Decifrando dados linguísticos: análise comparativa dos lematizadores para língua portuguesa

Mariana Gonçalves da Costa[1], **Sergio Serra**[1], **Jorge Zavaleta**[1]

[1]Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ – Brasil

marianag.costta@gmail.com, serra@pet-si.ufrrj.br, zavaleta@pet-si.ufrrj.br

Abstract. *This study comparatively analyzes the application of lemmatizers in the syntactic processing of Brazilian Portuguese texts. Lemmatization is the process of reducing inflected words to their base or dictionary form, eliminating inflections. In this paper, we analyze the application of three lemmatizers available for Portuguese: (i) the language model of the spaCy library; (ii) the method based on universal dependencies of the simplemma package; and (iii) the approach by lexicographic documents (Portilexicon-UD: POeTiSA Project). The results indicate that lemmatization methods that consider the linguistic context can improve the performance and the adequacy of the output, especially in models based on machine learning.*

Resumo. *Este estudo analisa comparativamente a aplicação de lematizadores no processamento sintático de textos em português brasileiro. A lematização refere-se ao processo de reduzir palavras flexionadas à sua forma base ou canônica, eliminando flexões. Neste trabalho, analisamos a aplicação de três lematizadores disponíveis para o português: (i) o modelo de linguagem da biblioteca spaCy; (ii) o método baseado em dependências universais do pacote simplemma; e (iii) a abordagem por documentos lexicográficos (Portilexicon-UD: Projeto POeTiSA). Os resultado indicam que métodos de lematização que consideram o contexto linguístico podem melhorar tanto o desempenho quanto a adequação da saída do lematizador, especialmente em modelos baseados em aprendizado de máquina.*

1. Introdução

Os repositórios de dados linguísticos desempenham um papel fundamental no avanço das pesquisas linguísticas baseadas no uso, sendo essenciais para a preservação e o compartilhamento de registros coletados por pesquisadores. Mais do que simples espaços digitais de armazenamento compartilhado, esses repositórios fortalecem a transparência e a reprodutibilidade da pesquisa linguística, alinhando-se aos princípios da Ciência Aberta (Fundação Oswaldo Cruz, 2020). No entanto, como amplamente discutido por pesquisadores das áreas de ciências humanas (Freitag et al., 2021; Machado Vieira; Barbosa, 2022), existem barreiras significativas no compartilhamento de dados de pesquisa. Uma das principais dificuldades é a ausência de padrões ou diretrizes claras para a coleta e o tratamento de dados. Isso frequentemente resulta em dados mal estruturados e em uma grande diversidade de técnicas empregadas na construção dos *corpora*, dificultando sua reutilização e interoperabilidade.

Dessa forma, é essencial fomentar uma cultura aprimorada de gerenciamento de dados que valorize o compartilhamento e a reutilização. Isso garante que os dados sejam acessíveis não apenas para a equipe que os coletou, mas também para uma comunidade mais ampla, composta por outros pesquisadores e interessados. Essa prática contribui para o avanço do conhecimento científico, ao possibilitar o uso sustentável e colaborativo dos recursos linguísticos.

Para assegurar a replicabilidade de uma pesquisa acadêmica, é fundamental oferecer um tratamento de dados transparente, que possibilite a reprodução da análise sem grandes variações nos resultados. No caso das pesquisas linguísticas, o processamento sintático dos dados deve ser devidamente documentado, garantindo sua reprodutibilidade. Nesse sentido, tratamentos de dados que envolvem trabalho manual excessivo também dificultam a replicabilidade de um estudo, o que abre espaço para colaborações entre linguistas e cientistas de dados. No âmbito desse estudo, busca-se explorar e implementar métodos eficazes para o tratamento de textos, facilitando a recuperação de dados linguísticos. Esse processamento considera as especificidades dos dados, que geralmente são não estruturados e organizados de maneiras heterogênea, o que demanda abordagens flexíveis e eficientes para a análise, categorização e organização desses recursos, assegurando sua usabilidade e valor científico.

Vale ressaltar que pesquisas linguísticas raramente se limitam ao estudo de itens isolados e estritamente definidos, como uma única palavra. Em vez disso, abrangem elementos mais amplos e complexos, como fonemas, morfemas, afixos, classes gramaticais, estruturas sintáticas, aspectos lexicais ou até conceitos mais abstratos, como a semântica e a análise do discurso. Ferramentas de busca simples baseadas em caracteres não atendem adequadamente a esses objetos de análise, levantando a questão de como estruturar e tratar os dados para garantir sua recuperabilidade pelo usuário. Nesse contexto, bancos de dados que não oferecem mecanismos eficazes de recuperação de informações textuais impõem um esforço adicional aos linguistas. Isso resulta em menos tempo disponível para a análise dos dados, prejudica a reprodutibilidade e replicabilidade dos estudos e aumenta a possibilidade de erros humanos durante as etapas de coleta e tratamento dos dados.

