



Léxico Digital de Guimarães Rosa: uma Aplicação em Processamento de Linguagem Natural para o Ensino de Linguística Computacional

Rodrigo Oliveira do Carmo, PPGCC/UEL, rodrigo.oliveira@uel.br,
<https://orcid.org/0000-0003-0872-1622>

Cinthyán Renata Sachs Camerlengo de Barbosa, PPGCC/UEL, cinthyán@uel.br,
<https://orcid.org/0000-0001-6248-6630>

Lucas Pereira Cassanho, PPGCC/UEL, lucas.cassanho@uel.br,
<https://orcid.org/0009-0004-6232-0461>

Resumo: Este artigo apresenta uma ferramenta computacional para análise léxico-morfológica das obras de Guimarães Rosa, com foco em aplicações educacionais e no ensino de linguística. A partir de um corpus com 63.868 palavras extraídas de suas obras foram integradas informações gramaticais, morfológicas e etimológicas, com base no *Léxico de Guimarães Rosa* e em técnicas de Processamento de Linguagem Natural. O sistema permite explorar o vocabulário por classe gramatical, origem etimológica, frequência e contexto. Para otimizar a busca foram testadas oito funções hash, com destaque para a Zobrist, que obteve o melhor desempenho. Os resultados indicam o potencial de contribuir para o ensino da Língua Portuguesa ao oferecer recursos interativos para a análise linguística do vocabulário literário.

Palavras-chave: processamento de linguagem natural, Guimarães Rosa, funções hash, análise léxico-morfológica.

Guimarães Rosa's Digital Lexicon: a Natural Language Processing Application for Teaching Computational Linguistics

Abstract: This paper presents a computational tool for the lexical-morphological analysis of Guimarães Rosa's works, with a focus on educational applications and linguistics instruction. Using a corpus of 63,868 words extracted from his writings, the tool integrates grammatical, morphological, and etymological data based on the *Léxico de Guimarães Rosa* and Natural Language Processing techniques. The system enables users to explore vocabulary by grammatical class, etymological origin, frequency, and contextual usage. To enhance search efficiency, eight hash functions were tested, with Zobrist hashing demonstrating optimal performance. The findings suggest that this tool can contribute to Portuguese language education by providing interactive resources for linguistic analysis of literary vocabulary.

Keywords: natural language processing, Guimarães Rosa, hash functions, lexical-morphological analysis.

1. Introdução

O avanço das tecnologias de Processamento de Linguagem Natural (PLN) tem possibilitado novas formas de interação com textos e línguas, especialmente no campo da Educação. O PLN é uma subárea da Inteligência Artificial (IA) que se dedica à interpretação e manipulação automática da linguagem humana por computadores (Barbosa, 2004).

Diante da popularização da IA, torna-se cada vez mais necessário que diferentes áreas do conhecimento avancem em estudos voltados à representação e recuperação da informação, fundamentais para o desenvolvimento de sistemas inteligentes e aplicáveis

a contextos educacionais e linguísticos (Izo e Oliveira, 2023).

Técnicas como análise morfológica, reconhecimento de entidades, lematização (redução de palavras flexionadas à forma canônica, como “cantavam” → “cantar”) e organização semântica viabilizam a exploração de grandes volumes de dados textuais de maneira eficiente (Manning, Raghavan e Schütze, 2009).

O campo das Humanidades Digitais tem se consolidado como uma área interdisciplinar que aplica métodos computacionais para análise e interpretação de textos literários (Burdick *et al.*, 2012), (Terras, 2012). A análise automatizada de corpus possibilita diferentes perspectivas na literatura, como demonstra Moretti (2005) em sua proposta de “*distant reading*” e Jockers (2013) em macroanálise textual. No âmbito do PLN, técnicas como tokenização, lematização e análise sintático-morfológica são fundamentais para estruturar léxicos digitais que suportem estudos linguísticos e literários (Manning e Schütze, 1999), (Jurafsky e Martin, 2025).

Em particular, a linguística computacional aplicada à literatura tem se beneficiado de abordagens híbridas que combinam extração automática e revisão manual, otimizando a qualidade dos dados (Cimiano e Völker, 2005), (Panchendrarajan e Zubiaga, 2024).

No contexto da literatura brasileira, o escritor João Guimarães Rosa se destaca pela complexidade e originalidade do uso da linguagem. Médico, diplomata e escritor mineiro, Rosa revolucionou a literatura nacional ao introduzir uma escrita inovadora, rica em regionalismos, arcaísmos, estrangeirismos e, sobretudo, neologismos, que são palavras inventadas ou modificadas que ampliam o potencial expressivo do idioma (Bosi, 1994), (Romero, 2022).

Obras como *Grande Sertão: Veredas* e os contos de *Sagarana* e *Primeiras Estórias* são exemplos emblemáticos de um vocabulário que desafia leitores e estudiosos, exigindo abordagens interpretativas que vão além das análises tradicionais (Martins, 2020).

