DOI: 10.5433/1981-8920.2025v30n3p521

PREPARAÇÃO AUTOMATIZADA DE CORPUS TEXTUAL PARA PESQUISAS QUALITATIVAS COM IRAMUTEQ

AUTOMATED PREPARATION OF TEXTUAL CORPUS FOR QUALITATIVE RESEARCH WITH IRAMUTEQ

Roberta de Oliveira Barbosa[a]
Júlio de Oliveira Júnior[b]
Luciano Cássio Lugli[c]
Deise Aparecida Peralta[d]

RESUMO
Objetivo: Apresentar um método automatizado para a preparação de corpus textual em pesquisas qualitativas utilizando o IRaMuTeQ, reduzindo o tempo e esforço na limpeza e formatação de textos. **Metodologia**: Foi desenvolvido um script em Python, executado no Google Colab, capaz de extrair o texto de arquivos PDF por meio da biblioteca pdfminer.six, remover caracteres incompatíveis com o IRaMuTeQ, substituir elementos de formatação proibidos (como aspas, hífens e símbolos especiais), unificar quebras de linha e padronizar a estrutura do corpus. Cada documento foi processado individualmente, gerando arquivos de texto no formato .txt já estruturados com linhas de comando específicas (subcorpus), conforme as normas do software. Após essa etapa, foi realizada uma revisão manual para corrigir fragmentações de palavras e eliminar elementos residuais, como numeração de páginas. **Resultados**: A abordagem automatizada eliminou inconsistências na formatação dos textos e reduziu significativamente o tempo necessário para a preparação do corpus, tornando a análise qualitativa mais eficiente e precisa. A metodologia permitiu maior padronização e replicabilidade no processamento dos dados textuais. **Conclusões**: A automação proposta facilita a adoção do IRaMuTeQ em pesquisas qualitativas de grande escala, eliminando barreiras técnicas e permitindo análises mais detalhadas. A integração de ferramentas computacionais à análise textual otimiza a organização dos dados, melhorando a qualidade e confiabilidade dos resultados.

a Doutora em Educação para a Ciência pela Universidade Estadual Paulista (Unesp). São Paulo, Brasil. E-mail: oliveira.barbosa@unesp.br.

b Doutorando no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) no Instituto de Pesquisas Energéticas e Nucleares (IPEN) pela Universidade de São Paulo (USP). São Paulo, Brasil. E-mail: julio.oliveira.oliveira@usp.br.

c Doutor em Engenharia Mecânica pela Universidade de São Paulo (USP). São Paulo, Brasil. E-mail: lc.lugli@unesp.br.

d Doutora em Educação Para a Ciência pela Universidade Estadual Paulista (Unesp). Docente Programa de Pós-Graduação em Educação para a Ciência na Universidade Estadual Paulista (Unesp). São Paulo, Brasil. E-mail: deise.peralta@unesp.br.

Descritores: Análise documental. Revisão bibliográfica. Google Colab. Python.

1 INTRODUÇÃO

A análise de dados qualitativos desempenha um papel fundamental na pesquisa científica, permitindo uma compreensão aprofundada de fenômenos complexos. Com o avanço da tecnologia, a utilização de *softwares* especializados tem se tornado cada vez mais comum, facilitando o processamento e a interpretação desses dados. A análise de dados qualitativos tem se beneficiado do avanço das ferramentas computacionais, permitindo não apenas a organização e processamento de grandes volumes de informações, mas também a extração de padrões e *insights* relevantes. No contexto da Informática na Educação, Baker, Isotani e Carvalho (2011) destacam a importância da mineração de dados educacionais como uma estratégia para identificar tendências e apoiar a tomada de decisões baseadas em evidências. Embora seu foco esteja voltado para dados quantitativos, muitas das técnicas discutidas podem ser complementares à análise qualitativa automatizada, como a realizada pelo IRaMuTeQ, ampliando as possibilidades de investigação e interpretação dos dados textuais em pesquisas educacionais.

Nascimento, Santos e Saraiva (2022) em sua revisão de literatura demonstram que o uso de softwares para análise de dados qualitativos tem ganhado espaço nas pesquisas qualitativas, e dão exemplos de alguns softwares como o MaxQDA, WebQDA, NVivo e Atlas Ti, os quais apresentam a desvantagem comum de serem softwares proprietários, ou seja, softwares não-livres, ou pagos, que restringe o uso, modificação e redistribuição.

Existem ainda softwares livres que auxiliam a análise de dados qualitativos, que podem ser utilizados, copiados, estudados, modificados e redistribuídos de forma gratuita sem restrições, dentre os quais Schlosser, Frasson e Cantorani (2019) evidenciam algumas opções, tais como o AQUAD 7, Cassandre, Digital Replay System, IRaMuTeQ, KH Coder, KNIME, Transcriber AG, e o Textométrie. Entretanto, dentre esses softwares, os únicos que tem bibliotecas e dicionários que possibilitam seu uso em língua portuguesa são o AQUAD 7, IRaMuTeQ, KH Coder e KNIME.

Dentre esses *softwares*, destaca-se que o IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*) é amplamente utilizado enquanto ferramenta de análise em pesquisas qualitativas, principalmente no Brasil, país que mais pesquisa o termo IRaMuTeQ no mundo no buscador Google, o que sugere uma ampla utilização conforme apontado por Mendes, Proença e Pereira (2022). O *software* de análises léxicas e gramaticais desenvolvido por Pierre Ratinaud em 2008, faz análises estatísticas de textos e funciona como uma interface do R, tendo o diferencial de analisar não apenas as palavras, mas o contexto no qual as palavras aparecem no texto.

