Artigo Original

Análise lexicométrica com IRaMuTeQ: recomendações metodológicas para preparação de corpus textual

Lexicometric analysis using IRaMuTeQ: methodological recommendations for textual corpus preparation

Análisis lexicométrico con IRaMuTeQ: recomendaciones metodológicas para la preparación del corpus textual

Clarissa Coelho Vieira Guimarães
coelho.clarissa@unifesp.br
https://orcid.org/0000-0002-7713-7182

Revista de Pesquisa Cuidado é Fundamental Online vol. 18 14805 2026

Universidade Federal do Estado do Rio de Janeiro
Brasil

Recepción: 25 Marzo 2026
Aprobación: 10 Abril 2026

Resumo: **Objetivo:** sistematizar recomendações metodológicas para a preparação e organização de corpus textual destinado à análise lexicométrica em pesquisas qualitativas na área da saúde. **Método:** trata-se de um estudo metodológico, de natureza teórico-aplicada, fundamentado na literatura científica sobre análise de dados qualitativos e no uso do software IRAMUTEQ. Foram sistematizados procedimentos relacionados à organização do corpus, padronização linguística, definição de variáveis analíticas, segmentação textual e estruturação do material para processamento computacional. **Resultados:** foram descritas etapas operacionais e recomendações metodológicas para a preparação do corpus textual, incluindo padronização lexical, organização das unidades de contexto inicial, definição de variáveis e codificação do material textual, visando garantir maior rigor metodológico, rastreabilidade e reprodutibilidade das análises lexicométricas. **Conclusão:** a preparação do corpus textual constitui etapa metodológica fundamental para a análise lexicométrica no IRAMUTEQ, influenciando diretamente a qualidade, a estabilidade e a confiabilidade dos resultados obtidos, devendo ser compreendida como uma etapa estruturante da pesquisa qualitativa apoiada por software.

Palavras-chave: Pesquisa qualitativa, Mineração de dados, Processamento de linguagem natural, Enfermagem, Métodos de pesquisa..

Abstract: **Objective:** to systematize methodological recommendations for the preparation and organization of textual corpus for lexicometric analysis in qualitative health research. **Method:** this is a methodological study of a theoretical-applied nature, based on the scientific literature on qualitative data analysis and on the use of the IRAMUTEQ software. Procedures related to corpus organization, linguistic standardization, definition of analytical variables, textual segmentation and structuring of the material for computational processing were systematized. **Results**: Operational steps and methodological recommendations for textual corpus preparation were described, including lexical standardization, organization of initial context units, definition of variables and coding of textual material, aiming to ensure greater methodological rigor, traceability and reproducibility of lexicometric analyses. **Conclusion:** The preparation of the textual corpus constitutes a fundamental methodological step for lexicometric analysis in IRAMUTEQ, directly influencing the quality, stability and reliability of the results obtained, and should be understood as a structuring step of qualitative research supported by software.

Keywords: Qualitative research, Data mining, Natural language processing, Nursing, research methods.

Resumen: **Objetivo**: sistematizar recomendaciones metodológicas para la preparación y organización de corpus textual destinado al análisis lexicométrico en investigaciones cualitativas en el área de la salud. **Método:** se trata de un estudio metodológico, de naturaleza teórico-aplicada, fundamentado en la literatura científica sobre análisis de datos cualitativos y en el uso del software IRAMUTEQ. Se sistematizaron procedimientos relacionados con la organización del corpus, estandarización lingüística, definición de variables analíticas, segmentación textual y estructuración del material para procesamiento computacional. **Resultados:** se describieron etapas operativas y recomendaciones metodológicas para la preparación del corpus textual, incluyendo estandarización léxica, organización de las unidades de contexto inicial, definición de variables y codificación del material textual, con el fin de garantizar mayor rigor metodológico, trazabilidad y reproducibilidad de los análisis lexicométricos. **Conclusión:** la preparación del corpus textual constituye una etapa metodológica fundamental para el análisis lexicométrico en IRAMUTEQ, influyendo directamente en la calidad, estabilidad y confiabilidad de los resultados obtenidos, y debe ser comprendida como una etapa estructurante de la investigación cualitativa apoyada por software.

Palabras clave: Investigación cualitativa, Minería de datos, Procesamiento del lenguaje natural, Enfermería, Métodos de investigación.

INTRODUÇÃO

A análise de dados qualitativos desempenha papel fundamental na produção de conhecimento nas ciências sociais e da saúde, ao possibilitar a compreensão de significados, experiências e processos sociais que não podem ser plenamente apreendidos por abordagens quantitativas. Diferentes estratégias metodológicas têm sido utilizadas para análise de dados textuais, incluindo análise de conteúdo, análise temática e abordagens lexicométricas apoiadas por ferramentas computacionais.[1-4]

Com o avanço das tecnologias digitais aplicadas à pesquisa científica, o uso de softwares de análise textual tem se ampliado, permitindo o processamento de grandes volumes de dados qualitativos por meio de técnicas de análise lexical e estatística textual.[1,5] Entre essas ferramentas, destaca-se o IRAMUTEQ® (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*), software baseado na linguagem estatística R, que possibilita a realização de diferentes análises lexicais, como estatísticas textuais clássicas, análise de similitude, Classificação Hierárquica Descendente e Análise Fatorial de Correspondência.[5-6] Essas análises permitem identificar padrões lexicais, estruturas semânticas e relações entre segmentos textuais, contribuindo para a interpretação sistematizada de dados qualitativos.

