palavras de polaridade negativa, os resultados revelam que essa associação não é constante ao longo de toda a obra. A baixa correlação global, em contraste com os picos de forte correlação localizados (conforme evidenciado na

Figura 6), sugere que o uso conjunto de advérbios e expressões negativas não é uma característica presente em toda a obra, mas sim algo que acontece em momentos específicos da narrativa. O exemplo da UT 381 ilustra como o método destaca os trechos onde a combinação desses elementos pode indicar uma a construção atmosférica de crítica, incerteza ou desespero, auxiliando na interpretação e associação interníveis que poderiam passar despercebidas em uma leitura manual. Esses achados reforçam a utilidade da análise segmentada para a compreensão de como diferentes níveis linguísticos interagem dentro do texto.

A partir do MIVES (Carvalho; Loula; Queiroz, 2020) foi possível identificar sentenças métricas no livro Os Sertões. Já com o uso do ALLPRO (Lima; Loula; Queiroz, 2021), identificaram-se as repetições fonéticas em diferentes partes do texto. A utilização dessas duas ferramentas permitiu localizar trechos do texto nos quais marcações do nível fonético e marcações do nível métrico ocorrem em uma mesma sentença.

A Figura 7 apresenta a correlação segmentada entre as características “repetição silábica” (nível Fonético) e “versos de sentença completa” (nível Métrico), calculada com o coeficiente de Pearson. O gráfico evidencia janelas (representadas pelos pontos em vermelho) nas quais essas duas características apresentam variação semelhante, sugerindo padrões textuais em que estruturas fonológicas e métricas podem aparecer de forma simultânea no texto, inclusive em uma mesma sentença.

Figura 7 – Correlação segmentada entre as características “repetição silábica” e “versos de sentença completa” utilizando a correlação de Pearson



Fonte: elaborada pelos autores (2025).

Como exemplo, o método computacional identificou o trecho destacado no Quadro 4, se observa a presença de repetições silábicas (identificadas via ALLPRO) em sentenças métricas (identificadas via MIVES).

Quadro 4 – Sentenças 453 a 455, presentes na UT 45

Sentença	Escansão	Classificação métrica	Repetições silábicas
Ao passo que a caatinga o afoga;	Ao/ p#a/sso/ que a/ caa/t#in/ga o a/f#o/ga;	Octossílabo	Ao passo que [a] ca[a]tinga o [a]foga;
abrevia-lhe o olhar;	a/bre/v#i/a-/lhe/ o o/lh#ar;	Heptassílabo	[a]brevi[a]- lhe o olhar;
agride-o e estonteia-o	a/gr#i/de-/o e es/ton/t#e/i/a-/o.	Heptassílabo	-

Fonte: elaborado pelos autores (2025).

A presença de repetições silábicas em Os Sertões não é um fenômeno isolado. O método computacional para identificação de paralelismos fonológicos na prosa literária brasileira, que serve de base para algumas das características extraídas neste estudo, foi aplicado a outras obras importantes da literatura nacional. O estudo de Lima, Loula e Queiroz (2022) evidencia a presença de sequências repetidas não apenas em Os Sertões, mas também em Triste Fim de Policarpo Quaresma, de Lima

Barreto, e em Macunaíma, de Mário de Andrade. A análise comparativa revelou que, embora o número absoluto de sequências varie entre as obras, a densidade relativa de paralelismos fonológicos é mais alta em Os Sertões e Triste Fim de Policarpo Quaresma. Tal achado sugere que este não é um fenômeno restrito a um único autor, mas uma característica da prosa em língua portuguesa, cuja frequência e distribuição podem ser utilizadas para caracterizar estilos de autores e até mesmo movimentos literários, abrindo caminho para comparações mais abrangentes.

O método computacional proposto, entretanto, não realiza a análise textual qualitativa de forma automática. Cabe ao usuário interpretar os resultados e verificar se o padrão identificado é de fato relevante ou se trata apenas de uma coincidência, um recurso estilístico ou algo característico do gênero textual ao qual a obra se classifica. O valor do método reside precisamente na capacidade de evidenciar, de forma quantitativa, a coocorrência de fenômenos linguísticos. Ao invés de deixar a descoberta de tais padrões para a intuição de uma leitura manual, a ferramenta computacional direciona o usuário para os trechos onde esses fenômenos se manifestam com maior intensidade, transformando uma intuição em um ponto de partida guiado por dados. Assim, o método não apenas assiste a análise crítica, mas a enriquece ao possibilitar a formulação de novas perguntas e a elaboração de hipóteses sobre a construção estilística e o significado da obra.