O presente trabalho foca exclusivamente no processamento de dados do português, com ênfase na lematização. Essa escolha decorre das limitações de técnicas como *stemming* – ou radicalização –, que reduz palavras a seus radicais, mas não identifica relações gramaticais em línguas morfologicamente ricas como o português, onde um verbo como *querer* pode apresentar mais de cinquenta formas diferentes. Em contraste, no inglês, o verbo *to want* possui apenas quatro formas. Devido a essa complexidade, muitas ferramentas de lematização não oferecem suporte adequado para essas línguas, como é o caso da biblioteca NLTK (*Natural Language Toolkit*) em Python.

O estudo avalia três técnicas de lematização – modelo de linguagem, dependências universais e arquivos lexicográficos (os dois primeiros envolvendo aprendizado de máquina) – tendo como parâmetros de avaliação a similaridade com a lematização manual, flexibilidade em diferentes aplicações e conformidade com a definição de lematização. Também abordamos medidas para mitigar perdas de informações linguísticas, que podem resultar em associações incorretas ou geração de termos inexistentes, impactando diretamente a análise textual. Essas questões são especialmente relevantes em contextos de pesquisa em que o texto é tratado como objeto de estudo e não apenas como recurso.

2. Referencial Teórico

A lematização é definida como uma tarefa do Processamento de Linguagem Natural (PLN) que reduz as formas flexionadas de determinadas palavras em suas formas de raiz morfologicamente corretas, o lema (Zhang; Mao; Cambria, 2023). Em outras palavras, a lematização trata do processo de remoção de afixos e desinências para reduzir o vocabulário do texto às formas sem flexão, ou seja, as encontradas no dicionário. Dessa maneira, uma palavra como *correndo* seria substituída por *correr*. Em contraste, no *stemming*, procura-se atingir algo relativo ao radical da palavra, logo, *correr* se tornaria *corr*. Além da possibilidade de reduzir palavras distintas a um mesmo stem, há ainda a preocupação com a remoção do significado do item reduzido, a impossibilidade de encontrar tal forma no conjunto de dados e a falta de amparo do contexto textual.

Estudos (Lagus; Klami, 2021) identificam que a lematização tende a ter melhor adequação em línguas morfologicamente ricas. As MRLs (*Morphologically Rich Languages*) seriam aquelas que incluem informações, como caso, gênero, número e modo, como parte dos vocábulos. Línguas analíticas, como o inglês, tendem a expressar essas marcas externamente à palavra, o que facilita o processo de redução dos itens gramaticas (não-lexicais, ou "vazios") que não contribuem à semântica da frase. (Lagus; Klami, 2021) ressaltam, ainda, que uma morfologia rica resulta em um vocabulário mais amplo e em uma menor frequência dos itens flexionados. Essa característica traz a essas línguas uma piora nos resultados em procedimentos que envolvem processamento de linguagem, como *word embedding*.

Em Ciência de Dados, há diferentes estudos em lematização que propoem a aplicação da técnica após a transformação da palavra em vetor. (Lagus; Klami, 2021) propoem a lematização após a etapa de vetorização das palavras, tratando a marcação flexional como um "viés" a ser tratado nos vetores. (Rosa; Zabokrtský, 2019) propoem a lematização por clusterização através de aprendizado não-supervisionado a fim de lidar com lacunas na representação de idiomas, mas sinalizam a dificuldade encontrada em vetorizar os dados sem considerar o contexto e indicam como possível solução a aplicação de *word embeddings* contextuais. Nessa linha, (Akhmetov et al., 2020) propoem um lematizador multilinguístico – que, na teoria, funcionaria independente do idioma – por classificação através de aprendizado supervisionado a partir do modelo de classificação de floresta aleatória (ou *Random Forest classification model*), apresentando resultados promissores.

Entretanto, apesar dos resultados interessantes dos estudos citados anteriormente, a aplicabilidade dessas técnicas possui a limitação de que o processo de *word embeddings* não permite a transdução contrária, ou seja, não é possível recuperar a palavra após seu tratamento como vetor. Seguindo os objetivos deste estudo, exploraremos apenas aplicações que resultem em um lema e que permita a comparação direta com a palavra original.