Na área da saúde e da enfermagem, o uso do IRAMUTEQ® tem sido crescente, sendo aplicado em estudos que analisam narrativas, documentos institucionais, políticas públicas e produções científicas, evidenciando seu potencial como ferramenta de apoio à pesquisa qualitativa.[7-9] Entretanto, a preparação do corpus textual constitui uma das etapas mais desafiadoras no desenvolvimento de pesquisas que utilizam análise lexicométrica, especialmente em estudos desenvolvidos no contexto da formação de pesquisadores, como trabalhos de conclusão de curso, dissertações e teses.

Dificuldades relacionadas à organização dos dados, à padronização linguística, à definição de variáveis analíticas e à estruturação do corpus para processamento no software são frequentemente observadas, podendo comprometer a qualidade das análises e a reprodutibilidade dos estudos. Apesar da crescente utilização do IRAMUTEQ®, muitos estudos descrevem de forma limitada os procedimentos metodológicos relacionados à preparação do corpus textual, o que evidencia a necessidade de sistematização de recomendações metodológicas para essa etapa da pesquisa.

Diante desse contexto, torna-se relevante discutir estratégias que contribuam para o rigor metodológico na preparação do corpus textual, favorecendo a qualidade das análises lexicométricas e a

reprodutibilidade científica. Assim, o presente estudo tem como objetivo sistematizar recomendações metodológicas para a preparação de corpus textual destinado à análise lexicométrica no software IRAMUTEQ® em pesquisas qualitativas na área da saúde.

MÉTODO

Trata-se de um estudo metodológico, de natureza teórico-aplicada, que objetivou sistematizar recomendações metodológicas para a preparação de material textual destinado à análise lexicométrica no software IRAMUTEQ® em pesquisas qualitativas na área da saúde. Estudos metodológicos caracterizam-se pela proposição, sistematização e aprimoramento de procedimentos, técnicas e estratégias de investigação, contribuindo para o rigor metodológico, a padronização de métodos e a reprodutibilidade científica.[1-3] O desenvolvimento do estudo seguiu recomendações relacionadas à transparência e à qualidade do relato científico, conforme diretrizes da EQUATOR Network.[1-4]

A análise lexical foi realizada com o auxílio do software IRAMUTEQ®, ferramenta computacional integrada à linguagem estatística R, utilizando procedimentos de estatística textual baseados no método de Reinert.[1-2] Inicialmente, foram geradas estatísticas textuais clássicas para identificação da frequência e distribuição das formas lexicais no corpus textual. Em seguida, realizaram-se análises de similitude, Classificação Hierárquica Descendente (CHD) e Análise Fatorial de Correspondência (AFC), com o objetivo de identificar padrões de coocorrência lexical, classes semânticas e estruturas de organização do vocabulário.

Para a interpretação das classes lexicais, foram considerados os valores de qui-quadrado (χ^2), adotando-se como critério de associação significativa das formas lexicais às classes valores de $\chi^2 \geq 3{,}84$ e $p < 0{,}05$, conforme pressupostos do método de Reinert.[1-2] A estabilidade da Classificação Hierárquica Descendente foi avaliada por meio da taxa de retenção dos segmentos de texto, sendo considerados adequados valores iguais ou superiores a 70%.

A construção do estudo fundamentou-se na análise de publicações científicas que abordam a análise textual e o uso do IRAMUTEQ em pesquisas qualitativas, incluindo artigos metodológicos, estudos empíricos e manuais de utilização da ferramenta, com a finalidade de identificar recomendações relacionadas à preparação, organização e padronização do corpus textual para análise lexicométrica.[5-8]

O corpus textual utilizado para fins ilustrativos foi estruturado em 40 segmentos de texto, organizados de acordo com variáveis descritivas relacionadas ao perfil profissional e à área temática. Cada segmento textual foi codificado conforme as orientações do software,

utilizando a marcação iniciada por cinco asteriscos (*****), seguida das variáveis de identificação, permitindo a realização de análises comparativas entre categorias e grupos de interesse. O material textual utilizado para fins ilustrativos encontra-se disponibilizado em repositório de acesso aberto no Zenodo[9-11], acessível por meio do DOI: https://doi.org/10.5281/zenodo.19008125.

A partir da análise das publicações e do material textual, foram sistematizadas recomendações metodológicas relacionadas à preparação do corpus textual, contemplando procedimentos de transcrição, revisão textual, padronização lexical, definição de variáveis analíticas, segmentação textual e estruturação do corpus para processamento no software. Para fins ilustrativos, foram geradas representações gráficas a partir de material textual previamente analisado em pesquisa qualitativa na área da saúde, com o objetivo de demonstrar as possibilidades analíticas do software e exemplificar a aplicação das etapas de preparação do corpus textual.

A preparação do corpus textual constitui etapa metodológica estruturante das análises lexicométricas, uma vez que influencia a retenção dos segmentos de texto, a estabilidade das classes lexicais e a consistência das análises estatísticas realizadas pelo software. Como produto metodológico deste estudo, foi elaborada uma sistematização das etapas para preparação de corpus textual para análise lexicométrica.

RESULTADOS

Preparação do corpus textual

Os resultados deste estudo evidenciam que a preparação do corpus textual constitui uma etapa metodológica determinante para a qualidade das análises lexicométricas realizadas com o software IRAMUTEQ®, influenciando diretamente a retenção dos segmentos de texto, a estabilidade das classes lexicais e a consistência das análises estatísticas.