Nesse contexto, ferramentas de PLN podem ampliar o acesso ao vocabulário literário de Guimarães Rosa, oferecer novas perspectivas interpretativas ainda pouco exploradas e contribuir para o ensino de linguística ao facilitar a análise sistematizada de obras complexas (Alhawiti, 2014). Além disso, podem servir como apoio valioso para estudantes que desejam compreender com mais profundidade as escolhas lexicais e estilísticas do autor (Romero, 2022).

Este artigo tem como objetivo apresentar uma ferramenta computacional que automatiza a análise léxico-morfológica das obras de Guimarães Rosa, permitindo a identificação, classificação e contextualização de seu vocabulário.

A partir de uma base textual contendo 63.868 palavras extraídas das obras do autor, o sistema foi enriquecido com informações de 8000 verbetes extraídos do *Léxico de Guimarães Rosa* de Martins (2020), além de dados etimológicos, morfológicos e exemplos de uso em trechos originais.

O trabalho também contempla a avaliação do desempenho de oito funções hash (algoritmos que convertem dados de entrada de tamanho variável em uma saída de tamanho fixo, conhecida como hash, para facilitar buscas rápidas), selecionadas com base em sua performance no estudo de Moreno (2017) e Moreno, Barbosa e Manfio (2021), aplicadas à estruturação e recuperação eficiente dos dados léxicos. A análise comparativa considera métricas como tempo de processamento, taxa de colisões e eficácia geral das funções.

A ferramenta visa oferecer um recurso educacional robusto para estudantes e

pesquisadores de Linguística, permitindo explorar o léxico rosiano em múltiplas dimensões: frequência de uso, distribuição por obra, classe gramatical, origem etimológica, formas variantes e contextos de ocorrência. Com isso, busca-se contribuir para o ensino de morfologia, etimologia e leitura crítica, ampliando o acesso e a compreensão da linguagem singular de Guimarães Rosa.

Este artigo está organizado da seguinte forma: a Seção 2 aborda os Materiais e Métodos; a Seção 3 traz os Resultados; a Seção 4 há a Discussão dos Resultados e a Seção 5 apresenta as Conclusões.

2. Material e Métodos

O processo de desenvolvimento desta ferramenta se dividiu em quatro etapas principais como mostra a Figura 1: (1) constituição do corpus textual, (2) processamento linguístico, (3) enriquecimento léxico-morfológico, e (4) implementação de funções hash para indexação e análise de desempenho.



Figura 1. Etapas do desenvolvimento da ferramenta. Fonte: os Autores.

2.1. Corpus textual

O corpus deste trabalho foi composto por obras de João Guimarães Rosa que constam no Léxico de Guimarães Rosa, de Martins (2020), o qual serviu como principal fonte de enriquecimento linguístico para a ferramenta desenvolvida. A escolha dessas obras visou garantir compatibilidade com os verbetes descritos no léxico, assegurando coerência entre os dados extraídos do corpus e as informações complementares associadas. A seguir são apresentadas as obras utilizadas, com suas respectivas siglas, edições e tamanhos (em número de palavras totais após o pré-processamento):

- S – Sagarana (Nova Fronteira, 2012) – 75556 palavras
- MM – Manuelzão e Miguilim (Nova Fronteira, 2001) – 53234 palavras
- UP – No Urubuquaquá, no Pinhém (Nova Fronteira, 2001) – 59520 palavras
- NS – Noites do Sertão (Nova Fronteira, 2013) – 67314 palavras
- GSV – Grande Sertão: Veredas (Companhia das Letras, 2019) – 138545 palavras
- PE – Primeiras Estórias (Nova Fronteira, 2001) – 36495 palavras
- T – Tutaméia (Nova Fronteira, 2009) – 39180 palavras
- EE – Estas Estórias (Nova Fronteira, 2013) – 63151 palavras
- AP – Ave, Palavra (Nova Fronteira, 2009) – 47512 palavras
- M – Magma (Nova Fronteira, 1997) – 7088 palavras

As obras foram convertidas para arquivos em formato .json, padronizadas e submetidas a um processo de limpeza e validação textual que incluiu a remoção de caracteres não textuais e a delimitação manual das seções relevantes, com o objetivo de excluir notas editoriais. Essa preparação visou garantir a integridade dos dados linguísticos a serem analisados nas etapas posteriores (Tabassum e Patil, 2020). Ao final do pré-processamento, o corpus totalizou 63.868 palavras distintas para análise.

2.2. Processamento linguístico

O pré-processamento textual foi realizado com a biblioteca spaCy¹ como da linguagem de programação Python, utilizando o modelo para português (pt_core_news_lg). As etapas incluíram: tokenização (separação das palavras), lematização (redução à forma base) e análise morfológica (atribuição de categoria gramatical, gênero, número, tempo verbal, entre outros).