3. Descrição do dataset

O dataset analisado neste estudo é o Corpus D&G, desenvolvido pelo Grupo de Estudos Discurso e Gramática (D&G) da Universidade Federal Fluminense. Esse conjunto de dados reúne informações provenientes de cinco cidades brasileiras (Rio de Janeiro, Rio Grande, Niterói, Natal e Juiz de Fora), abrangendo cinco tipos de textos orais e versões escritas derivadas desses textos. Durante as entrevistas, os participantes foram solicitados a produzir os seguintes gêneros textuais: (1) narrativa de experiência pessoal; (2) narra-

Juiz de Fora	53.653
Natal	155.941
Niterói	38.028
Rio de Janeiro A	76.155
Rio de Janeiro B	91.351
Rio Grande	32.162
Base Conjugada	447.290

Tabela 1. Número de tokens resultantes do pré-processamento

Juiz de Fora	28.969
Natal	87.808
Niterói	21.156
Rio de Janeiro A	43.758
Rio de Janeiro B	49.838
Rio Grande	17.787
Base Conjugada	249.316

Tabela 2. Número de tokens após remoção de *stopwords*

tiva recontada; (3) descrição de local; (4) relato de procedimento; e (5) relato de opinião. A seleção dos falantes considerou diferentes níveis de escolaridade, organizados nas categorias: (a) alunos de classes de alfabetização infantil; (b) alunos da 4ª série do Ensino Fundamental; (c) alunos da 8ª série do Ensino Fundamental; (d) alunos da 3ª série do Ensino Médio; e (e) estudantes do último ano do Ensino Superior.

O *corpus* Discurso e Gramática foi escolhido devido à sua disponibilidade gratuita on-line, associada a uma documentação detalhada do processo de coleta, fator essencial para estudos linguísticos. Além disso, os dados transcritos apresentam características que diferem dos textos comumente utilizados por algoritmos de lematização, o que pode gerar resultados inovadores e revelar erros difíceis de identificar em textos originalmente escritos. Por exemplo, as transcrições incluem falas interrompidas, marcadores discursivos (como "aí", "i", "ah", "oh", "eh"), expressões com função discursiva (como “quer dizer”, “sabe?”, “ah tá”) e demais traços de oralidade. Apenas os caracteres utilizados para a transcrição do texto foram removidos.

4. Proveniência

No contexto de pesquisas científicas, a proveniência refere-se ao rastreamento da origem, história e transformações de dados, assim como dos resultados e métodos usados no estudo. Trata-se da documentação minuciosa do percurso de como um dado ou resultado foi obtido, incluindo as fontes utilizadas, os processos aplicados e as decisões tomadas durante a pesquisa. Isso garante transparência, reprodutibilidade e credibilidade aos resultados científicos.

Neste trabalho, o PROV-model, desenvolvido pelo *World Wide Web Consortium* (W3C), foi o padrão adotado para documentar e estruturar a proveniência. O modelo foi escolhido devido a sua padronização, rastreabilidade, reprodutibilidade e interoperabilidade, o que aumenta a confiabilidade da proveniência e a aproxima dos princípios FAIR que norteiam o movimento de Ciência Aberta.