No Brasil, a utilização do software começou a se popularizar em 2013, quando os pesquisadores do Laboratório de Psicologia Social da Comunicação e Cognição da Universidade Federal de Santa Catarina (LACCOS/UFSC) em parceria com o Centro Internacional de Estudos em Representações Sociais e Subjetividade - Educação, da Fundação Carlos Chagas (CIERS-ed/FCC); e com o grupo de pesquisa Valores, Educação e Formação de Professores da Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) desenvolveram um dicionário para o português (Camargo; Justo, 2013).

Apesar da ampla utilização do *software* IRaMuTeQ nas pesquisas qualitativas no Brasil, o *software* apresenta a desvantagem de precisar ser preparado manualmente para ser processado, uma etapa que pode demandar muito tempo de quem realiza as pesquisas. Este artigo propõe um método automatizado para a preparação de corpus textual, integrando o uso de Python e Google Colab. Essa abordagem visa proporcionar uma alternativa eficiente ao processo manual, facilitando a formatação e organização dos dados textuais para análises no IRaMuTeQ.

2 OS DESAFIOS DE USO DO IRAMUTEQ

Apesar da ampla utilização do *software* no Brasil, um dos limitadores do uso do software é a necessidade de preparação do corpus textual, ou a preparação do texto de acordo com normas que garantem que o arquivo seja lido corretamente, disponíveis no próprio site do Iramuteq. A chamada “limpeza do texto”, pode ser um desafio para utilizar essa ferramenta na análise de

documentos extensos, como no caso de documentos ou de íntegras de artigos e outros textos acadêmicos, visto que além de o software não realizar a leitura de documentos em formato PDF ou .DOC, apenas TXT, algumas especificidades precisam ser respeitadas, tais como não utilizar caracteres especiais como aspas ("), apóstrofo ('), hífen (-), cifrão ($), percentagem (%) ou asterisco (*), que é utilizado apenas na identificação dos subcorupus, além disso, o texto não pode utilizar recursos como itálico, negrito ou sublinhado, fatores comuns em documentos de todas as naturezas.

Essas e outras regras de formatação, para serem realizadas de forma manual, demandam muito tempo e diversas leituras da íntegra do material, o que torna pouco viável a análise de documentos extensos com o uso do software . Talvez, esse seja um dos motivos da maior parte das pesquisas realizadas utilizando essa ferramenta, serem relativas a análises de entrevistas, resumos, comentários de redes sociais e outros conjuntos de dados relativamente curtos, apesar do software não ter uma limitação de tamanho para os documentos analisados.

Como exemplo dessas pesquisas podemos mencionar a pesquisa de Ramos, Lima e Rosa (2018), que evidencia as potencialidades do *software* para análise textual discursiva, e realiza uma pesquisa utilizando respostas de um questionário aberto, a mesma ferramenta de coleta de dados utilizada no estudo Almeida *et al.* (2023) específica num movimento similar, Tinti, Barbosa e Lopes (2021) fazem a análise de um conjunto de narrativas.

Esses exemplos de pesquisa utilizam respostas de questionários ou entrevistas, ou seja, trechos relativamente curtos que torna menos exaustivo o processo de preparação do corpus textual pelo pesquisador. No campo da enfermagem por exemplo, o software é amplamente utilizado na análise textual, e como explicam Acauan *et al.* (2020), o tipo de *corpus* mais utilizado é o de respostas a entrevistas em pesquisas originais, apesar de também serem realizadas pesquisas de revisão integrativa, análise documental e estudos observacionais. A pesquisa de Souza (2018) utiliza como fonte de dados entrevistas semiestruturadas, e é um exemplo dessa aplicação no campo da enfermagem.

Porém, o *software* também é utilizado em diversas áreas para revisões bibliográficas, um tipo de pesquisa que por natureza lida com uma quantidade mais robusta de diferentes arquivos e páginas de conteúdo, o que poderia ser um desafio de preparação de dados. O estudo de Silva e Souza (2020) é um exemplo no qual ao realizar uma pesquisa bibliográfica com o uso do IRaMuTeQ, as pessoas autoras selecionaram apenas alguns elementos dos textos selecionados, nesse caso, as palavras-chave e os resumos. Silva e Riberio (2021) também utilizam o resumo de textos para elaborar o *corpus* textual, e descrevem detalhadamente o processo manual de limpeza do texto para não interferir no processamento dos dados, assim como Melo, Vasconcelos e Lima (2023), que também organizou o corpus textual a partir do resumo das publicações elencadas para seu estudo. Exemplos similares são os casos das pesquisas de Souza, Carvalho e Ramos (2020) que analisam resumos e títulos de teses e dissertações, e de Motta e Fiúza (2022) que utilizaram resumos e considerações finais de artigos. Silva e Dell'Agio (2023) também fez a análise de resumos de artigos com auxílio do IRaMuTeQ, assim como Santos *et al.* (2022), Bueno (2018), e Hoffmann, Alvarez e Martí-Lahera (2020).

3 UMA ALTERNATIVA PROCESSUAL PARA ELABORAÇÃO DO CORPUS TEXTUAL

Essa problemática da demanda de tempo para a elaboração do corpus textual, que pode ser um limitante que leva os pesquisadores a selecionarem apenas trechos dos textos ao realizar análises bibliográficas ou documentais, foi superada com o auxílio da linguagem de programação Pyhton pelos autores. Foi gerado um código de programação base, que ao ser rodado na plataforma Google Colabs realiza a leitura e limpeza do texto, automatizando esse processo e transformando o arquivo original em um arquivo limpo e no formato exigido pelo IRaMuTeQ, que após breve revisão de eventuais erros está pronto para uso. Essa alternativa foi utilizada no desenvolvimento de parte da tese de doutorado de uma das autoras do presente texto. Após realizar a limpeza do texto com o uso do código em Python, é necessário a realização de uma revisão dos conteúdos gerados para garantir que seguiam as normas de construção de

um corpus textual, bem como para criar linhas de comando que diferenciam cada seção do documento, facilitando a leitura das análises e a busca dos trechos nos documentos oficiais.