Seguindo práticas similares às de Abonizio, Barbosa e Artoni (2019), que empregaram o spaCy na identificação automática de heterônimos de Fernando Pessoa, cada palavra foi enriquecida com atributos como classe gramatical, lema, frequência absoluta e relativa por obra, e conjunto de traços morfológicos conforme o padrão *Universal Dependencies*.

Essa abordagem visa garantir uma representação linguística detalhada e padronizada das palavras, servindo de base para o enriquecimento posterior com dados etimológicos e contextuais.

2.3. Enriquecimento léxico

As palavras identificadas foram enriquecidas com informações extraídas do Léxico de Guimarães Rosa de Martins (2020). Cada item do léxico foi estruturado como um dicionário contendo informações linguísticas e contextuais associadas a uma palavra específica. A estrutura inclui os seguintes campos:

- token (palavra): a forma textual original encontrada no corpus;
- frecuencia_total: número total de ocorrências da palavra nas obras analisadas;
- frecuencia_por_obra: distribuição das ocorrências por obra, utilizando as siglas previamente definidas;
- frases: lista de trechos extraídos diretamente das obras, contendo a palavra em seu contexto original, com indicação da sigla da obra de origem;
- formas_variantes: variantes morfológicas da palavra, caso identificadas (por exemplo: plural, feminino, formas conjugadas);
- significados: espaço destinado à inserção de definições semânticas, quando disponíveis;
- tokens_similares: lista de palavras semanticamente ou morfolologicamente próximas que são úteis para análise comparativa ou sugestões de estudo;
- classe_gramatical: a categoria gramatical da palavra (por exemplo, substantivo, adjetivo, verbo, etc.);
- lema: a forma base ou canônica da palavra;
- morfologia: conjunto de atributos morfológicos associados à palavra, extraídos via spaCy conforme o padrão *Universal Dependencies*. Pode incluir informações como:
 - Gênero (*Gender*): masculino, feminino, comum;
 - Número (*Number*): singular ou plural;
 - Tempo (*Tense*): presente, passado, futuro;
 - Modo (*Mood*): indicativo, subjuntivo, imperativo;
 - Pessoa (*Person*): 1^a, 2^a ou 3^a pessoa;
 - Forma verbal (*VerbForm*): infinitivo, particípio, gerúndio, etc.;

¹ <https://spacy.io>

- Grau (*Degree*): positivo, comparativo ou superlativo;
- Tipo de pronome (*PronType*): pessoal, demonstrativo, relativo, indefinido etc.;
- Entre outros, conforme a natureza da palavra analisada.

A estrutura permite buscas linguísticas/contextuais e serve de base para análises quantitativas e qualitativas, especialmente no que se refere à ocorrência e distribuição de neologismos, formas raras e construções gramaticais complexas.

A estrutura final foi organizada em um Léxico Digital, com campos específicos para cada dimensão linguística, seguindo os princípios de organização encontrados em bancos lexicais computacionais, como o WordNet (Fellbaum, 1998).

2.4. Funções de dispersão

O uso de funções de dispersão em tarefas linguísticas é recorrente em aplicações como compiladores, analisadores léxicos e estruturas de dados otimizadas para busca (Manning, Raghavan e Schütze, 2009), permitindo ganhos significativos de desempenho em sistemas baseados em grandes volumes textuais.

Visando a eficiência no armazenamento e na recuperação das entradas do léxico, foram implementadas 8 (oito) funções hash clássicas e especializadas, selecionadas entre as 14 avaliadas no estudo de Moreno (2017) e Moreno, Barbosa e Manfio (2021), que analisou comparativamente o desempenho dessas funções em diferentes contextos computacionais. A escolha das funções levou em consideração critérios como aplicabilidade textual, diversidade de estratégias de dispersão e potencial para otimização de buscas em estruturas lexicais.

A presente aplicação amplia esse escopo ao testar as funções em um léxico com características morfológicas e etimológicas complexas, voltado à análise linguística de textos literários. Com isso, busca-se não apenas desempenho computacional, mas também viabilidade prática em contextos educacionais e de investigação textual.

As funções utilizadas foram: Mix 32 bits, Mix 64 bits, One At a Time, PJW Hash, AHO Hash, Lookup3 64 bits, Lookup3 32 bits e Zobrist.

Essas funções foram aplicadas a todas as chaves léxicas do dicionário (palavras ou expressões) e testadas com base em métricas de desempenho, incluindo tempo total de processamento e tempo médio em milissegundos por operação, taxa de colisões e uso de memória.

2.5. Ambiente e ferramentas

Os experimentos foram conduzidos na plataforma Google Colab, um ambiente de computação em nuvem que fornece acesso a máquinas virtuais com recursos configuráveis. Durante os testes utilizou-se um ambiente com sistema operacional Linux (Ubuntu 22.04), processador virtual Intel Xeon e 13 GB de memória RAM.

As implementações foram realizadas em Python 3.11, com o uso das bibliotecas spaCy, pandas, numpy, matplotlib, seaborn e implementações próprias em Python para a aplicação e comparação das funções hash.