Como exemplo, tem-se a correlação entre as características “Verso de final de sentença” e “Frequência de Nojo”, cuja correlação global de 0,0764. Entretanto, em um determinado trecho foi observada uma correlação segmentada de 0,8512, indicando aumento simultâneo tanto na quantidade de versos de final de sentença quanto na frequência de nojo (respectivamente destacados em azul e laranja):

> “Revoluteia, brabo e corado , o sertanejo moço .Desafios Nos intervalos travam-se os desafios .Enterreiram-se , adversários , dois cantores rudes .As rimas saltam e casam-se em quadras muita vez belíssimas .Nas horas de Deus , amém , Não é zombaria , não !Desafio o mundo inteiro Pra cantar nesta função !O adversário retruca logo , levantando-lhe o último verso da quadra : Pra cantar nesta função , Amigo , meu camarada , Aceita teu desafio”

Ao analisar esse resultado, nota-se que o aumento da frequência de nojo se deve à repetição da palavra “desafio”, considerada como um possível marcador dessa emoção. Esse fator pode gerar resultados que não representam necessariamente um padrão de correlação entre as características, o que reforça a importância da visualização do texto e da identificação dos trechos correspondentes para a correta interpretação dos dados.

Por fim, embora o método computacional tenha a capacidade de identificar uma vasta quantidade de padrões interníveis, esta seção se concentra na apresentação de um conjunto seleto de exemplos. A escolha desses padrões visou ilustrar a diversidade de relações possíveis entre alguns dos diferentes níveis linguísticos utilizados e demonstrar as diferentes formas da aplicabilidade do método. Os exemplos foram escolhidos para: (a) demonstrar a capacidade de identificar padrões que podem não ser imediatamente óbvios à leitura manual, como, por exemplo, a interação entre fonética e métrica em prosa; (b) ilustrar a dinâmica das correlações ao longo da narrativa, onde padrões fortes podem emergir em trechos específicos apesar de uma correlação global fraca; e (c) abrir caminho para discussões sobre a interpretação de fenômenos linguísticos em seu contexto literário, mesmo quando a correlação é influenciada por aspectos gramaticais ou lexicais, como no caso dos advérbios e palavras negativas.

Assim, o objetivo não foi esgotar todos os padrões possíveis ou listar os mais recorrentes em termos absolutos, mas sim oferecer uma amostra representativa das potencialidades analíticas do método proposto, incentivando futuras investigações mais aprofundadas sobre esses e outros padrões detectados.

Conclusão

Os resultados apresentados aqui não apenas demonstram a aplicabilidade do método proposto na detecção de padrões textuais interníveis, mas também estabelecem as bases para uma abordagem quantitativa e sistemática na análise de textos literários. Ao permitir a correlação entre características de diferentes níveis linguísticos o método amplia o escopo das análises possíveis, fornecendo elementos para novas visões e interpretações de obras literárias em prosa.

Ao analisar o livro "Os Sertões" de Euclides da Cunha, foi possível identificar uma variedade de padrões textuais interníveis distribuídos ao longo da obra. Por exemplo, observou-se a coincidência de versos com estrutura métrica regular e repetições fonológicas em uma mesma sentença, bem como a correlação entre o uso de advérbios e palavras associadas à emoção medo em trechos específicos. Os exemplos demonstrados ilustram a diversidade dos resultados obtidos e evidenciam o potencial da metodologia para destacar particularidades estilísticas e estruturais.

Apesar de termos apresentado apenas alguns exemplos de trechos com padrões interníveis 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 padrões interníveis presentes em textos literários.