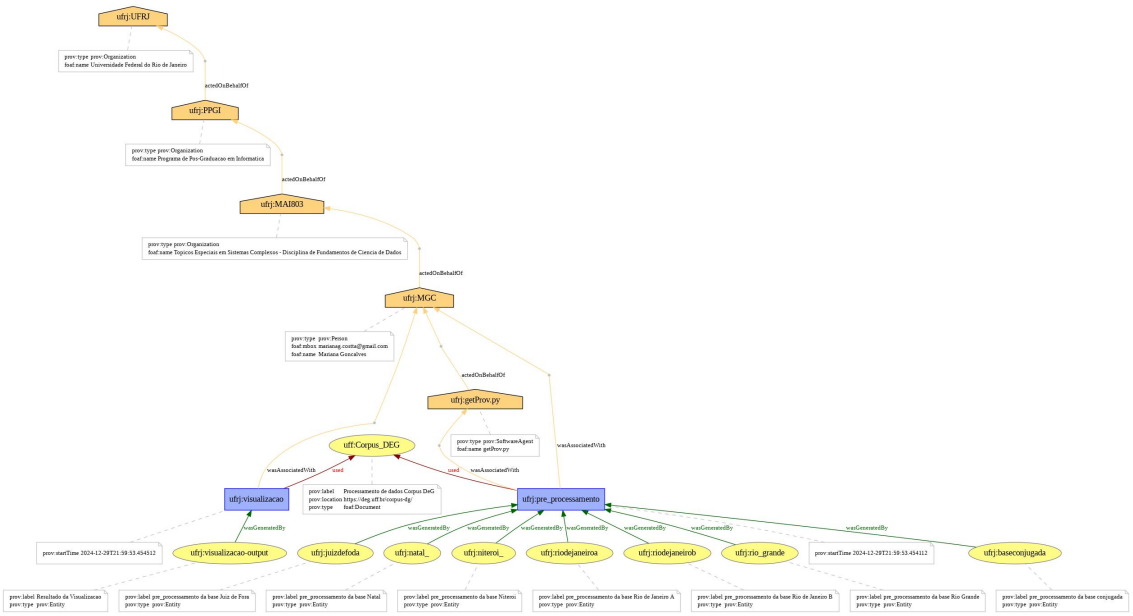

Figura 1. Pré-processamento do Corpus D&G

4.1. Grafos de Proveniência do Pré-processamento

A fim de facilitar a visualização da Figura 1, optou-se por remover da imagem as informações das entidades parte do Corpus D&G, ou seja, o *subcorpus* Juiz de Fora, Natal, Niterói, Rio de Janeiro A e Rio de Janeiro B – todos associados ao agente UFF (Universidade Federal Fluminense) responsável pela coleta e segmentação dos dados. A versão original da imagem pode ser encontrada no GitHub do trabalho.

No pré-processamento, foram realizadas as etapas de remoção de pontuação e caracteres de transcrição, normalização das palavras para minúsculas e tokenização. Com o objetivo de avaliar a influência da remoção de *stopwords* nas saídas do processo de lematização, optou-se por não incluir essa etapa no pré-processamento, diferentemente do que é comumente realizado em tarefas de PLN.

4.2. Grafos de Proveniência da Lematização

Devido à extensão dos dados disponibilizados pelo Grupo Discurso e Gramática, apenas a lematização do subset Rio Grande será analisada descritiva e quantitativamente. Como mencionado anteriormente, a remoção de *stopwords* foi realizada após o pré-processamento, com o objetivo de avaliar as interferências causadas pela eliminação de palavras indicativas de contexto gramatical. O lematizador da biblioteca spaCy, em particular, destaca-se por sua sensibilidade ao contexto linguístico dos itens lematizados. Portanto, é esperado que a remoção de palavras gramaticais influencie os resultados do lematizador.

5. Resultados e Discussão

Embora exista uma ampla variedade de materiais para tratamento de texto em Python, ainda é possível identificar lacunas significativas ao lidar com línguas morfologicamente ricas e pouco representadas, como o português. As ferramentas e recursos disponíveis

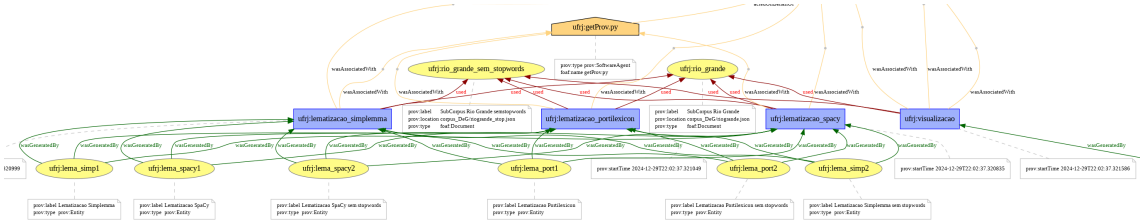

Figura 2. Lematização do subcorpus Rio Grande

frequentemente introduzem interferências que comprometem a integridade dos dados, especialmente quando etapas de mineração e processamento textual são concebidas como fases preliminares de análises mais abrangentes. Nesses casos, tais interferências poderiam ser minimizadas ou desconsideradas. Contudo, as fontes de erros frequentemente se perdem em meio às diferentes etapas de manipulação e generalização.

5.1. Remoção de *stopwords*

Na seleção das *stopwords* a serem removidas dos dados, foi realizada inicialmente uma busca por listas amplamente utilizadas em PLN, visando alinhar o processamento de dados às práticas da comunidade. Entretanto, o pacote de *stopwords* da biblioteca NLTK, um dos mais utilizados em PLN, apresenta limitações: contém 207 palavras consideradas *stopwords* em português, incluindo verbos de ligação, pronomes, preposições, verbos impessoais (como ter e haver), advérbios e adjetivos (só), além de termos inexistentes na língua portuguesa (tém, houverei, houverão, houveriam, houvéramos) ou pouco utilizados (fôramos, tivéramos, hajamos). Por outro lado, pronomes oblíquos (-lo, -la, -los, -las) e outros termos de alta frequência foram omitidos.