3. Resultados

Os resultados obtidos a partir da construção e análise do Léxico Digital de Guimarães Rosa evidenciam tanto a riqueza linguística da obra quanto a eficácia do processamento computacional aplicado. A seguir os dados são organizados em três

frentes principais: distribuição lexical e morfológica, análise de neologismos, e avaliação do desempenho das funções hash.

3.1. Exploração linguística e pedagógica

A estrutura do léxico e os dados linguísticos extraídos permitem o desenvolvimento de recursos educacionais digitais inovadores, como demonstram iniciativas que empregam léxicos especializados para apoiar estudantes em domínios específicos do conhecimento, integrando visualizações e consultas para fins didático-pedagógicos (Faria e Barbosa, 2021).

Entre os recursos possíveis, destacam-se:

- exploração dinâmica por classe gramatical, origem etimológica e frequência por obra, incentivando a descoberta de padrões linguísticos e a contextualização histórica e cultural do vocabulário;
- estudo comparativo de palavras com formas variantes, arcaicas ou semelhantes, promovendo reflexões sobre a formação e transformação lexical da língua;
- análises morfológicas sistematizadas que podem ser utilizadas em atividades de sala de aula voltadas ao ensino de morfologia, etimologia e variação linguística.

3.2. Análise linguística do léxico

A base léxica construída a partir das obras de Guimarães Rosa resultou em 63.868 tokens únicos, enriquecidos com dados morfológicos, etimológicos e gramaticais. A classificação gramatical dos termos é apresentada na Figura 2, destacando a predominância de verbos (24.079 ocorrências) e substantivos (21.617 ocorrências), seguidos de adjetivos (8.323) e advérbios (1.206).

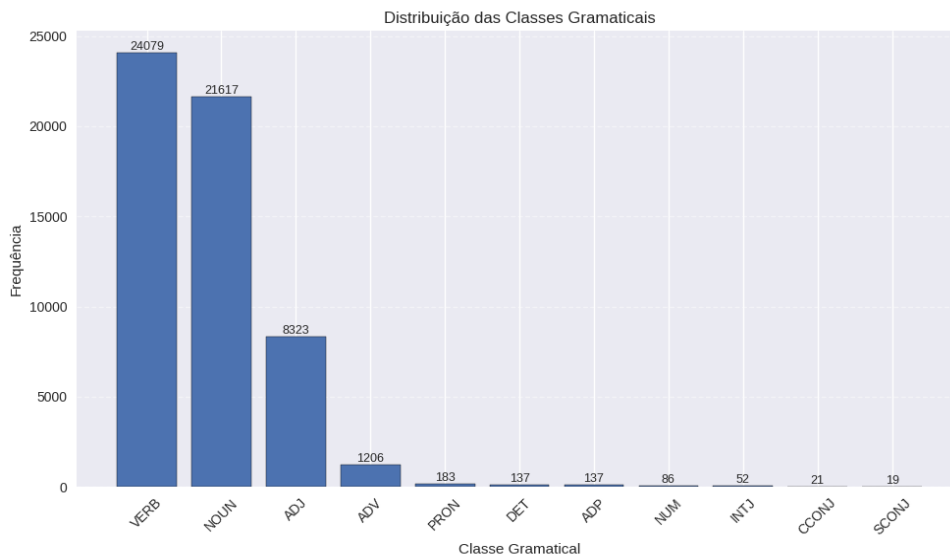


Figura 2. Frequência absoluta das classes gramaticais no corpus analisado. Abreviações: VERB (verbo), NOUN (substantivo), ADJ (adjetivo), ADV (advérbio), PRON (pronome), DET (determinante), ADP (preposição), NUM (numeral), INTJ (interjeição), CCONJ (conjunção coordenativa), SCONJ (conjunção subordinativa). Fonte: os Autores.

A análise etimológica, baseada no Léxico de Guimarães Rosa de Martins (2020), revela que a maioria das palavras com origem identificada remonta ao latim, seguida por termos de origem indígena, francesa e grega, conforme ilustrado na Figura 3. É importante destacar que foram consideradas apenas as palavras para as quais o referido léxico fornece a definição etimológica, desconsiderando-se a origem "portuguesa" como

categoria, por se tratar da língua base comum às demais.