Embora este estudo não tenha realizado uma mensuração cronográfica sistemática comparando o tempo do processo manual e do automatizado, observou-se, durante os testes com diferentes conjuntos de dados, que a metodologia proposta reduziu de forma significativa o tempo de preparação do corpus e a ocorrência de erros de formatação. A automação das etapas de extração, limpeza e padronização do texto eliminou a necessidade de múltiplas leituras e ajustes manuais, aumentando a padronização do corpus e diminuindo o esforço cognitivo do pesquisador. Além disso, o procedimento mostrou-se acessível a usuários com pouca familiaridade com programação, favorecendo a autonomia e a replicabilidade. Estudos futuros estão sendo planejados para quantificar de maneira precisa esses benefícios, incluindo métricas de tempo economizado, taxa de inconsistências eliminadas e usabilidade percebida.

3.1 UTILIZANDO A FERRAMENTA GOOGLE COLABS

3.1.1 Preparando o Ambiente

a) Acesse o Google Colab:

Vá para Google Colab e crie um novo notebook.

b) Instale a biblioteca pdfminer.six (necessária para extrair texto de PDFs):

Execute o seguinte comando em uma célula do Colab:

```
!pip install pdfminer.six
```

c) Carregue o arquivo PDF: Utilize o código abaixo para carregar o arquivo PDF manualmente:

```
from google.colab import files
uploaded = files.upload()
```

3.1.2 Código Python Completo

Copie e cole o código fornecido em uma célula do Colab, substituindo o título do PDF:

```
import re
from pdfminer.high_level import extract_text
from google.colab import files

def limpar_texto(texto):

    # Remove caracteres especiais que o IRAMUTEQ não lida bem
    texto = re.sub(r'[“”"\'‘’]', '', texto)  # Aspas
    texto = re.sub(r'[-—–]', ' ', texto)  # Substitui hífens por espaço
    texto = re.sub(r'(\w+)-(\w+)', r'\1_\2', texto)  # Une palavras hifenizadas com underline
    texto = re.sub(r'\b(\w+)-me\b', r'me \1', texto)  # Converte pronomes para próclise

    # Remove todas as quebras de linha e substitui por espaço simples
    texto = re.sub(r'\n+', ' ', texto)

    # Adiciona a linha de comando "**** *texto1" apenas no início do texto seguida de quebra de linha
    texto = texto.strip()
    texto = '**** *texto1\n' + texto  # Coloca a linha de comando seguida do texto corrido

    return texto

def extrair_texto_pdf(arquivo_pdf):
    # Extrai o texto do PDF
    texto = extract_text(arquivo_pdf)

    # Limpa o texto extraído
    texto_limpo = limpar_texto(texto)

    return texto_limpo

# Nome do arquivo PDF carregado manualmente no Google Colab
filename = 'curriculo-paulista-fundamental.pdf'[1]

# Extrai e limpa o texto
texto_limpo = extrair_texto_pdf(filename)
```

[1] No código fornecido, o nome do arquivo PDF é especificado como: 'curriculo-paulista-fundamental.pdf'. **Este título é apenas um exemplo utilizado para demonstrar o processo**. Ao utilizar este código, o pesquisador deve substituir o nome do arquivo PDF pelo **nome correspondente ao seu próprio documento** entre aspas simples, como no exemplo.

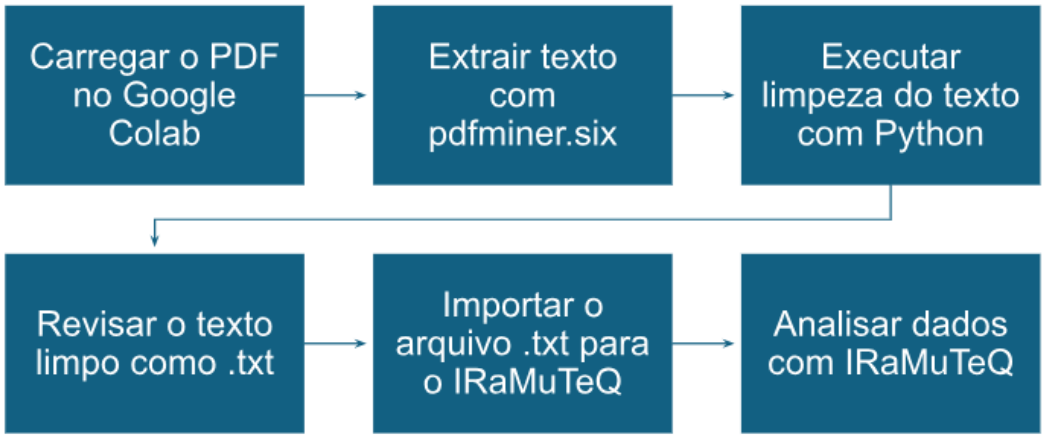
```
# Salva o texto limpo em um novo arquivo
with open('texto_limpo.txt', 'w', encoding='utf-8') as f:
    f.write(texto_limpo)

# Baixar o arquivo de saída automaticamente
files.download('texto_limpo.txt')

print("Texto extraído e limpo salvo em 'texto_limpo.txt' e download iniciado.")
```

3.1.3 Download Automático:

Ao final da execução, o arquivo texto_limpo.txt será baixado automaticamente para o seu computador.

O fluxograma abaixo (figura 1) ilustra as etapas necessárias para a aplicação do método:

Figura 1: Diagrama de fluxo da automação da preparação do corpus textual



Fonte: Elaborado pelos autores (2024).