Um dos destaques do nosso 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. Ao fornecer uma ferramenta para a detecção de padrões interníveis, o método capacita o pesquisador a investigar a dinâmica e a variabilidade estilística da prosa. É importante ressaltar, contudo, que a caracterização estilística completa de um autor ou a atribuição de autoria de uma obra, embora sejam potenciais aplicações futuras, exigem análises comparativas e um escopo mais abrangente que vai além da proposta do presente trabalho. Nosso objetivo foi, portanto, apresentar e validar o método em um estudo de caso, estabelecendo as bases para que futuras pesquisas possam explorar essas e outras aplicações.

Mesmo não encontrando uma pesquisa ou estudo de uma ferramenta ou método computacional que proponha a aplicação da correlação para fins de identificação de padrões textuais interníveis, 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.

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.; ABAURRE, M. B. M.; PONTARA, M. *Gramática: Texto: análise e construção de sentido.* 2. ed. São Paulo: Moderna, 2011.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. *Journal of machine Learning research*, [S.I.] v. 3, n. Jan, p. 993-1022, 2003.

BUSA, Roberto A. Foreword: Perspectives on the digital humanities. In: SCHREIBMAN, Susan; SIEMENS, Raymond George; UNSWORTH, John (Orgs.). *A companion to digital humanities.* Malden, MA: Blackwell, 2004. p. xvi–xxi.

CARVALHO, R.; LOULA, A. C.; QUEIROZ, J. Identificação computacional de estruturas métricas de versificação na prosa de Euclides da Cunha. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 28, n. 1, p. 41, jan. 2020.

CORSO, G.; FOSSA, C. R.; OLIVEIRA, G. B. de. Uma aplicação da teoria de redes à estilometria: comparando Machado de Assis e Tribuna do Norte. *Revista Brasileira de Ensino de Física*, v. 27, p. 389-393, 2005.

DE ROC BORONAT, C.; WANNER, L. On the relevance of syntactic and discourse features for author profiling and identification. In: *Conference of the European Chapter of the Association for Computational Linguistics*, 15., p.681–687, 2017.

DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. 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*, p. 189–197.

EDER, M.; PIASECKI, M.; WALKOWIAK, T. An open stylometric system based on multilevel text analysis. *Cognitive Studies| Études cognitives*, Warsaw, n. 17, 2017.

FERREIRA, J. J.; OLIVEIRA, H. G; RODRIGUES, R. J. Improving NLTK for Processing Portuguese. *Symposium on Languages, Applications and Technologies*, p. 9, 1 jan. 2019.

GALINA, R.; FLORES, D.; KOMATI, K. Comparação de Atributos Estilométricos para Identificação de Autoria de Escrita: Um Estudo de Caso de Guimarães Rosa versus Clarice Lispector. In: *ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC)*, 16. , 2019, Salvador. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019 . p. 353-364. ISSN 2763-9061. DOI: https://doi.org/10.5753/eniac.2019.9297.

GOODRICH, R. A. On Poetic Function: Jakobson's Revised 'Prague' Thesis. *Literature & Aesthetics*, v. 7, 1997.

JACOBS, A. M. Sentiment analysis for words and fiction characters from the perspective of computational (neuro-) poetics. *Frontiers in Robotics and AI*, [S.I.], v. 6, art. 53, 2019.

JAKOBSON, R.; POMORSKA, K. *Diálogos*. Tradução: Elisa A. Kossovitch. São Paulo: Cultrix, 1985.

JOCKERS, Matthew L. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

LAGUTINA, K.; LAGUTINA, N.; BOYCHUK, E.; PARAMONOV, I. The influence of different stylometric features on the classification of prose by centuries. In: *Conference of Open Innovations Association (FRUCT)*, 27., 2020, p. 108–115. IEEE.

LARSON, R.; FARBER, B. *Estatística Aplicada.* Tradução: José Fernando Pereira Gonçalves. São Paulo: Pearson Education do Brasil, 2015.

LIMA, L.; LOULA, A. C.; QUEIROZ, J. Computational identification of phonological parallelisms in Brazilian literary prose. *Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022)*, p. 47–52, 2022.

MIN, S.; PARK, J. Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, [S.I.], v.14, n. 12, p. e0226025, 2019.

MOHAMMAD, Saif M.; TURNEY, Peter D. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, v. 29, n. 3, p. 436-465, 2013.