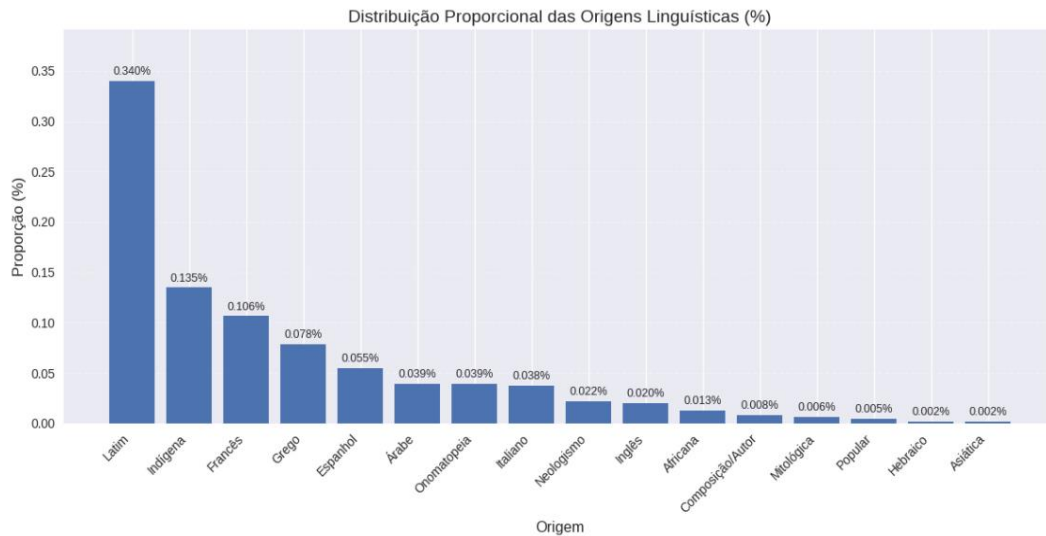


Figura 3. Distribuição proporcional das origens linguísticas, cujos valores indicam a proporção relativa (%) de palavras.
Fonte: os Autores

Além da estrutura léxica e dos dados morfológicos, foi desenvolvido um protótipo de interface interativa que possibilita a visualização dinâmica das palavras no contexto das obras analisadas (Figura 4).



Figura 4. Interface interativa de visualização do léxico: distribuição, variantes e contexto da palavra "Veredas".
Fonte: os Autores.

3.3. Avaliação das funções hash

Para a organização e indexação eficiente do léxico digital de Guimarães Rosa, foram aplicadas oito funções hash. A análise comparativa baseou-se em quatro critérios principais: tempo total de processamento, tempo médio por hash, taxa de colisões e espalhamento dos valores gerados.

Com relação ao tempo total de processamento do léxico, a função Zobrist apresentou o melhor desempenho, com tempo total de apenas 112,44 ms, em contraste com a função *One-at-a-Time* que registrou 961,58 ms (Figura 5).

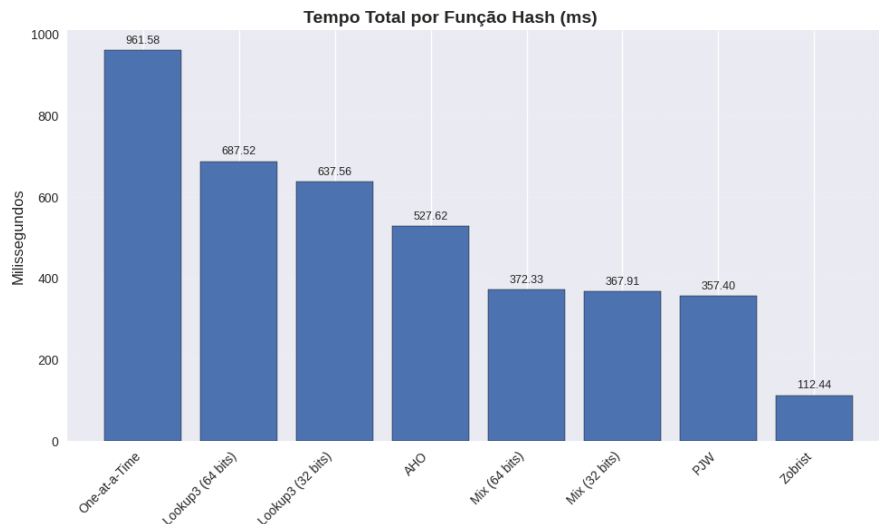


Figura 5. Tempo total de execução das funções hash com valores expressos em milissegundos (ms). Fonte: os Autores.

Já o tempo médio por operação, Zobrist também obteve o menor tempo médio por hash (0,0018 ms), conforme podemos ver na Figura 6, evidenciando seu desempenho na conversão de chaves lexicais em índices.