4 POSSIBILIDADES DE ANÁLISES A PARTIR DOS DADOS GERADOS PELO IRAMUTEQ

As análises possíveis com o IRaMuTeQ oferecem uma ampla gama de aplicações bem descritas por Camargo e Justo (2013) tanto para revisões bibliográficas quanto para análises documentais. Cada método proporciona uma maneira específica de interpretar os dados textuais, permitindo desde a visualização das palavras mais frequentes até a identificação de padrões

complexos de sentido e agrupamento temático. Apesar da proposta de automatização de preparo do corpus textual, o uso dos dados gerados a partir do software devem passar pelo crivo das pessoas pesquisadoras, que deve julgar caso a caso, qual ferramenta do software apresentou resultados que podem contribuir para a resolução do seu problema de pesquisa. A seguir, o quadro 1 apresenta as principais possibilidades de análise com o IRaMuTeQ, exemplificando como cada uma delas pode ser utilizada em diferentes contextos e os tipos de dados que podem ser extraídos em cada caso.

Quadro 1: Aplicações das Análises do IRaMuTeQ em Revisões Bibliográficas e Análises Documentais

Tipo de Análise	Descrição	Aplicações	Exemplos de Dados Extraídos
Estatísticas Textuais Clássicas	Calcula frequências de palavras, identifica hapax (palavras que aparecem uma única vez) e realiza lematização (redução de palavras às suas formas básicas).	Fornece uma visão geral do corpus, identificando termos mais frequentes e padrões lexicais.	Lista de palavras com suas frequências absolutas e relativas; identificação de palavras exclusivas.
Pesquisa de Especificidades de Grupos	Identifica palavras ou expressões características de subgrupos dentro do corpus, auxiliando na comparação entre diferentes segmentos textuais.	Comparação de discursos entre diferentes grupos ou categorias presentes no corpus.	Palavras ou termos que são estatisticamente significativos para determinados grupos ou categorias.
Classificação Hierárquica Descendente (CHD)	Segmenta o texto em unidades menores e as agrupa em classes temáticas com vocabulário semelhante, conforme o método descrito por Reinert.	Identificação de temas ou tópicos principais presentes no corpus, facilitando a compreensão da estrutura temática.	Classes de segmentos textuais com vocabulário homogêneo; dendrogramas representando a relação entre classes.
Análise de Similitude	Constrói redes de coocorrência de	Visualização de conexões entre	Grafos que representam a

	palavras, mostrando como os termos se associam dentro do texto, evidenciando a estrutura das representações sociais.	conceitos e identificação de núcleos centrais em discursos.	proximidade e ligação entre palavras ou conceitos-chave.
Nuvem de Palavras	Gera uma representação visual das palavras mais frequentes no corpus, onde o tamanho de cada palavra é proporcional à sua frequência.	Identificação rápida de temas predominantes e termos mais utilizados no corpus.	Imagem visual destacando as palavras mais frequentes, facilitando a identificação de tópicos centrais.
Análise Fatorial de Correspondência (AFC)	Analisa a correspondência entre linhas e colunas de uma tabela de contingência, permitindo a visualização das relações entre categorias de variáveis qualitativas.	Exploração de associações entre diferentes categorias ou variáveis presentes no corpus.	Mapas fatoriais que mostram a proximidade ou distância entre categorias, auxiliando na interpretação de relações entre variáveis.

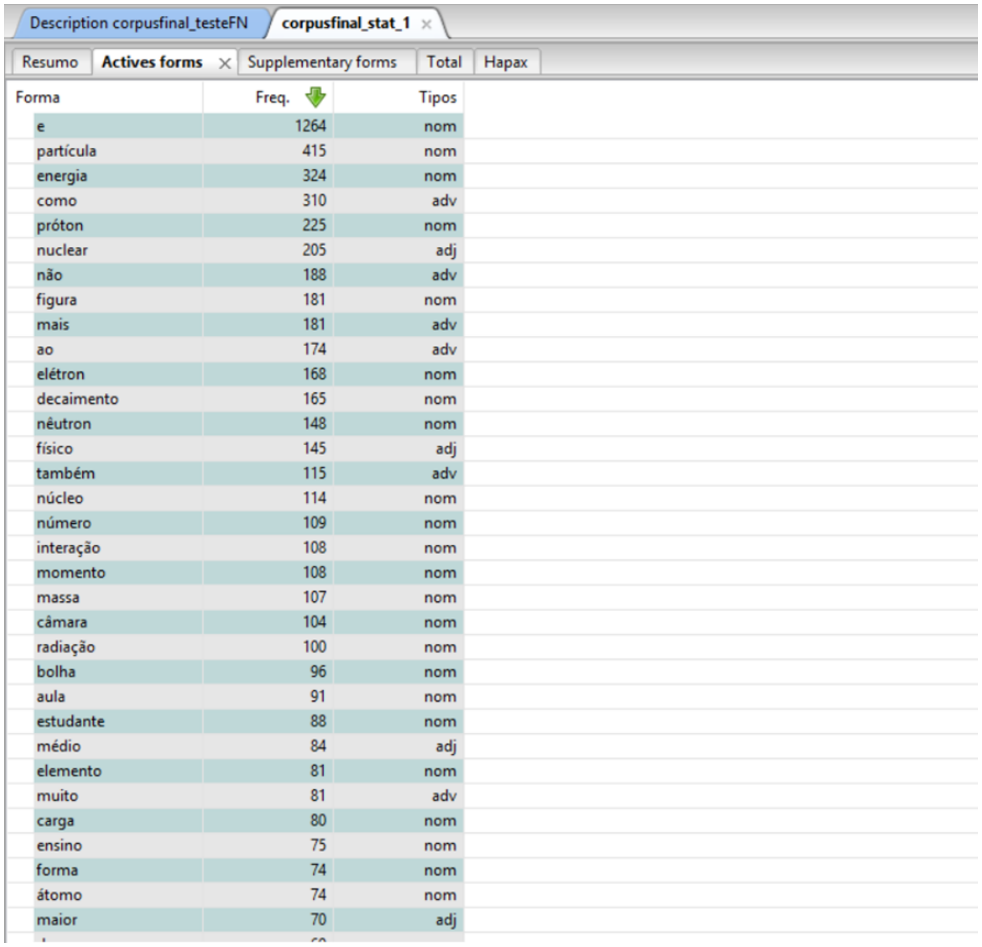
Fonte: Elaborado pelos autores (2024).