De maneira geral, as listas de *stopwords* fornecidas pelas bibliotecas existentes selecionam palavras de forma aparentemente arbitrária, ainda que incluam muitos itens de alta frequência e baixo valor semântico. Além disso, o formato de lista única do NLTK dificulta a manipulação de palavras conforme as necessidades específicas do processamento de dados. Com isso em mente, desenvolvemos um `dict` de *stopwords* voltado para o português, organizado por categorias gramaticais. Essa estrutura permite selecionar *stopwords* relevantes para a limpeza de dados, de acordo com os objetivos específicos do processamento textual.

As categorias definidas no dicionário incluem: artigos, pronomes, preposições, advérbios, conjunções, verbos (de ligação, como ser e estar, e de existência, como ter e haver), marcadores discursivos e números. Neste estudo, não removemos os verbos. Além disso, a função de seleção de *stopwords* oferece flexibilidade, permitindo a ativação de categorias específicas e a exclusão de palavras por meio do argumento `remove_list`, que aceita uma lista personalizada de termos a serem removidos. Verifique o material em: github.com/MarianaGCosta/Processamento-de-dados-em-portugues-brasileiro.

5.2. Lematização

5.2.1. Lematização spaCy

A função `token.lemma_` da biblioteca spaCy é amplamente utilizada no processo de lematização. Tem como base um modelo de linguagem que pode ser selecionado e baixado

(’que’, ’eu’)	132
(’a’, ’gente’)	116
(’que’, ’tu’)	109
(’e’, ’o’)	84
(’que’, ’é’)	84
(’e’, ’eu’)	81
(’relato’, ’de’)	75
(’e’, ’a’)	74
(’e’, ’e’)	74
(’a’, ’minha’)	66

Tabela 3. Ranqueamento de bigramas anterior à remoção de *stopwords*

(’alguma’, ’coisa’)	62
(’experiência’, ’pessoal’)	43
(’narrativa’, ’experiência’)	40
(’relato’, ’opinião’)	39
(’narrativa’, ’recontada’)	37
(’relato’, ’procedimento’)	36
(’descrição’, ’local’)	35
(’é’, ’é’)	27
(’11’, ’93’)	26
(’todo’, ’mundo’)	25)

Tabela 4. Ranqueamento de bigramas posterior à remoção de *stopwords*

diretamente do site da biblioteca. Neste estudo, utilizamos o modelo `pt_core_news_sm`, que é formado por dados extraídos da Wikipédia e pelo arquivo de dependências universais "UD Portuguese Bosque"na versão 2.8. A biblioteca spaCy afirma que a acurácia da lematização com esse modelo é de 0,97, mas oferece poucas informações sobre os detalhes de como a lematização é realizada. Pode-se presumir, com ressalvas, que o spaCy utilize um treinamento semi-supervisionado dos dados na criação do modelo, já que os modelos recebem dados anotados e não-anotados, mas não é possível identificar o grau de *feedback* disponibilizado durante treinamento.

Nos dados processados pela biblioteca spaCy, foram identificados três tipos principais de erros: (a) lematização insuficiente ou ausente (ex.: “liguei” → “liguei”, “alugá-la” → “alugá la”, “disser” → “disser”); (b) troca de classe gramatical (ex.: “lista” → “listar”, “narrativa” → “narrativo”); (c) surgimento de um item inexistente no léxico do português (ex.: “entrevistadora” → “entrevistadorar”). Cada tipo de erro apresenta uma gravidade e impacto diferentes nos resultados. Em alguns casos, a lematização produziu duas palavras distintas, como no exemplo de “no”, que se tornou “em o”, ou seja, a preposição “em” seguida do artigo definido masculino “o”. Além disso, a lematização ocorreu de maneira a mudar o número de tokens no dataset, o que dificulta a contabilização automática da acurácia.

É importante destacar que, em diversos algoritmos de PLN, como no spaCy, a lematização é considerada uma técnica de normalização que atua tanto na camada morfológica quanto na semântica do texto. Além de remover desinências e afixos para identificar uma forma comum, ela também agrupa palavras com significados equivalentes, como “superior” e “alto” ou “mínimo” e “pequeno”. Esse processo visa assegurar que documentos com significados semelhantes ao da busca sejam recuperados. Mais adiante, discutiremos os possíveis conflitos gerados por essa técnica.

A função realiza a substituição de palavras por seus sinônimos, como "superior"→ "alto"e "mínimo"→ "pequeno". Embora esse processo tenha como objetivo melhorar a recuperação de documentos com significados próximos ao da busca, a seleção das palavras a serem substituídas por sinônimos não é documentada e não pode ser desativada. No contexto da lematização neste estudo, essa abordagem pode prejudicar a qualidade dos resultados, uma vez que, do ponto de vista linguístico, a generalização que envolve o agrupamento de palavras sinônimas ocorre de forma arbitrária, o que resulta em uma perda semântica considerável.