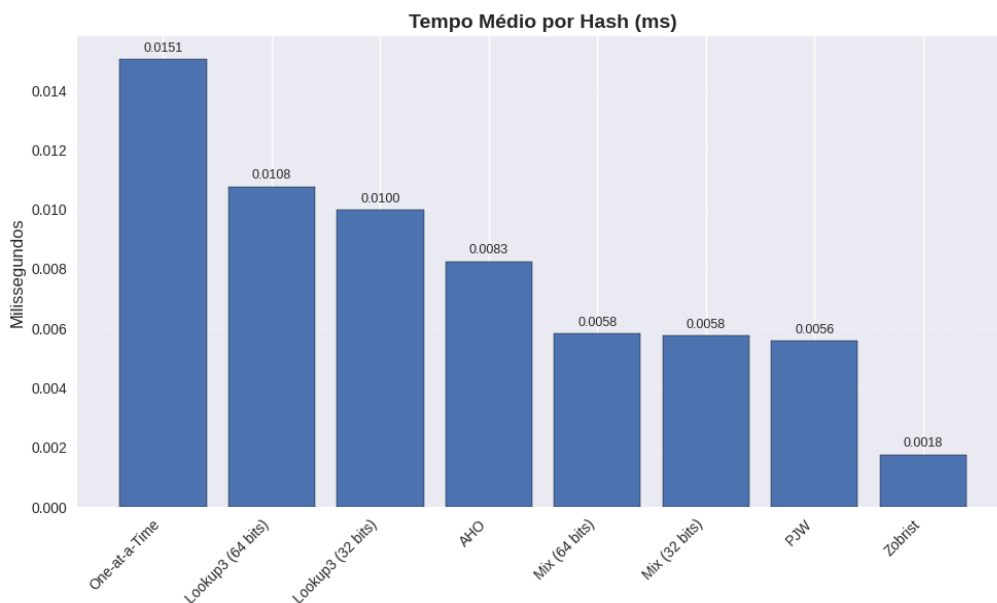


Figura 6. Tempo médio por função hash com valores expressos em milissegundos (ms). Fonte: os Autores.

E quanto à taxa de colisões, como mostra a Figura 7, as funções Mix (32 e 64 bits) apresentaram a maior quantidade de colisões absolutas (59.804), enquanto Zobrist teve 21.400 colisões. Lookup3 e *One-at-a-Time* não apresentaram colisões neste experimento.

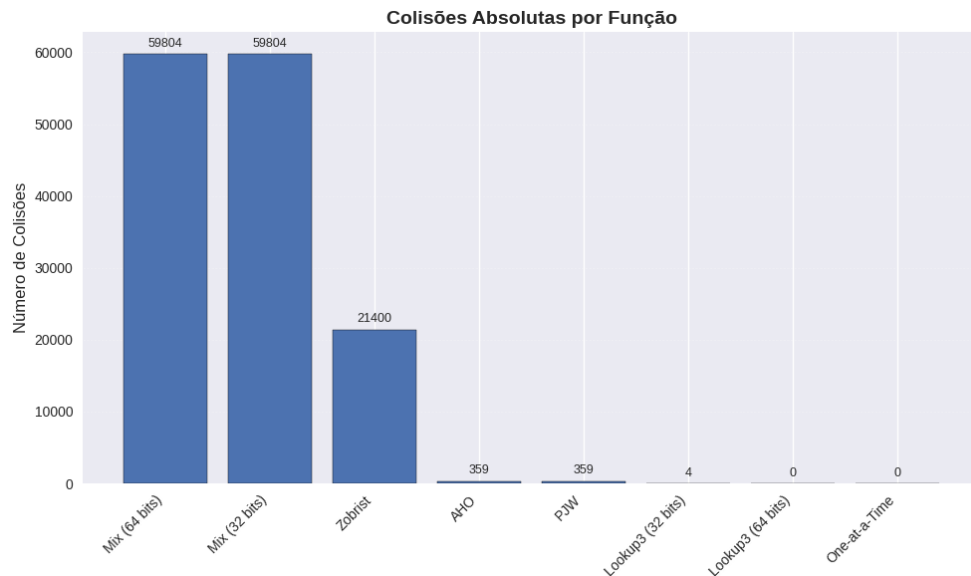


Figura 7. Colisões absolutas por função hash, as quais representam o número total de colisões detectada para cada função testada. Fonte: os Autores.

No quesito eficiência computacional e baixa taxa de colisões, a função Zobrist se destacou como a função mais eficiente entre as testadas, sendo especialmente adequada para aplicações linguísticas que exigem grande volume de dados e acesso rápido às informações. A Figura 8 apresenta a distribuição dos valores gerados pela função Zobrist, demonstrando uniformidade na frequência dos valores, o que corrobora sua alta qualidade como função de dispersão.