5 EXEMPLO DE DADOS EXTRAÍDOS EM REVISÃO BIBLIOGRÁFICA

Para validação da metodologia proposta, foram realizadas análises documentais e bibliográficas. As análises documentais compõem o conjunto de dados analisados em outro artigo, parte do relatório de tese de doutoramento de uma das autoras do presente artigo, e as análises da revisão bibliográfica são apresentadas aqui de forma breve, para ilustrar de que forma essa abordagem pode enriquecer abordagens tradicionais de análise. Foram selecionados três artigos que abordam o ensino de física nuclear na educação básica. Os textos foram limpos de acordo com a metodologia descrita acima, e em seguida, unidos em um único artigo .TXT, revisados e processados no IRaMuTeQ.

As estatísticas textuais clássicas, conforme exemplificadas na figura 2, podem demonstrar assim como as nuvens de palavras, quais os termos mais frequentes, o que pode indicar tendências de pesquisas no campo investigado no caso de análises bibliográficas, ou ainda servir como um banco de dados para a busca da ocorrência de termos específicos, seja em análises bibliográficas ou documentais.

Figura 2 - Estatísticas textuais clássicas



Description corpusfinal_testeFN | corpusfinal_stat_1

Resumo | Actives forms | Supplementary forms | Total | Hapax

Forma	Freq.	Tipos
e	1264	nom
partícula	415	nom
energia	324	nom
como	310	adv
próton	225	nom
nuclear	205	adj
não	188	adv
figura	181	nom
mais	181	adv
ao	174	adv
elétron	168	nom
decaimento	165	nom
nêutron	148	nom
físico	145	adj
também	115	adv
núcleo	114	nom
número	109	nom
interação	108	nom
momento	108	nom
massa	107	nom
câmara	104	nom
radiação	100	nom
bolha	96	nom
aula	91	nom
estudante	88	nom
médio	84	adj
elemento	81	nom
muito	81	adv
carga	80	nom
ensino	75	nom
forma	74	nom
átomo	74	nom
maior	70	adj

Fonte: Elaborado pelos autores (2024).

No que diz respeito a Análise de Especificidades, que identifica palavras ou expressões que ocorrem com frequência significativamente maior em subgrupos de um corpus, como por exemplo, na Imagem 3, em três artigos diferentes, essa ferramenta é capaz de revelar particularidades lexicais associadas a categorias específicas. Em pesquisas de análise documental em especial, a identificação desses padrões linguísticos pode auxiliar na

identificação, por exemplo, a ocorrência ou não de termos de interesse do pesquisador em diferentes documentos, possibilitando a comparação de políticas públicas em diferentes lugares, ou a evolução de determinada questão em documentos de diferentes tempos, entre outras abordagens que dependem das necessidades da pessoa pesquisadora.

No caso do exemplo aqui demonstrado, a ocorrência de determinados termos em determinados artigos deve se dar apenas a diferença temática, mas em outro formato de organização, a análise pode oferecer dados valiosos também a pesquisas de revisão bibliográfica.

Figura 3 – Análise de especificidades

Description corpusfinal_testeFN | corpusfinal_stat_1 | Specificities - corpusfinal_spec_1

Formas | Formas comuns | Tipos | Forms frequencies | Tipos de frequências | Frequência relativa das formas | Tipos de frequências relativas | AFC

formes	*artigo_1	*artigo_2	*artigo_3
nêutron	20.0562	-4.1969	-14.2178
aula	16.0156	-1.2057	-15.1267
nuclídeos	14.9515	-2.2335	-11.4542
nuclídeo	12.2819	-2.4859	-8.8173
fissão	11.3311	-1.6928	-8.6807
ligação	11.0898	-1.6567	-8.4959
que	9.9598	-1.9484	-7.2935
nuclear	9.8228	4.0871	-23.2078
científico	9.6422	-1.4405	-7.387
fusão	9.201	-0.6396	-9.2353
núcleo	8.5139	-0.7233	-7.8941
mais	8.2596	-1.6767	-6.0975
4he	7.9538	-1.1883	-6.0935
sequência	7.9132	-2.0893	-5.3498
átomo	7.8783	-1.2245	-6.2533
rutherford	7.4715	-1.1162	-5.724
elemento	7.1233	1.255	-13.3254
acima	7.005	-1.2603	-5.1717
essa	6.7784	-0.2697	-7.5857
usina	6.7759	-0.6287	-6.2783
carbono	6.748	-1.0082	-5.1698