Entrada: *narrativa recontada e tá joão e tem alguma história que alguém tenha te contado e que tu não tenha participado e que tu soubeste através de alguém i eh eu não sei se chega a ser o que tu está querendo mas isso aconteceu sexta feira também é mais pra mais um vale datado o cara chegou estava eh almoçando a mãe dele ligou pra ele assim está com um pé de cada sapato aí o o colega chegou e disse ué pé de cada sapato você acha que eu sou louco pôr um pé de sapato um pé de cada sapato olhou pra baixo um sapato com uma listinha e outro sem lista riso de e foi trabalhar com um pé de cada sapato pegamos no pé dele lá*

Saída: *narrativa recontar e tá João e ter algum história que alguém ter te contar e que tu não ter participar e que tu soubeste através de alguém i eh eu não saber se chegar a ser o que tu estar querer mas isso acontecer sexta feira também ser mais pra mais*

um vale datar o cara chegar estar eh almoçar o mãe de ele ligar pra ele assim estar com um pé de cada sapato aí o o colega chegar e dizer ué pé de cada sapato você achar que eu ser louco pôr um pé de sapato um pé de cada sapato olhar pra baixo um sapato com um listinha e outro sem lista riso de e ir trabalhar com um pé de cada sapato pegar em o pé de ele lá

5.2.2. Lematização simplemma

A função `simplemma` utiliza o arquivo de dependências universais do português "UD-PT-GSD" e afirma ter uma acurácia de 0,92. Seu objetivo é oferecer uma abordagem simples e multilíngue para lematização que, ao contrário de soluções mais complexas – como aquelas baseadas em modelos de linguagem –, seria fácil e rapidamente instalada e aplicada, sem a necessidade de grandes quantidades de informações já que seu treinamento é não-supervisionado. A documentação disponível no GitHub inclui uma definição clara da abordagem e da lematização, além de um tutorial completo de instalação, uso e otimização. Um ponto positivo é que a função não realiza a generalização das palavras por grupos de sinônimos.

A aplicação do `simplemma` superou as expectativas em termos de praticidade e acurácia. Embora tenha cometido alguns erros, como "coisa"→ "coisas", "parte"→ "partir"ou "quer"→ "quer", esses erros ocorreram em pequena escala, e a função se destacou como a de mais fácil aplicação. Os erros parecem estar relacionados à dificuldade de reconhecimento do contexto linguístico das palavras, com a função se baseando principalmente nas terminações para determinar se uma palavra deve ser lematizada ou não. Um teste mais aprofundado é necessário para verificar se a acurácia dos resultados diminui quando aplicada a um conjunto maior de dados.

Entrada: *narrativa recontada e tá joão e tem alguma história que alguém tenha te contado e que tu não tenha participado e que tu soubeste através de alguém i eh eu não sei se chega a ser o que tu está querendo mas isso aconteceu sexta feira também é mais pra mais um vale datado o cara chegou estava eh almoçando a mãe dele ligou pra ele assim está com um pé de cada sapato aí o o colega chegou e disse ué pé de cada sapato você acha que eu sou louco pôr um pé de sapato um pé de cada sapato olhou pra baixo um sapato com uma listinha e outro sem lista riso de e foi trabalhar com um pé de cada sapato pegamos no pé dele lá*

Saída: *narrativo recontado e tá joão e ter algum história que alguém ter te contar e que tu não ter participar e que tu saber através de alguém ir eh eu não saber se chegar o ser o que tu estar querer mas isso acontecer sexto feirar também ser mais pra mais um vale datar o caro chegar estar eh almoçar o mãe delir ligar pra ele assim estar com um pé de cada sapato aí o o colega chegar e dizer ué pé de cada sapato você achar que eu ser louco pôr um pé de sapato um pé de cada sapato olhar pra baixo um sapato com umar listinha e outro sem listar riso de e ir trabalhar com um pé de cada sapato pegar o pé delir lá*

5.2.3. Lematização por arquivos lexicográficos

Nesta abordagem, utilizou-se como base o arquivo lexicográfico "Portilexicon-UD", desenvolvido pelo Projeto POeTiSA. A função `lematizacao_portilexicon` recebe um texto tokenizado e um arquivo lexicográfico — sendo o Portilexicon-UD como padrão — e retorna uma lista de tokens lematizados. O processo de lematização ocorre por meio da busca do token na primeira coluna do arquivo e sua substituição pelo item correspondente na segunda coluna, caso encontrado. Se a palavra não for encontrada, ela é retornada sem alterações. O principal problema dessa abordagem está no fato de o arquivo estar organizado em ordem alfabética. Dessa forma, a palavra alvo será sempre a primeira a ser listada, independentemente da classe gramatical à qual pertence, como nos casos: "ensino" $\rightarrow$ "ensinar", "queira" $\rightarrow$ "queiro" ou "chega" $\rightarrow$ "chega".