A preparação do corpus textual constitui uma etapa estruturante na realização de análises lexicométricas com o software IRAMUTEQ®. Essa fase envolve um conjunto de procedimentos metodológicos voltados à organização, padronização e estruturação dos dados textuais que serão posteriormente submetidos ao processamento estatístico do software. A qualidade dessa etapa influencia diretamente a consistência das análises e a confiabilidade das classes lexicais geradas.

Inicialmente, realiza-se a transcrição dos dados empíricos, etapa que corresponde à transformação de entrevistas, narrativas ou documentos em material textual passível de processamento computacional. Recomenda-se que essa transcrição seja realizada de

forma fiel ao conteúdo original, preservando o sentido das falas ou registros analisados.

Após a transcrição, procede-se à revisão textual do corpus, com o objetivo de corrigir inconsistências ortográficas, padronizar grafias e uniformizar termos recorrentes. Essa etapa é fundamental para reduzir a dispersão lexical, situação que pode ocorrer quando diferentes grafias são utilizadas para representar um mesmo conceito.

Outra fase importante refere-se à padronização linguística do corpus textual. Nesse momento são eliminados elementos que podem comprometer a leitura pelo software, como caracteres especiais, símbolos gráficos ou formatações incompatíveis com arquivos de texto simples. A padronização também inclui a uniformização de termos técnicos e expressões utilizadas ao longo do corpus.

Recomenda-se a padronização de expressões compostas que representam um único conceito analítico. No software IRAMUTEQ®, palavras separadas por espaço são interpretadas como unidades lexicais independentes, o que pode fragmentar o significado de determinados termos durante a análise lexicométrica e aumentar a dispersão lexical do corpus. Assim, expressões compostas podem ser unificadas por meio do uso do caractere sublinhado (_), preservando sua integridade semântica e favorecendo a consistência das análises estatísticas realizadas pelo software. Exemplos de padronização lexical de expressões compostas são apresentados no Quadro 1.

Quadro 1

Exemplos de padronização lexical em corpora para análise no IRAMUTEQ®.

Expressão original	Forma padronizada
trabalho em equipe	trabalho_em_equipe
Segunda-feira	Segunda_feira
segurança do paciente	segurança_do_paciente

Elaborado pela autora (2026).

Posteriormente, realiza-se a definição das variáveis analíticas que irão identificar os segmentos de texto. Essas variáveis permitem a realização de análises comparativas entre grupos, categorias ou características dos participantes, constituindo um elemento fundamental para a exploração analítica do corpus. Concluídas essas etapas, o corpus textual é estruturado em um único arquivo de texto (.txt), organizado em segmentos precedidos por linhas de comando que identificam as variáveis analíticas do estudo.

Tabela 1

Etapas para preparação de corpus textual para análise no IRAMUTEQ®

Etapa	Descrição	Recomendações metodológicas
Transcrição dos dados	Conversão de entrevistas, narrativas ou documentos para formato textual	Realizar transcrição fiel do material empírico
Revisão textual	Verificação da consistência linguística do corpus	Corrigir erros ortográficos e padronizar grafias
Padronização lexical	Uniformização de termos e expressões	Evitar variações linguísticas
Limpeza do corpus	Remoção de elementos que dificultam o processamento	Excluir caracteres especiais
Definição de variáveis	Identificação de categorias analíticas	Utilizar linhas de comando padronizadas
Estruturação do corpus	Organização do arquivo textual	Construir um único arquivo .txt
Verificação final	Conferência do corpus antes da análise	Revisar manualmente o arquivo

Elaborado pela autora (2026).

A sistematização dessas etapas contribui para reduzir inconsistências na organização do corpus textual e favorecer a estabilidade das análises lexicométricas. Para assegurar a consistência estatística das análises, recomenda-se observar a taxa de retenção dos Segmentos de Texto (ST) processados na Classificação Hierárquica Descendente (CHD), sendo recomendadas taxas superiores a 70% para garantir estabilidade das classes lexicais.

Durante o processamento do corpus textual, o software IRAMUTEQ® realiza procedimentos de lematização, agrupando diferentes formas de uma mesma palavra em uma forma lexical base. Nesse processo, verbos são convertidos para o infinitivo e substantivos e adjetivos são reduzidos à forma masculina singular, o que contribui para a padronização do vocabulário e para a identificação de relações semânticas entre os termos presentes no corpus analisado.

Desafios na organização do corpus

Durante o processo de preparação do corpus textual podem surgir diferentes desafios metodológicos, especialmente em pesquisas qualitativas desenvolvidas no contexto da formação de pesquisadores. A organização inadequada do corpus constitui uma das principais causas de dificuldades na execução das análises lexicométricas no software IRAMUTEQ®, podendo comprometer tanto o processamento dos dados quanto a interpretação dos resultados obtidos.

Entre os desafios mais frequentes destacam-se inconsistências na transcrição dos dados, ausência de padronização lexical e fragmentação inadequada dos segmentos textuais. Problemas na

revisão ortográfica ou na uniformização de termos podem gerar dispersão lexical, dificultando a identificação de padrões semânticos e reduzindo a consistência estatística das classes lexicais formadas.

Outro desafio recorrente refere-se à utilização de grafias distintas para palavras com o mesmo significado, situação que pode provocar fragmentação na frequência lexical e interferir na organização das classes semânticas identificadas na análise.

Adicionalmente, dificuldades relacionadas à estruturação das variáveis analíticas também são frequentemente observadas. A definição inadequada das variáveis ou a inconsistência na codificação dos segmentos textuais pode comprometer a organização do corpus e limitar a realização de análises comparativas entre diferentes categorias de participantes ou contextos analisados.