MORETTI, Franco. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

PANG, Bo; LEE, Lillian. Opinion Mining and Sentiment Analysis**.** *Information Retrieval*, [S.I.] v. 2, n. 1-2, p. 1-135, 2008.

PENNEBAKER, James W.; BOOTH, Roger J.; FRANCIS, Martha E. *Linguistic Inquiry and Word Count: LIWC2001*. Mahway: Lawrence Erlbaum Associates, 2001.

ROMMEL, T. Literary studies. In: SCHREIBMAN, Susan; SIEMENS, Raymond; UNSWORTH, John (Orgs.) *A Companion to Digital Humanities*. Oxford: Blackwell, 2004, p.89.

SANTOS, D. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, v. 1, n. 1, p. 25-58, 6 abr. 2009.

SANTOS, D.; PIRES, E.; FREITAS, C.; FUÃO, R. S.; LOPES, J. M. Periodização automática: Estudos linguistíco-estatísticos de literatura lusófona. *Linguamática*, v. 12, n. 1, p. 81-95, 29 jun. 2020.

SILVA, Mário J. et al. Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis.Technical Report. TR 10-08. University of Lisbon, Faculty of Sciences. 2010.

SOUZA, Marlo; VIEIRA, Renata. Sentiment analysis on twitter data for portuguese language. In: *International Conference on Computational Processing of the Portuguese Language*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 241-247.

UNDERWOOD, Ted. *Distant horizons: digital evidence and literary change*. Chicago: University of Chicago Press, 2019.

WAUMANS, M. C.; NICOD`EME, T.; BERSINI, H. Topology analysis of social networks extracted from literature. *PloS one*, [S.I.], v. 10, n. 6, p. e0126470, 2015.

NOTAS DE AUTORIA

Angelo Loula (angelocl@uefs.br): Doutor em Engenharia Elétrica pela Unicamp (2011) e mestre em Engenharia Elétrica pela Unicamp (2004), com pesquisas em Inteligência Artificial. Graduação em Engenharia Elétrica pela Universidade Federal da Bahia (2000), graduação em Processamento de Dados pela Faculdade Ruy Barbosa (1997). Atualmente é professor da Universidade Estadual de Feira de Santana, atuando na graduação em Engenharia de Computação (UEFS) e no mestrado em Ciência da Computação (UEFS). Tem experiência na área de Engenharia e Ciência da Computação, com ênfase em Inteligência Artificial, atuando em pesquisa principalmente nos seguintes temas: ciência de dados, mineração de dados e textos, aprendizado de máquina, humanidades digitais, cognição artificial, sistemas cognitivos incluindo representação e comunicação e aplicações interdisciplinares da computação. Tem atuação em educação empreendedora, geração de *startups*, inovação tecnológica e transferência de tecnologia, seja no Núcleo de Inovação Tecnológica da UEFS, em disciplinas, em projetos de extensão, em cursos de extensão e em eventos de capacitação e divulgação.

Luciano Alves Machado Júnior (lucianoamjunior@gmail.com): Mestre em Ciência da Computação pela Universidade Estadual de Feira de Santana (UEFS). Especialista em Projetos de Aplicativos Móveis Multiplataforma pela Faculdade Descomplica. Especialista em Big Data pela Faculdade Descomplica. Graduado em Tecnologia em Análise e Desenvolvimento de Sistemas pelo Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA - Campus Irecê). Técnico em Informática pelo Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA - Campus Irecê). Professor EBTT em Ciência da Computação no IFBA Campus Irecê. Tem interesse nas áreas de Inteligência Artificial, Interação Humano-Computador, Desenvolvimento de Software, Desenvolvimento de Jogos Digitais, Computação Gráfica em desenvolvimento de jogos e animações.

Como citar este artigo de acordo com as normas da revista?
LOULA, Angelo; MACHADO JÚNIOR, Luciano Alves. Identificação computacional de padrões interníveis em textos da Literatura Brasileira. *Texto Digital*, Florianópolis, v. 21, n. 1, p. 284-322, 2025.

Contribuição de autoria
Não se aplica.

Financiamento

Não se aplica.

Consentimento de uso de imagem
Não se aplica.

Aprovação de comitê de ética em pesquisa
Não se aplica.

Histórico
Recebido em: 10/05/2025.
Aprovado em: 20/05/2025.