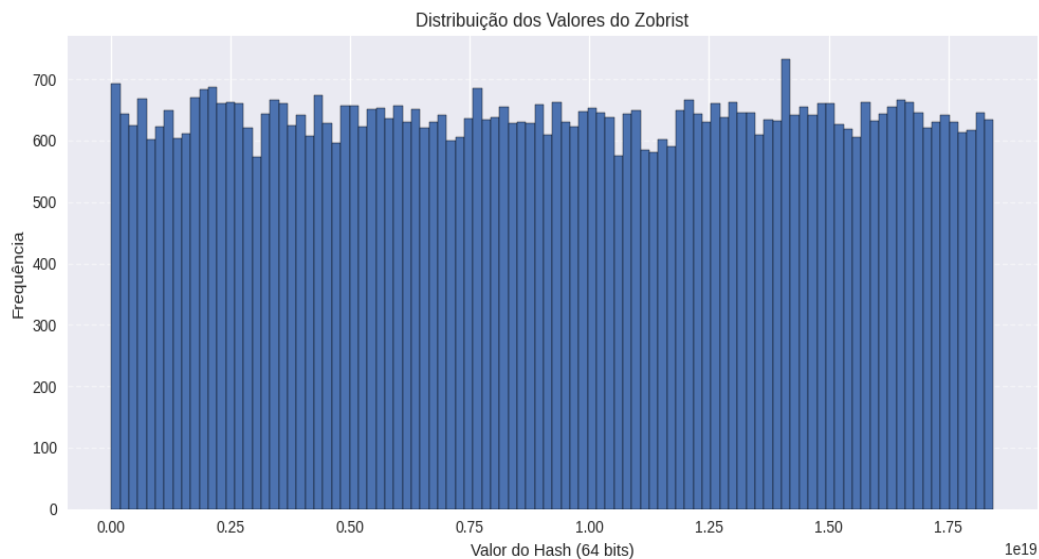


Figura 8. Distribuição dos valores de hash (64 bits) gerados pela função Zobrist. Fonte: os Autores.

4. Discussões dos Resultados

Esta seção apresenta os principais resultados da análise léxico-morfológica, destacando suas aplicações educacionais no ensino de linguística e literatura. A ferramenta desenvolvida é discutida como recurso pedagógico que integra tecnologia e humanidades, com foco na exploração da obra de Guimarães Rosa. Também são

analisadas limitações do processo, indicando possíveis caminhos para aprimoramento.

4.1. Aplicações educacionais

O uso de tecnologias de Processamento de Linguagem Natural (PLN) no ensino de Linguística tem potencial significativo para enriquecer práticas pedagógicas (Alhawiti, 2014). A ferramenta desenvolvida neste trabalho oferece um ambiente exploratório que permite investigar o vocabulário de Guimarães Rosa com base em critérios linguísticos variados, como classe gramatical, origem etimológica, frequência e variantes lexicais.

Entre os usos educacionais possíveis, destacam-se atividades como:

- exploração de neologismos. Estudantes podem identificar palavras inventadas pelo autor e analisar sua composição morfológica, discutindo seu impacto estilístico no contexto narrativo;
- estudo de etimologia aplicada. A ferramenta permite mapear palavras de origem tupi, árabe ou grega, promovendo reflexões sobre a formação histórica do português brasileiro e a intersecção entre léxico e identidade cultural;
- leitura crítica com apoio morfológico. Ao comparar a distribuição de substantivos, verbos ou adjetivos entre diferentes obras, alunos podem associar escolhas lexicais a características estruturais e temáticas da escrita.

Essas aplicações viabilizam um ensino mais investigativo, no qual o estudante assume papel ativo na construção do conhecimento linguístico, alinhando-se aos princípios das Humanidades Digitais (Burdick *et al.*, 2012) e à valorização da literatura como objeto de análise estruturada.

4.2. Limitações

Apesar do uso do Léxico de Guimarães Rosa (Martins, 2020) como principal fonte de enriquecimento deste trabalho, sua limitação com cerca de 8000 verbetes impediu a cobertura total do corpus analisado, que inclui um número significativamente maior de ocorrências lexicais extraídas das obras de Rosa, o que reduziu a uniformidade das análises. Além disso, algumas entradas do léxico carecem de informações completas. Essas ausências não comprometem a proposta, mas evidenciam a necessidade e a relevância de estratégias complementares de análise.

A delimitação textual, feita manualmente, pode ter gerado imprecisões pontuais na preparação dos dados. Tais fatores afetam a padronização geral do léxico, mas não interfere significativamente nos resultados principais.

5. Conclusões

Este trabalho demonstrou que é possível aplicar técnicas de Processamento de Linguagem Natural para estruturar e explorar o vocabulário de Guimarães Rosa de forma automatizada e educativa. A ferramenta desenvolvida permitiu a análise léxico-morfológica de um corpus com 63.868 palavras, enriquecido com informações gramaticais, morfológicas, etimológicas e contextuais extraídas do Léxico de Guimarães Rosa (Martins, 2020).

Os resultados confirmam o potencial da proposta para apoiar o Ensino de Linguística Computacional, oferecendo recursos para a investigação de padrões lexicais, variações morfológicas e origens etimológicas em um ambiente interativo.

Além disso, a aplicação e comparação de oito funções hash evidenciaram o bom desempenho da função *Zobrist* que se destacou pela eficiência computacional e

baixa taxa de colisões, reforçando sua adequação para a indexação de grandes volumes de dados linguísticos.

Confrontando os resultados com os objetivos iniciais, conclui-se que a ferramenta cumpre sua proposta de integrar literatura, linguística e computação em uma aplicação com valor pedagógico e potencial de expansão.