Outro aspecto relevante diz respeito à delimitação dos segmentos de texto. Segmentos excessivamente longos ou excessivamente fragmentados podem interferir no desempenho das análises lexicométricas, uma vez que a Classificação Hierárquica Descendente depende da distribuição equilibrada do vocabulário entre os segmentos textuais analisados.

Diante desses desafios, torna-se fundamental que o pesquisador realize uma revisão cuidadosa do corpus textual antes da execução das análises no software IRAMUTEQ®. A adoção de critérios claros para transcrição, padronização linguística, definição de variáveis e segmentação do texto contribui para maior consistência metodológica e para a robustez das análises lexicométricas realizadas.

Quadro 2 — Exemplo de estrutura de codificação do corpus textual para análise no software IRAMUTEQ®

Quadro 2

Exemplo de estrutura de codificação do corpus textual para análise no software IRAMUTEQ®

****** *Entrevista_01 *Sexo_Feminino *Profissao_Enfermeira**
A experiência no cuidado ao paciente em situação crítica exige preparo técnico e apoio da equipe. Muitas vezes o profissional precisa tomar decisões rápidas durante a assistência.
****** *Entrevista_02 *Sexo_Masculino *Profissao_Enfermeiro**
O trabalho em equipe é fundamental para garantir a segurança do paciente. A comunicação entre os profissionais facilita a tomada de decisão e a organização do cuidado.
****** *Entrevista_03 *Sexo_Feminino *Profissao_Enfermeira**
Durante o atendimento emergencial, a atuação do enfermeiro envolve avaliação clínica contínua e monitoramento dos sinais vitais do paciente.

Elaborado pela autora (2026).

A adoção de padrões de codificação contribui para a organização sistemática do corpus textual e para a adequada identificação das variáveis analíticas exploradas nas análises lexicométricas.

Aplicações analíticas do IRAMUTEQ

Após a preparação e organização do corpus textual, o software IRAMUTEQ® possibilita a realização de diferentes análises lexicais que contribuem para a exploração sistematizada de dados textuais em pesquisas qualitativas. Essas análises permitem identificar padrões lexicais, estruturas semânticas e relações entre segmentos de texto, auxiliando o pesquisador na organização e interpretação do material qualitativo.

Entre as análises mais utilizadas destacam-se as estatísticas textuais clássicas, a análise de similitude, a Classificação Hierárquica Descendente (CHD) e a Análise Fatorial de Correspondência (AFC). Cada uma dessas análises apresenta finalidades específicas e contribui de forma complementar para a interpretação dos dados textuais.

As estatísticas textuais clássicas constituem a etapa inicial da análise lexicométrica e permitem identificar indicadores importantes para a avaliação da qualidade do corpus, como o número de textos, o número de ocorrências, o número de formas lexicais e a taxa de retenção dos segmentos de texto. Esses indicadores permitem ao pesquisador avaliar se o corpus apresenta condições adequadas para a realização das análises lexicométricas, sendo a taxa de retenção considerada um dos principais parâmetros de qualidade do corpus.

A análise de similitude baseia-se na teoria dos grafos e permite identificar relações de coocorrência entre palavras, possibilitando a visualização da estrutura de conexões entre os termos presentes no corpus textual, sendo amplamente utilizada em análises lexicométricas para a identificação de estruturas semânticas em dados textuais.[16] Essa técnica contribui para a identificação de núcleos semânticos e para a compreensão da organização lexical do discurso analisado. A Figura 1 apresenta um exemplo de análise de similitude gerada pelo software IRAMUTEQ®, ilustrando a estrutura de relações entre os termos presentes no corpus.