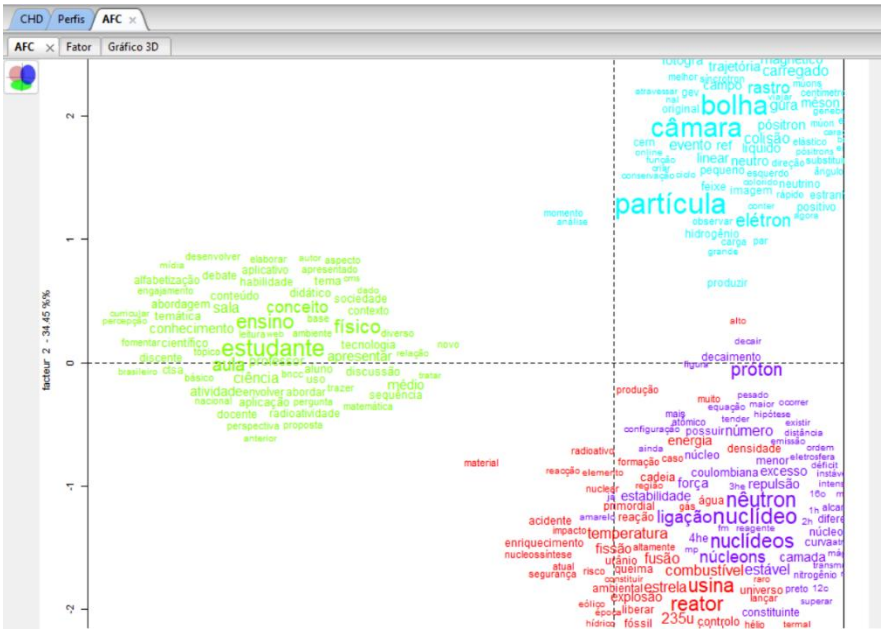
Fonte: Elaborado pelos autores (2024).

A classificação pelo método de Reinert, também conhecido como Classificação Hierárquica Descendente (CHD), é uma abordagem eficaz para análises bibliográficas e documentais, pois permite identificar e agrupar segmentos de texto com vocabulário semelhante, revelando estruturas temáticas subjacentes nos dados, como por exemplo, revelando que determinados termos são mais frequentes e agrupados conjuntamente em determinados documentos, indicando temas ou preocupações particulares de grupos de documentos ou artigos, além do mapeamento de relações temáticas proporcionados pela análise de AFC, que demonstra como conceitos se associam a categorias específicas,

dando recursos para identificar a estrutura temática do corpus textual analisado.

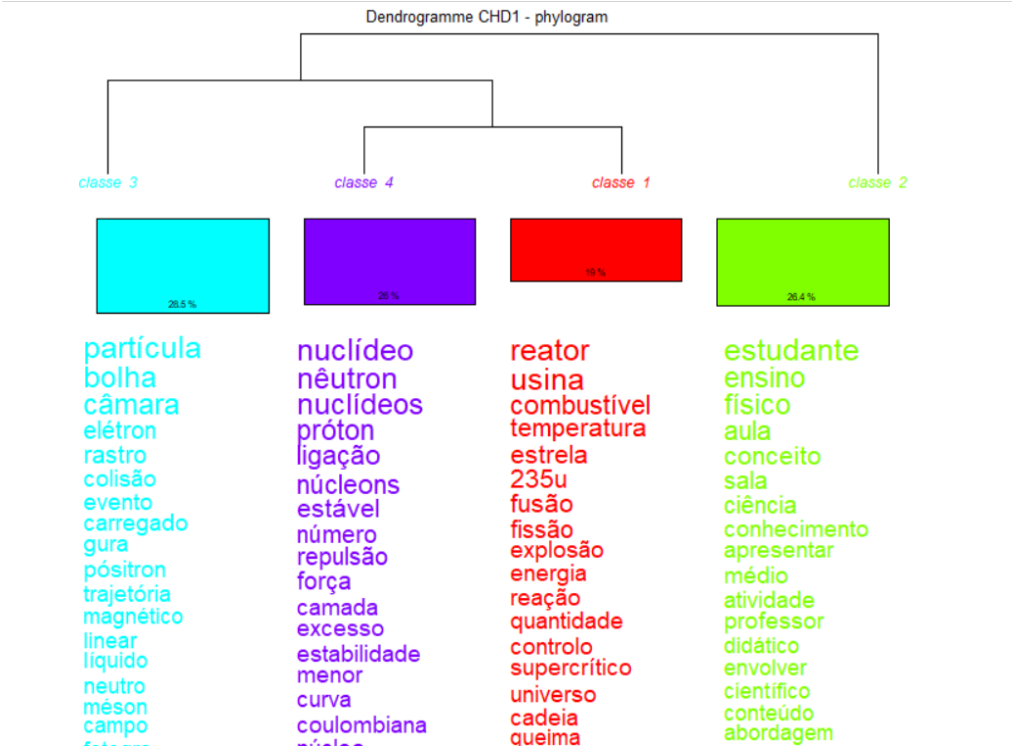
Em outras palavras, é possível observar sob qual abordagem teórica um determinado assunto é tratado, ao considerarmos as palavras que compõem categorias específicas, e como essas categorias se relacionam entre elas, como pode ser visualizado na figura 4, ou mesmo quais temáticas são evidenciadas e quais são negligenciadas no conjunto de dados, visto que uma porcentagem é apontada nas análises conforme visualizado na figura 5.

Figura 4 – Classificação CHD



Fonte: Elaborado pelos autores (2024).

Figura 5 – Dendograma



Fonte: Elaborado pelos autores (2024).

Já por meio da análise de similitude, é possível visualizar, em forma de grafo, como os termos do corpus em questão se relacionam, destacando núcleos centrais e suas ramificações periféricas. Em pesquisas de análise documental e bibliográfica, a análise de similitude pode revelar, por exemplo, no caso da figura 6 que o termo "partícula" está centralmente conectado a palavras como "energia", "elétron" e "como", que se desdobra em palavras como “nuclear”, “próton” e “estudante” respectivamente, indicando que esses conceitos estão intimamente ligados no contexto estudado, facilitando a identificação de padrões semânticos e a estruturação dos conteúdos presentes nos documentos analisados.

Figura 6 – Dendograma

Fonte: Elaborado pelos autores (2024).