Inicialmente, considerou-se utilizar a anotação sintática (*POS-tagging*) disponível no Portilexicon como peso na seleção do lema. No entanto, algoritmos de anotação sintática no português apresentam baixa acurácia e podem demandar grande tempo de limpeza e correção. Além disso, a natureza dos textos trabalhados dificultaria a anotação, pois as orações são frequentemente interrompidas ou repetidas. Vale considerar a dificuldade de adaptar o algoritmo a línguas diferentes, especialmente caso dependa da anotação sintática. Além da possibilidade de os textos utilizarem padrões diferentes de anotação, a acurácia dessa anotação pode variar substancialmente.

Assim, ao considerar possíveis otimizações do algoritmo, deve-se levar em conta a replicabilidade, a acurácia e a adequação a arquivos lexicográficos de outros idiomas. A operação mantém a organização alfabética do arquivo lexicográfico original já que esse é o formato padrão de dicionários, embora um ranqueamento por frequência possa permitir uma execução mais rápida. Vale ressaltar que a complexidade de pior caso do algoritmo ocorre quando a palavra a ser lematizada está no final do arquivo lexicográfico ou quando não é encontrada no arquivo — o segundo caso sendo mais frequente. Portanto, alterar a organização do arquivo não reduzirá a complexidade de pior caso e afetará diretamente a substituição do Portilexicon-UD por documentos de outros idiomas.

Entrada: *narrativa recontada e tá joão e tem alguma história que alguém tenha te contado e que tu não tenha participado e que tu soubeste através de alguém i eh eu não sei se chega a ser o que tu está querendo mas isso aconteceu sexta feira também é mais pra mais um vale datado o cara chegou estava eh almoçando a mãe dele ligou pra ele assim está com um pé de cada sapato aí o o colega chegou e disse ué pé de cada sapato você acha que eu sou louco pôr um pé de sapato um pé de cada sapato olhou pra baixo um sapato com uma listinha e outro sem lista riso de e foi trabalhar com um pé de cada sapato pegamos no pé dele lá*

Saída: *narrativo recontar e estar joão e ter algum história que alguém ter te contar e que tu não ter participar e que tu saber através de alguém i eh eu não saber se chegar o ser o que tu estar querer mas isso acontecer sexto feirar também ser mais para mais um valer datar o caro chegar estar eh almoçar o mãe delir ligar para ele assim estar com um pé de cada sapato aí o o colega chegar e dizer ué pé de cada sapato você achar que eu ser louco pôr um pé de sapato um pé de cada sapato olhar para baixo um sapato com um listinho e outro sem listar riso de e ser trabalhar com um pé de cada sapato pegar*

6. Validação dos resultados

Na etapa de validação, foi selecionada uma amostra dos dados do Rio Grande (1.000 tokens) e realizada uma lematização manual seguindo a definição proposta por Zhang et al. (2023). Essa lematização manual foi utilizada como referência para comparar o output esperado dos algoritmos. A acurácia de cada resultado foi calculada por meio de uma simples divisão entre o número de acertos e o tamanho da amostra.

A acurácia A se refere à lematização da base Rio Grande, a acurácia B se refere à lematização realizado nos dados sem *stopwords* e o tempo de execução se refere à lematização da base conjugada, conforme apresentado na tabela abaixo:

Lematizador	Acurácia A	Acurácia B	Tempo de execução
spaCy	0.869	0.866	1m
simplemma	0.823	0.862	0s
Portilexicon	0.8	0.806	1s

Pode-se perceber um aumento pequeno na acurácia dos lematizadores Simplemma e Portilexicon e uma baixa redução na acurácia do lematizador da biblioteca spaCy. Ao considerar que a lematização do spaCy realiza a duplicação de determinados itens ("do"para "de o", etc), como citado anteriormente, poderia se supor um aumento na acurácia quando tais itens fossem removidos, porém a falta de indicadores contextuais prejudica diretamente o processo de lematização contextual.