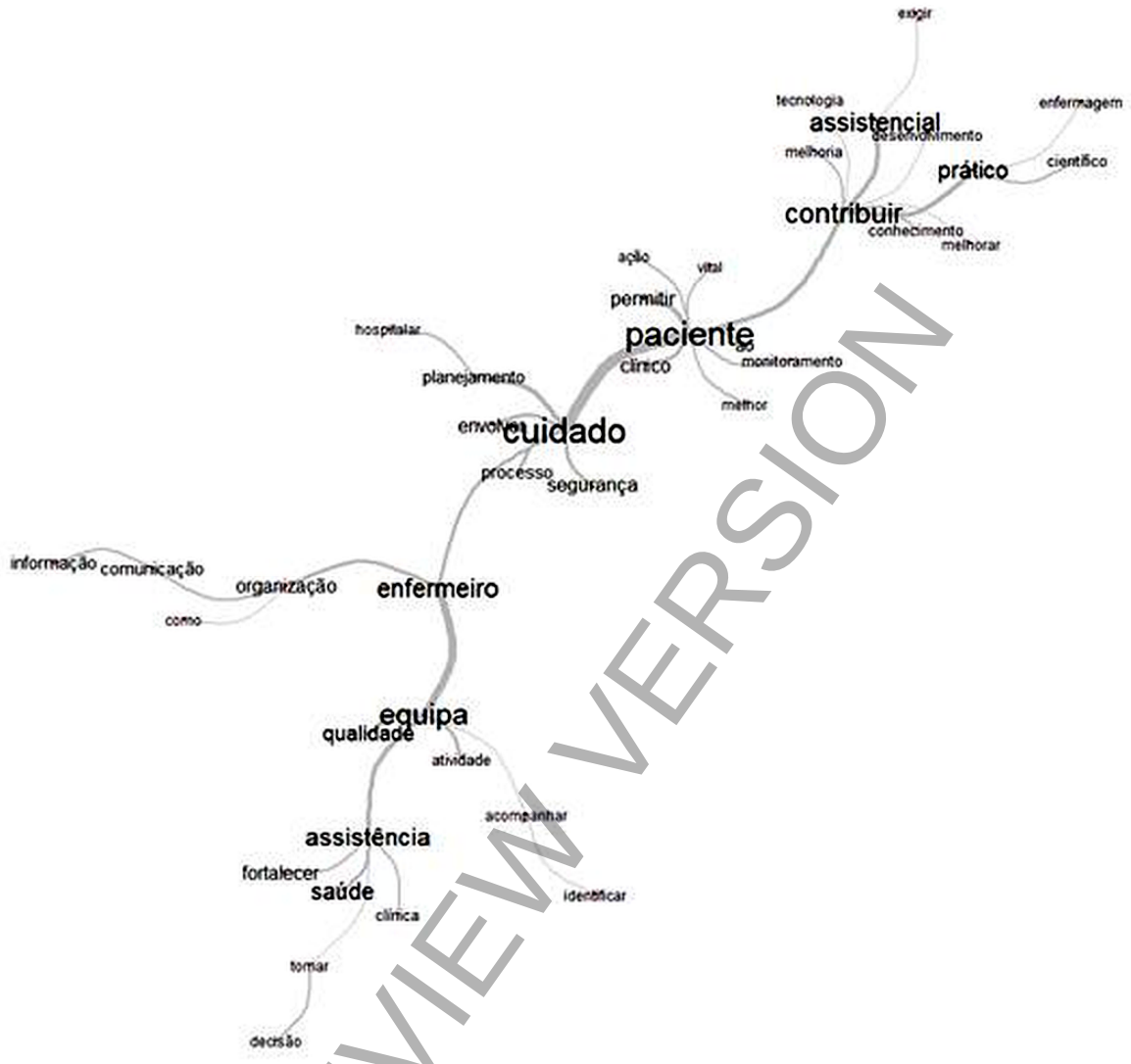

Figura 1

Análise de similitude representando as relações de coocorrência entre termos presentes no corpus textual analisado. Rio de Janeiro, Brasil, 2026.

Elaborado pela autora (2026).

A Classificação Hierárquica Descendente constitui uma das análises mais utilizadas no software IRAMUTEQ®, permitindo identificar classes lexicais formadas por segmentos de texto com vocabulário semelhante. Essa técnica utiliza procedimentos estatísticos para agrupar segmentos textuais de acordo com a distribuição do vocabulário, possibilitando a identificação de classes temáticas e estruturas discursivas presentes no corpus textual. A Figura 2 apresenta um exemplo de dendrograma gerado pelo software, evidenciando a divisão do corpus em classes lexicais.

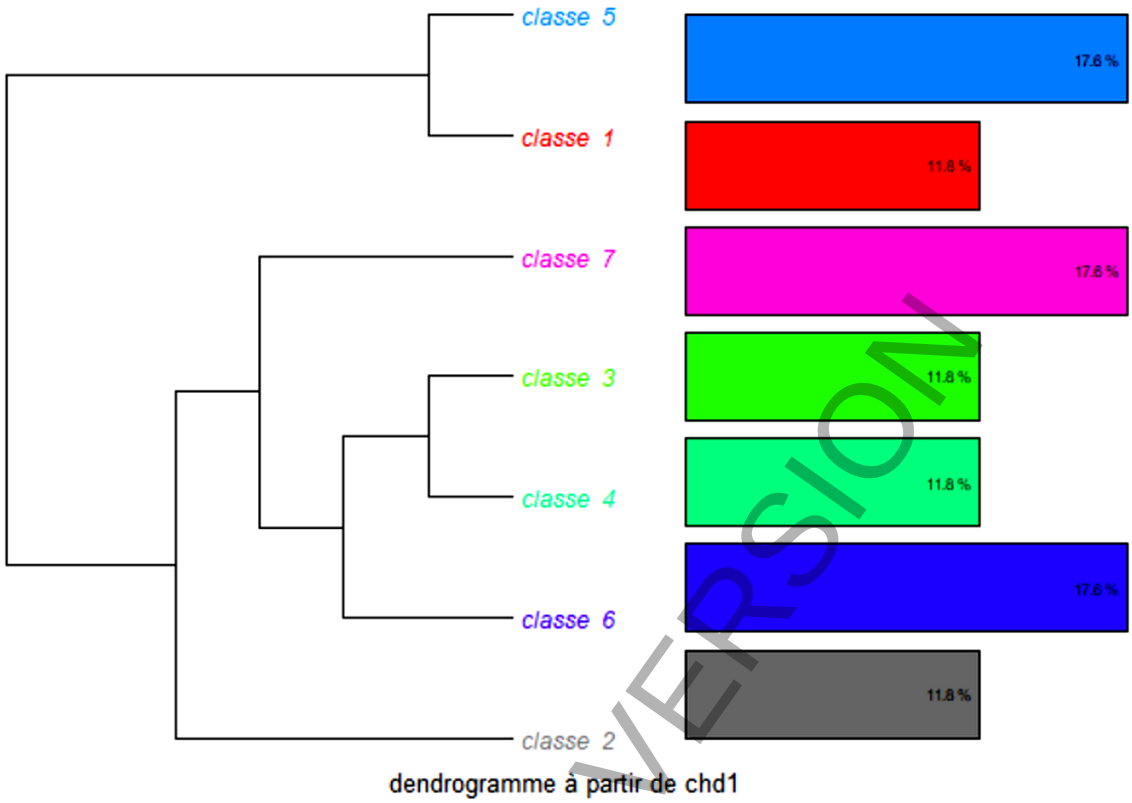

Figura 2

Dendrograma da Classificação Hierárquica Descendente (CHD) gerado pelo software IRAMUTEQ®, evidenciando as classes lexicais identificadas no corpus textual analisado. Rio de Janeiro, Brasil, 2026.