Além da aplicação em revisões bibliográficas, a metodologia proposta foi testada com êxito em *corpus* de natureza documental, confirmando sua versatilidade e robustez. Um exemplo relevante é sua utilização no estudo “A (In)visibilidade da Maternidade em Documentos Curriculares: uma análise com o Software Iramuteq”, parte integrante da tese de doutorado de Barbosa (2025). Nesse caso, foram processados dois volumes do Currículo Paulista (São Paulo, 2019; 2020), totalizando mais de 300 páginas de conteúdo normativo. A automação mostrou-se eficiente mesmo frente à extensão e complexidade de formatação dos documentos, mantendo a integridade textual e eliminando caracteres incompatíveis com o IRaMuTeQ. O processamento automatizado possibilitou análises lexicais e de frequência que evidenciaram, por exemplo, a ausência de termos como “maternidade”, “paternidade” e “amamentação” em ambos os documentos, revelando um silenciamento temático relevante para a discussão sobre gênero e cuidado na política curricular. Esses resultados

reforçam que o método pode ser aplicado a diferentes gêneros e formatos textuais, incluindo produções científicas e documentos institucionais de grande porte, ampliando seu potencial de uso em diversos campos da pesquisa qualitativa.

6 CONCLUSÃO

A utilização de ferramentas como Python e Google Colab oferece uma solução prática e eficiente para automatizar a preparação de corpus textual para o IRaMuTeQ, superando desafios relacionados à limpeza e formatação de textos extensos. Esse processo automatizado não apenas reduz significativamente o tempo demandado para a formatação manual, mas também facilita a análise de grandes volumes de dados em pesquisas qualitativas, ampliando as possibilidades de aplicação do *software* em revisões bibliográficas e análises documentais. Assim, ao incorporar tecnologias de código aberto e automação ao processo de pesquisa, os pesquisadores podem se dedicar mais à interpretação dos dados, promovendo uma análise mais abrangente e produtiva.

Três dificuldades foram encontradas ao utilizar essa metodologia, e apesar de demandarem um cuidado especial no processo de revisão, ainda demandam uma quantidade de tempo menor do que a elaboração manual total do corpus. Uma dificuldade encontrada na utilização da metodologia proposta diz respeito ao fato de que a limpeza dos textos deve ser feita de maneira individualizada, PDF por PDF, e depois, caberá ao pesquisador unir todos os textos em um arquivo .TXT separando os textos por linhas de comando conforme descrito por outros autores, como Camargo e Justo (2013). Além disso, durante o processo de revisão, algumas palavras se encontravam divididas ao meio, com espaço simples entre elas. Isso ocorre quando o texto PDF tem, devido a sua formatação, palavras divididas em mais de uma linha. Essa situação pode ou não ocorrer, dependendo da diagramação do documento analisado. Durante a revisão, cabe ao pesquisador fazer a correção desses casos, para evitar que alguns termos não sejam contados pelo software. Por fim, em caso de textos com numeração de páginas, essas também ficam presentes como parte do texto. Essa numeração poderá ser excluída pelo pesquisador durante a revisão do

texto.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

REFERÊNCIAS

ACAUAN, L. V.; ABRANTES, C. V.; STIPP. M. A. C.; TROTTE, L. A. C.; PAES, G. O.; QUEIROZ, A. B. A. Utilização do software Iramuteq® para análise de dados qualitativos na Enfermagem: um ensaio reflexivo. **REME-Revista Mineira de Enfermagem**, [*S.l.*], v. 24, n. 1, 2020.

ALMEIDA, L. M. N.; GOULART, M.C.L.; GÓES, F.G. B.; PEREIRA-ÁVILA, F. M. V.; PINTO, C. B.; SILVA, A. C. S. S.; GARCIA, L. R.; BRUN, L. S. O. Continuidade do aleitamento materno no retorno ao trabalho: sentimentos, desafios e estratégias de enfermeiras nutrizes. **Revista Gaúcha de Enfermagem**, [*S.l.*], v. 44, p. 1-15, 2023.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de informática na educação**, v. 19, n. 02, p. 03, 2011.

BARBOSA, R. O. **Maternagem, educação e emancipação:** uma análise curricular. 2025. 206 f. Tese (Doutorado em Educação para a Ciência) - Faculdade de Ciências, Universidade Estadual Paulista, Bauru, 2025.

BUENO, A. J. A**. Uma análise por meio do software Iramuteq de teses e dissertações defendidas entre 2007 e 2017 com a temática filmes comerciais no ensino de ciências.** 2018. Dissertação (Mestrado em Ensino de Ciências e educação Matemática) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2018.

CAMARGO, B. V.; JUSTO, A. M. IRAMUTEQ: um software gratuito para análise de dados textuais. **Temas em psicologia**, [*S.l.*], v. 21, n. 2, p. 513-518, 2013.

SÃO PAULO (Estado). Secretaria da Educação. **Currículo Paulista**: etapa Ensino Médio. São Paulo: SEE, 2020.

SÃO PAULO (Estado). Secretaria da Educação. **Currículo Paulista**: etapa Educação Infantil e Ensino Fundamental. São Paulo: SEE, 2019.

HOFFMANN, Y. T.; ALVAREZ, E. B.; MARTÍ-LAHERA, Y. Análise textual com IRaMuTeQ de pesquisasrecentes em História da educação matemática no Brasil: um exemplo de Humanidades Digitais. **Investigación Bibliotecológica**: Archivonomía, Bibliotecología e Información, Cidade do México, v. 34, n. 84, p. 103-133, 2020.

MELO, N. M.; VASCONCELOS, A. M.; LIMA, T. N. Percepção Ambiental e Biofilia nos Parques Urbanos: Uma Revisão Bibliográfica. **Revista Pantaneira**, [*S.l.*], v. 22, p. 42-53, 2023.

MENDES, L. O. R.; PROENÇA, M. C.; PEREIRA, A. L. O" software" IRaMuTeQ na pesquisa qualitativa: ma revisão no campo da Educação Matemática. **Paradigma**, [*S.l.*], n. 2, p. 228-258, 2022.