7. Considerações Finais

Os testes implementados indicam que a lematização por modelo de linguagem do spaCy apresentou o melhor resultado, apesar de ainda inferior à acurácia indicada em sua documentação. Essa discrepância pode ser atribuída à natureza dos dados utilizados, uma vez que a semelhança entre os dados de treinamento e os dados de teste influencia diretamente a performance do algoritmo. Ademais, é essencial garantir uma definição clara da etapa de lematização uma vez que a aglomeração de palavras por semelhança de sentido (aquelas consideradas sinônimas) pode ser útil no tratamento de *corpora* volumosos, mas deve ser evitada em analises sensíveis às diferenças semânticas. Seria possível, por exemplo, acrescentar um parâmetro à função de lematizador do spaCy que ative ou desative a substituição dos sinônimos.

Apesar dos esforços contínuos de especialistas em linguística e processamento de linguagem natural (PLN), a língua portuguesa ainda é sub-representada em algoritmos de PLN. As poucas bibliotecas que oferecem lematização automática para o português frequentemente exibem acurácias inferiores às mencionadas em sua documentação oficial. Ainda assim, a acurácia dos lematizadores foi surpreendentemente alta considerando o contexto de teste, embora ainda inferior à descrita em sua documentação. Considerando a simplicidade da abordagem e os poucos recursos necessários, verifica-se que a qualidade dos dados de treinamento é mais relevante do que a quantidade. Métodos mais inteligentes e contextualmente informados de lematização podem contribuir para melhorar a performance e a adequação do processo em diferentes contextos linguísticos. Modelos

que levam em conta o contexto sintático e se baseiam em dependências universais catalogadas por pesquisadores apresentam grande potencial para aprimorar a lematização no português e em outras línguas de alta complexidade morfológica.

A partir das conclusões alcançadas neste estudo, os trabalhos futuros – já em desenvolvimento – visam avançar no treinamento de modelos de linguagem que representem de forma mais precisa e abrangente o repertório linguístico da língua portuguesa. Esses modelos serão projetados não apenas para permitir ativar ou não a aglomeração de sinônimos, mas também para disponibilizar a escolha do modelo mais adequado às características específicas dos dados analisados. Por exemplo, dados que se aproximam da oralidade poderão ser processados por meio de modelos treinados com *corpora* de transcrições, garantindo maior aderência e precisão nos resultados.

Além disso, será dada atenção especial às aplicações que envolvem a vetorização de palavras, explorando de maneira mais detalhada como as representações vetoriais podem contribuir para análises linguísticas e tarefas computacionais mais robustas. O desenvolvimento dessas aplicações inclui a investigação de métodos que maximizem a eficiência e a adaptabilidade dos modelos em diferentes contextos, como análise de sentimentos, processamento de linguagem natural e compreensão textual. Assim, os esforços futuros buscarão consolidar uma base tecnológica que promova avanços significativos no processamento automático da língua portuguesa, beneficiando tanto a pesquisa acadêmica quanto as aplicações práticas.

Referências

AKHMETOV, Iskander et al. Highly Language-Independent Word Lemmatization Using a Machine-Learning Classifier. **Computación y Sistemas**, v. 24, n. 3, p. 1353–1364, 2020.

FREITAG, R. M. K. et al. Challenges of Linguistic Data Management and Open Science. **Cadernos de Linguística**, v. 2, n. 1, 2021. Accessed: July 2024. Disponível em: https://cadernos.abralin.org/index.php/cadernos/article/view/307.

FUNDAÇÃO OSWALDO CRUZ, Fundação. **Política de gestão, compartilhamento e abertura de dados para pesquisa: princípios e diretrizes**. Rio de Janeiro, 2020. P. 19.

LAGUS, Jarkko; KLAMI, Arto. Learning to Lemmatize in the Word Representation Space. In_____. **Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)**. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, mai. 2021. P. 249–258. Disponível em: https://aclanthology.org/2021.nodalida-main.25/.

MACHADO VIEIRA, Marcia dos Santos; BARBOSA, Juliana Bertucci. Brazilian Datasets for Teaching Portuguese. In: VARIAÇÃO e Ensino de Português no Mundo: Variation et Enseignement de Portugais Dans le Monde. São Paulo: Blucher, 2022. P. 347–356.

ROSA, Rudolf; ZABOKRTSKÝ, Zdenek. Unsupervised Lemmatization as Embeddings-Based Word Clustering. **CoRR**, abs/1908.08528, 2019. arXiv: 1908.08528. Disponível em: http://arxiv.org/abs/1908.08528.

ZHANG, X.; MAO, R.; CAMBRIA, E. A survey on syntactic processing techniques. **Artificial Intelligence Review**, Springer, v. 56, p. 5645–5728, 2023. DOI: 10.1007/s10462-022-10300-7.