Elaborado pela autora (2026).

Para a realização da Classificação Hierárquica Descendente, o software considera as formas ativas do corpus, geralmente substantivos, adjetivos e verbos, e utiliza o valor do qui-quadrado como critério para associação das palavras às classes lexicais, sendo frequentemente adotado o valor de $\chi^2 \geq 3{,}84$ e $p < 0{,}05$ como parâmetro de significância estatística para associação das formas lexicais às classes, conforme pressupostos do método de Reinert.[12-16] Esse procedimento permite identificar as palavras mais representativas de cada classe lexical.

Complementarmente, a Análise Fatorial de Correspondência permite visualizar a distribuição das classes lexicais e das formas lexicais em um plano fatorial, possibilitando identificar aproximações e distanciamentos semânticos entre classes e categorias analisadas. Essa análise contribui para a interpretação das relações existentes entre as classes lexicais identificadas na Classificação Hierárquica Descendente. A Figura 3 apresenta um exemplo de plano fatorial obtido por meio da Análise Fatorial de Correspondência.

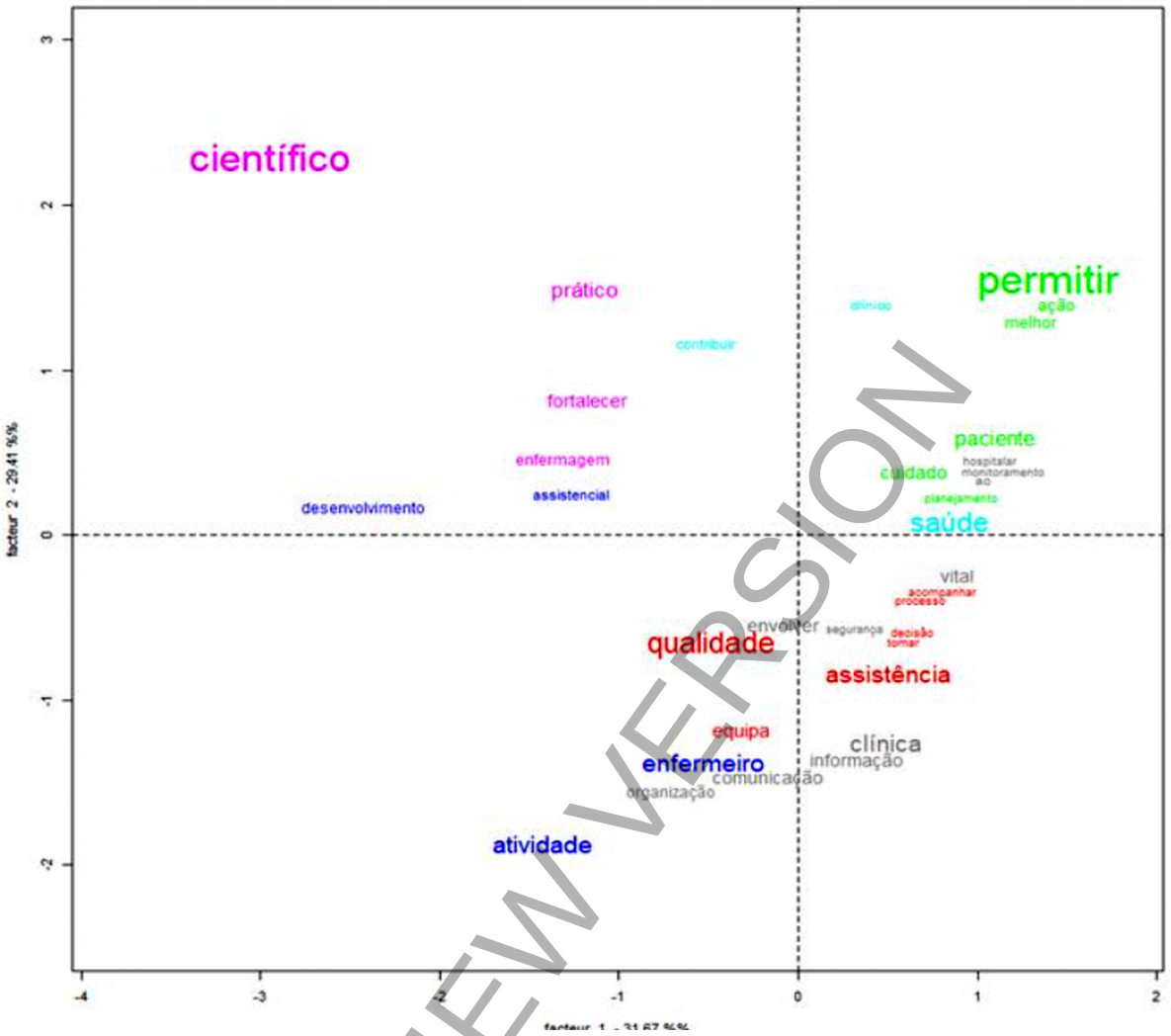

Figura 3

Plano fatorial obtido por meio da Análise Fatorial de Correspondência (AFC), representando a distribuição das formas lexicais no corpus textual analisado. Rio de Janeiro, Brasil, 2026.

Elaborado pela autora (2026).