MOTTA, J. A.; FIÚZA, A. L. C. Mulheres na ciência: uma análise sistematizada dos artigos científicos publicados no Brasil pós-década de 1990. **Cadernos de Gênero e Tecnologia**, [*S.l.*], v. 15, n. 46, p. 46-63, 2022.

NASCIMENTO, V. B.; SANTOS, L. A.; SARAIVA, R. S. A. Softwares de análise de dados qualitativos: revisão narrativa da literatura. **Revista Científica da Faculdade de Educação e Meio Ambiente**, [*S.l.*], v. 13, n. 1, p. 44-58, 2022.

RAMOS, M. G.; LIMA V. M. R; ROSA, M. P. Contribuições do software Iramuteq para a análise textual discursiva. *In*: COSTA, A. P.; SOUZA, D. N.; CASTRO, P. A.; SAAVEDRA, R. A.; SÁ, S. O. **Atas do 7o Congresso Ibero-Americano em Investigação Qualitativa em Educação**. Fortaleza: Universidade de Fortaleza; 2018. p. 505-14.

RATINAUD, P. **IRAMUTEQ**: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. Disponível em: http://www.iramuteq.org. 2008.

SANTOS, J. S.; ALMEIDA JUNIOR. E. R. B.; BRITO. A. A. NOGUEIRA, G. Tecnologia na enfermagem: uma revisão bibliográfica. **Research, Society and Development**, *[S. l.]*, v. 11, n. 3, p. e54811327051, 2022.

SCHLOSSER, D. F.; FRASSON, A. C.; CANTORANI, J. R. H. Softwares livres para análise de dados qualitativos. **Revista Brasileira de Ensino de Ciência e Tecnologia**, [*S.l.*], v. 12, n. 1, 2019.

SILVA, J. V.; SOUZA, P. A. R. O Ambiente de Gestão de Microempresas: uma análise a partir do software IRAMUTEQ. **Revista de tecnologia aplicada**, [*S.l.*], v. 8, n. 3, p. 54-66, 2020.

SILVA, S.; RIBEIRO, E. A. W. O software Iramuteq como ferramenta metodológica para análise qualitativa nas pesquisas em Educação Profissional e Tecnológica. **Cadernos de Educação Tecnologia e Sociedade**, [*S.l.*], v. 14, n. 2, p. 275–284, jun. 2021.

SILVA, T. R.; DELL'AGLIO, D. D. Abuso sexual contra crianças e adolescentes e as consequências psicológicas: Revisão bibliográfica com análise de similitude. **Psicologia e Saúde em debate**, [*S.l.*], v. 9, n. 2, p. 653-669, 2023.

SOUZA, M. A. R. O uso do software IRAMUTEQ na análise de dados em pesquisas qualitativas. **Revista da Escola de Enfermagem da USP**, [*S.l.*], v. 52, p. e03353, 2018.

SOUZA, R. F.; CARVALHO, P. R.; RAMOS, M. G. 50 anos do PPGCI IBICT: análise textual da produção científica com iramuteq. **Informação & Informação**, [*S.l.*], v. 25, n. 4, p. 117-141, 2020.

TINTI, D. S.; BARBOSA, G. C.; LOPES, C. E. O software IRAMUTEQ e a Análise de Narrativas (Auto) biográficas no Campo da Educação Matemática. **Bolema: Boletim de Educação Matemática**, [*S.l.*], v. 35, p. 479-496, 2021.

AUTOMATED PREPARATION OF TEXTUAL CORPUS FOR QUALITATIVE RESEARCH WITH IRAMUTEQ

ABSTRACT
Objective: To present an automated method for preparing textual corpora in qualitative research using IRaMuTeQ, reducing time and effort in text cleaning and formatting. **Methodology**: A Python script was developed and executed in Google Colab to automate text extraction from PDF files, remove incompatible characters, and structure the corpus according to IRaMuTeQ standards. The methodology was tested with different datasets to verify its effectiveness. **Results**: The automated approach eliminated inconsistencies in text formatting and significantly reduced the time required for corpus preparation, making qualitative analysis more efficient and accurate. The methodology enabled greater standardization and replicability in textual data processing. **Conclusions**: The proposed automation facilitates the adoption of IRaMuTeQ in large-scale qualitative research, eliminating technical barriers and enabling more detailed analyses. The integration of computational tools with textual analysis optimizes data organization, improving the quality and reliability of results.

Keywords: Document analysis. Literature review. Google Colab. Python.

PREPARACIÓN AUTOMATIZADA DE CORPUS TEXTUAL PARA INVESTIGACIONES CUALITATIVAS CON IRAMUTEQ

RESUMEN
Objetivo: Presentar un método automatizado para la preparación de corpus textual en investigaciones cualitativas utilizando IRaMuTeQ, reduciendo el tiempo y esfuerzo en la limpieza y el formato de textos. **Metodología:** Se desarrolló un código en Python, ejecutado en Google Colab, para automatizar la extracción de texto de archivos PDF, eliminar caracteres incompatibles y estructurar el corpus según los estándares exigidos

por IRaMuTeQ. La metodología fue probada con diferentes conjuntos de datos para verificar su eficacia. **Resultados:** El enfoque automatizado eliminó inconsistencias en el formato de los textos y redujo significativamente el tiempo necesario para la preparación del corpus, haciendo el análisis cualitativo más eficiente y preciso. La metodología permitió una mayor estandarización y replicabilidad en el procesamiento de datos textuales. **Conclusiones:** La automatización propuesta facilita la adopción de IRaMuTeQ en investigaciones cualitativas a gran escala, eliminando barreras técnicas y permitiendo análisis más detallados. La integración de herramientas computacionales con el análisis textual optimiza la organización de los datos, mejorando la calidad y confiabilidad de los resultados.

Descriptores: Análisis documental. Revisión bibliográfica. Google Colab. Python.

Recebido em: 07.02.2025
Aceito em: 08.09.2025