Como trabalhos futuros, propõe-se a ampliação do léxico, o enriquecimento dos verbetes usando modelos de linguagem de larga escala (LLMs – *Large Language Models*), bem como a disponibilização pública da ferramenta após mais validações em repositório aberto para fins acadêmicos e educacionais.

Referências Bibliográficas

- Abonizio, H. Q.; Barbosa, C. R. S. C. de; Artoni, A. A. Detecção Automática dos Heterônimos de Fernando Pessoa por Aprendizado de Máquina. In: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 22., 2019. **Anais...** Salvador: Sociedade Brasileira de Computação, 2019. p.144-151.
- Alhawiti, K. Natural Language Processing and its use in Education. **International Journal of Advanced Computer Science and Applications**, Bradford, Reino Unido, v.5, n.12, p.72-76, Dec., 2014.
- Barbosa, C. R. S. C. de. **Técnicas de Parsing para Gramática Livre de Contexto Lexicalizada da Língua Portuguesa**. São José dos Campos: Instituto Tecnológico da Aeronáutica, 2004. 171p. Tese (Doutorado em Engenharia Eletrônica e Computação).
- Bosi, A. **História concisa da Literatura Brasileira**. São Paulo: Cultrix, 1994. 528p.
- Burdick, A.; Drucker, J.; Lunenfeld, P.; Presner, T.; Schnapp, J. **Digital Humanities**. Cambridge, MA: MIT Press, 2012. 141p.
- Cimiano, P.; Völker, J. Text2Onto. In: Montoyo, A.; Muñoz, R.; Métails, E. (Ed.). Natural Language Processing and Information Systems. NLDB 2005. **Lecture Notes in Computer Science**, Berlin, v.3513, p. 227–238, Jun., 2005. Springer.
- Faria, C. R.; Barbosa, C. R. S. C. de. Ferramenta de Apoio aos Estudantes da Agricultura para Identificação de Invasores na Cultura da Soja. **Revista Novas Tecnologias na Educação**, Porto Alegre, v.18, n.2, p.161-170, jan., 2020. Disponível em: <<https://seer.ufrgs.br/renote/article/view/110214>>.
- Fellbaum, C. **WordNet: An electronic lexical database**. Cambridge: MIT Press, 1998. 423p.
- Izo, F.; Oliveira, E. Recuperação Automatizada de Documentos Científicos por meio da Análise de Planos de Ensino. **Revista Novas Tecnologias na Educação**, Porto Alegre, v.21, n.2, p.168-177, dez., 2023.
- Jockers, M. L. **Macroanalysis: digital methods and literary history**. 1ª. ed. Urbana: University of Illinois Press, 2013. 208p.
- Jurafsky, D.; Martin, J. H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models**. 2ª. ed. Upper Saddle River: Prentice Hall, 2008. 1024p.
- Manning, C. D.; Raghavan, P.; Schütze, H. **An Introduction to Information Retrieval**. Cambridge, England: Cambridge University Press, 2009. 544p.

Manning, C.; Schütze, H. **Foundations of statistical natural language processing**. Cambridge, MA: MIT Press, 1999. 680p.

Martins, N. S. **O léxico de Guimarães Rosa**. 3^a ed. São Paulo: Universidade de São Paulo, 2020. 536p.

Moreno, F. C. **Visual Tahs: Ferramenta para Analisar a Eficácia de Buscas das Funções Hash em um Léxico para Língua Natural**. Londrina: Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Londrina, 2017. 202p. Dissertação (Mestrado em Ciência da Computação).

Moreno, F. C.; Barbosa, C. R. S. C. de., Manfio, E. R. Tabelas Hash para um Léxico Digital. **Revista de Informática Teórica e Aplicada**, Porto Alegre, v.28, n.2, p. 25–38, ago. 2021.

Moretti, F. **Graphs, Maps, Trees: Abstract Models for a Literary History**. London; New York: Verso, 2005. 119p.

Panchendrarajan, R.; Zubiaga, A. Synergizing machine learning & symbolic methods: a survey on hybrid approaches to natural language processing. **Expert Systems with Applications**, Oxford, v.251, n.2, Oct., 2024.

Romero, S. Y. W. **Entre a risada e a alegria: estilística lexical e processos de transformação em Guimarães Rosa**. Londrina: Programa de Pós-Graduação em Estudos da Linguagem da Universidade Estadual de Londrina. 2022, 154p. Tese (Doutorado em Estudos da Linguagem).

Tabassum, A.; Patil, R. R. A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. **International Research Journal of Engineering and Technology**, Tiruchirappalli, v.7, n.6, p. 4864-4867, Jun., 2020.

Terras, M. The impact of social media on the dissemination of research: Results of an experiment. **Journal of digital humanities**, Fairfax, v.1, n.3, Sep., 2012. Disponível em: <<https://journalofdigitalhumanities.org/1-3/the-impact-of-social-media-on-the-dissemination-of-research-by-melissa-terras>>