O corpus textual utilizado neste exemplo foi composto por 40 segmentos de texto, totalizando 752 ocorrências e 256 formas lexicais, com média de 18,8 palavras por segmento textual. A análise resultou na retenção de 85% dos segmentos de texto na Classificação Hierárquica Descendente, indicando adequada estabilidade do corpus e consistência estatística para a formação das classes lexicais.

As representações gráficas apresentadas têm caráter ilustrativo e metodológico e têm como objetivo demonstrar as possibilidades analíticas do software IRAMUTEQ®, bem como evidenciar como a adequada preparação do corpus textual influencia a estabilidade das análises lexicométricas e a formação das classes lexicais. Dessa forma, as análises apresentadas não têm como finalidade a discussão temática do conteúdo do corpus, mas sim a demonstração dos procedimentos

analíticos e das possibilidades de interpretação oferecidas pelo software em pesquisas qualitativas.

Os resultados obtidos por meio das análises lexicométricas evidenciam que a organização adequada do corpus textual possibilita a obtenção de classes lexicais estáveis e semanticamente interpretáveis, reforçando que a preparação do corpus textual constitui uma etapa metodológica determinante para a qualidade das análises realizadas com o software IRAMUTEQ®.

DISCUSSÃO

Diferentemente de estudos que abordam a preparação do corpus sob a perspectiva operacional ou automatizada, o presente estudo propõe compreender essa etapa como componente estruturante do delineamento metodológico da pesquisa qualitativa apoiada por análise lexicométrica, influenciando diretamente a qualidade das análises e a reprodutibilidade científica.

Apesar da ampliação do uso de softwares de análise textual em pesquisas qualitativas, a literatura ainda apresenta número limitado de estudos que sistematizem, de forma detalhada, os procedimentos metodológicos relacionados à preparação do corpus textual. De modo geral, as publicações concentram-se na descrição das técnicas analíticas, como a Classificação Hierárquica Descendente, a análise de similitude e a Análise Fatorial de Correspondência, enquanto a etapa de organização prévia dos dados textuais permanece descrita de forma secundária ou pouco detalhada.[5-8] Essa lacuna metodológica pode comprometer a reprodutibilidade das análises e dificultar a formação de pesquisadores que utilizam análise textual assistida por software.

Os resultados deste estudo evidenciam que a qualidade das análises lexicométricas está diretamente relacionada à preparação adequada do corpus textual, especialmente no que se refere à padronização lexical, à segmentação dos textos, à definição de variáveis analíticas e à estruturação do material para processamento computacional. Esses elementos influenciam a distribuição do vocabulário nos segmentos de texto e, consequentemente, a formação das classes lexicais e a estabilidade estatística das análises realizadas pelo software.[5,13]

A principal contribuição deste estudo consiste na sistematização de recomendações metodológicas para a preparação do corpus textual, organizadas em etapas que incluem transcrição dos dados, revisão textual, padronização lexical, definição de variáveis analíticas, segmentação textual e estruturação do arquivo para processamento. Ao sistematizar essas etapas, o estudo contribui para a padronização de procedimentos metodológicos, para a transparência das análises e para a reprodutibilidade de pesquisas qualitativas que utilizam análise lexicométrica.

Além disso, a disponibilização do corpus textual em repositório de acesso aberto contribui para a transparência metodológica, para a rastreabilidade das análises e para a reprodutibilidade científica, aspectos cada vez mais valorizados na produção do conhecimento científico e nas diretrizes de ciência aberta.

Como limitação do estudo, destaca-se que o corpus textual utilizado para fins ilustrativos foi composto por material proveniente de uma área específica da saúde. Entretanto, as recomendações metodológicas sistematizadas podem ser aplicadas a diferentes contextos de pesquisa qualitativa que utilizem análise textual automatizada, uma vez que os princípios de organização do corpus textual são comuns a diferentes áreas do conhecimento.

Estudos recentes têm buscado sistematizar procedimentos relacionados à preparação de corpus textual para análise no software IRAMUTEQ®, com destaque para propostas de preparação automatizada do corpus textual.[14] Entretanto, observa-se que grande parte dessas abordagens concentra-se em aspectos operacionais e de automação do processo, enquanto o presente estudo propõe uma sistematização metodológica abrangente, contemplando etapas de transcrição, revisão textual, padronização lexical, definição de variáveis analíticas, segmentação textual e estruturação do corpus. Dessa forma, o estudo reforça a compreensão da preparação do corpus como etapa estruturante do delineamento metodológico em pesquisas qualitativas apoiadas por análise lexicométrica.

Nesse contexto, a preparação do corpus textual pode ser compreendida como uma etapa de pré-processamento de dados, fundamental em estudos de mineração de texto e análise estatística textual, influenciando diretamente a qualidade dos resultados analíticos obtidos.

CONCLUSÃO

Este estudo contribui para o campo metodológico das pesquisas qualitativas ao sistematizar recomendações para a preparação de corpus textual destinado à análise lexicométrica com o software IRAMUTEQ®, evidenciando que a organização do corpus constitui etapa metodológica central para a validade, confiabilidade e reprodutibilidade das análises.

Os resultados demonstram que a preparação do corpus influencia a estabilidade das classes lexicais, a retenção dos segmentos de texto e a consistência das análises estatísticas. A sistematização das etapas de preparação, incluindo transcrição, revisão textual, padronização lexical, definição de variáveis, segmentação textual e estruturação do arquivo, oferece subsídios metodológicos que contribuem para a padronização de procedimentos e a transparência das análises.

A preparação do corpus textual deve ser compreendida como etapa de pré-processamento de dados em pesquisas qualitativas que utilizam análise textual assistida por software, constituindo condição fundamental para a qualidade e a reprodutibilidade das análises lexicométricas.

DISPONIBILIDADE DOS DADOS

Os dados que suportam os achados deste estudo estão disponíveis em repositório de acesso aberto Zenodo, com o objetivo de garantir transparência, rastreabilidade e reprodutibilidade da pesquisa científica. O conjunto de dados pode ser acessado por meio do DOI: https://doi.org/10.5281/zenodo.19008125

CONTRIBUIÇÃO DA AUTORA

A autora foi responsável pela concepção e delineamento do estudo, organização do corpus textual, análise dos dados, interpretação dos resultados, redação do manuscrito, revisão crítica do conteúdo intelectual e aprovação final da versão a ser publicada.

CONFLITO DE INTERESSES

A autora declara não haver conflito de interesses.

FINANCIAMENTO

O presente estudo foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brasil.

AGRADECIMENTOS

A autora agradece às instituições de ensino e pesquisa que contribuíram para o desenvolvimento de sua formação científica e para a consolidação das reflexões metodológicas que fundamentam este estudo.

REFERÊNCIAS

1. Bardin L. Análise de conteúdo. Lisboa: Edições 70; 1977.

2. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. [Internet]. 2006 [cited 2026 Apr 10];3(2). Available from: https://doi.org/10.1191/1478088706qp063oa.

3. Gibbs GR. The analysis of qualitative data. London: SAGE Publications; 2012.

4. Miles MB, Huberman AM, Saldaña J. Qualitative data analysis: a methods sourcebook. 3rd ed. Thousand Oaks: SAGE Publications; 2014.

5. Camargo BV, Justo AM. IRAMUTEQ: um software gratuito para análise de dados textuais. Temas Psicol. [Internet]. 2013 [acesso em 10 de abril 2026];21(2). Disponível em: https://pepsic.bvsalud.org/.

6. Mazieri MR, Quoniam L, Reymond D, Cunha KCT. Uso do IRAMUTEQ para análise de conteúdo baseada em classificação hierárquica descendente e análise fatorial de correspondência. Braz J Mark. [Internet]. 2022 [acesso em 10 de abril 2026];21(2). Disponível em: https://www.researchgate.net/.

7. Rodríguez J, Reguant M, Ortega D. A practical case study of qualitative data analysis with IRAMUTEQ: lexicometric analysis of narratives of bisexual men and women. Rev Investig Educ. [Internet]. 2024 [cited 2026 Apr 10];42(2). Available from: https://doi.org/10.6018/rie.560411.

8. Chaves MMN, Santos APR, Santos NP, Larocca LM. Use of the software IRAMUTEQ in qualitative research: an experience report. In: Costa AP, Reis LP, Souza FN, Moreira A, editors. Computer supported qualitative research. Cham: Springer; 2017.

9. Soares SSS, Souza NVDO, Carvalho EC, Varella TCMML, Andrade KBS, Pereira SRM. Ensino do IRAMUTEQ para uso em pesquisas qualitativas segundo vídeos do YouTube: estudo exploratório-descritivo. Rev Esc Enferm USP. [Internet]. 2022 [acesso em 10 de abril de 2026];56:e20210407. Disponível em: https://www.scielo.br/.

10. Carvalho V, Silva EN, Sousa MS, Sampaio R, Barreto J. IRAMUTEQ analysis of Trastuzumab's public consultation in Brazil. Int J Technol Assess Health Care. [Internet]. 2018 [cited 2026 Apr 10];34(1). Available from: https://doi.org/10.1017/S0266462317004420 .

11. EQUATOR Network. Enhancing the quality and transparency of health research [Internet]. Oxford: EQUATOR Network; 2023 [cited 2026 Mar 23]. Available from: https://www.equator-network.org/.

12. Reinert M. Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. Cah Anal Données. 1983;8(2).

13. Guimarães CCV. Corpus metodológico de enfermagem para análise lexicométrica (CME-AL): exemplo de corpus textual para uso no software IRAMUTEQ [dataset]. Zenodo [Internet]. 2026 [acesso em 10 de abril 2026]. Disponível em: https://doi.org/10.5281/zenodo.19008125.

14. Barbosa RO, Oliveira Júnior J, Lugli LC, Peralta DA. Preparação automatizada de corpus textual para pesquisas qualitativas com IRAMUTEQ. Inf Inf. [Internet]. 2025 [acesso em 10 de abril 2026];30(3). Disponível em: https://doi.org/10.5433/1981-8920.2025v30n3p521

15. Lebart L, Salem A, Berry L. Exploring textual data. Dordrecht: Kluwer Academic Publishers; 1998.

16. Marchand P, Ratinaud P. L'analyse de similitude appliquée aux corpus textuels. In: Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles. Liège, Belgique; 2012.

Notas de autor

coelho.clarissa@unifesp.br

Información adicional

redalyc-journal-id: 5057

Clarissa Coelho Vieira Guimarães

Análise lexicométrica com IRaMuTeQ: recomendações metodológicas para preparação de corpus textual

Lexicometric analysis using IRaMuTeQ: methodological recommendations for textual corpus preparation

Análisis lexicométrico con IRaMuTeQ: recomendaciones metodológicas para la preparación del corpus textual

Revista de Pesquisa Cuidado é Fundamental Online
vol. 18, 14805, 2026
Universidade Federal do Estado do Rio de Janeiro, Brasil
rpcfo@unirio.br

ISSN-E: 2175-5361

DOI: https://doi.org/10.9789/2175-5361.rpcfo.v18.14805