Densidade Lexical de *abstracts* de aviação: um estudo comparativo

Lexical density of aviation abstracts: a comparative study

Fernanda Beatriz Caricari de Morais[1]
João Paulo Martins dos Santos[2]

Resumo
As densidades lexicais de um conjunto de *abstracts* provenientes dos trabalhos de conclusão de curso na Academia da Força Aérea Brasileira foram analisadas e comparadas com as densidades provenientes de *abstracts* de artigos do Air & Space Power Journal e Journal of Aviation/Aerospace Education and Research, revistas especializadas em publicações no contexto aeroespacial. A base de dados é composta pelos *abstracts* obtidos por Morais & Santos (2024) e pelos *abstracts* de artigos de trabalhos de conclusão de curso da AFA. A análise considera uma visão integrada dos pontos de vista estatístico e linguístico, tendo como suporte metodológico a Linguística Sistêmico-Funcional (HALLIDAY, 1994 e HALLIDAY & MATTHIESSEN, 2004, 2014), o conceito de densidade lexical de Ure (1971) e o de densidades por especialidade. Os resultados revelam uma diferença estatisticamente significativa na densidade lexical entre os *abstracts* de pesquisadores e cadetes com $z = 2,37, valorp = 0,0$, apoiando a hipótese de que aqueles escritos por pesquisadores apresentam um maior grau de concentração informacional. Isso se deve à necessidade de concisão da informação dentro de uma comunidade de discurso especializada, enquanto a menor densidade nos *abstracts* de cadetes provavelmente reflete um foco na clareza pedagógica e acessibilidade para um público mais amplo.

Palavras-chave: Densidade lexical. Linguagem acadêmica. Linguística Sistêmico-Funcional.

Abstract
The lexical densities of a set of abstracts from undergraduate thesis at the Brazilian Air Force Academy were evaluated and compared with those from the Air & Space Power Journal and the Journal of Aviation/Aerospace Education and Research, both specialized in publications within the aerospace field. The dataset consists of a set of articles obtained by Morais & Santos, (2024) and abstracts from graduation theses of cadets from the AFA. The analysis considers an integrated view of statistical and linguistic points of view, grounded in Systemic-Functional Linguistics (HALLIDAY, 1994 and HALLIDAY & MATTHIESSEN, 2004, 2014), the concept of lexical density as proposed by Ure (1971), and disciplinary variation in lexical density as well. The results indicate a statistically significant difference in lexical density between the abstracts written by researchers and those written by cadets, with a z-value of 2.37 and a p-value of 0.00. This supports the hypothesis that researcher-written abstracts have a higher level of informational concentration. This higher density is likely due to the need for conciseness within a specialized discourse community. In contrast, the lower lexical density in cadet abstracts may reflect an emphasis on pedagogical clarity and accessibility for a broader audience.

Keywords: Lexical density. Academic language. Systemic-Functional Linguistics.

1 Introdução

Este artigo é parte de um projeto de pesquisa em desenvolvimento em uma Instituição Militar de Ensino Superior, localizada no Estado de São Paulo, cujo objetivo é analisar as escolhas linguísticas

[1] Professora Adjunta IV da Academia da Força Aérea. Doutora em Linguística Aplicada e Estudos da Linguagem (PUC-SP). Pós-doutorado na UFU (PNPD/CAPES) e na PUC-SP (PDJ/CNPq). Orcid: https://orcid.org/0000-0001-6075-4101. E-mail: fernandafbcm@fab.mil.br.
[2] Professor Associado I da Academia da Força Aérea. Doutor em Ciências pela Escola de Engenharia de São Carlos - EESC-USP. Orcid: https://orcid.org/0000-0002-0957-7119. E-mail: joaopaulojpms1@fab.mil.br .

de artigos acadêmicos da área de aviação, escritos em Língua Inglesa, publicados em periódicos de notabilidade internacional.

Esse projeto de pesquisa está inserido na linha de pesquisa da Instituição “Poder Militar”, mais precisamente em “Estudos Linguísticos no Contexto Militar”. Fundamentado na Linguística Sistêmico-Funcional (HALLIDAY, 1994 e HALLIDAY & MATTHIESSEN, 2004, 2014), a pesquisa se beneficia da expertise do grupo SAL (Systemics Across Languages), que reúne pesquisadores de diversas partes do mundo, contribuindo com estudos sobre a descrição linguística, o ensino de línguas e o desenvolvimento de recursos pedagógicos focados nas necessidades dos aprendizes.

Além desse grupo internacional de linguística, esta pesquisa dialoga com uma iniciativa de desenvolvimento, o Grupo de Modelagem Matemática e Computacional (GMMC), que contribuiu para a constituição do corpus de pesquisa, formado por aproximadamente 800 artigos científicos escritos por pesquisadores do mundo todo, nativos e não-nativos na língua. O referido corpus possui 3 milhões de palavras, considerado de médio-grande porte (Berber-Sardinha, 2004), foi coletado na primeira etapa da pesquisa que concentrou-se na raspagem de dados dos periódicos *Air & Space Power Journal* e *Journal of Aviation/Aerospace Education and Research* (Morais & Santos, 2024).

Após a constituição do corpus de estudo, foi possível verificar as escolhas linguísticas mais frequentes feitas por pesquisadores da área de aviação por meio do tratamento dos dados pela ferramenta WordSmith Tools v.6 (SCOTT, 2018). A análise dos dados revelou uma alta frequência de adjuntos modais, especialmente nas seções de discussão. Essa tendência à construção de enunciados mais cautelosos, ao invés de afirmações categóricas, pode ser interpretada como uma estratégia para minimizar os riscos de generalização e garantir uma maior precisão nas afirmações. O que contribui com a argumentação do texto, com a clareza e com a objetividade, características valorizadas nos gêneros acadêmicos (Morais, 2023, 2024a, 2024b).

Tendo em mente as características da linguagem acadêmica que preza pela precisão e concisão das informações, observou-se na lista de palavras, gerada pelo WordSmith Tools v.6, a alta frequência do uso de substantivos abstratos e nominalizações (Morais & Barbara, 2018), que é um recurso importante para condensar informações em textos escritos de códigos elaborados (BERNSTEIN, 1990, 1996). Ao transformar verbos e adjetivos em substantivos, a nominalização permite tratar conceitos abstratos como entidades concretas, facilitando a análise, a classificação e a generalização. Essa transformação, que Halliday (2004) denomina metáfora gramatical, confere ao discurso uma estrutura mais lógica e objetiva, deixando a linguagem mais científica.

Essa transformação não se limita ao nível lexical, a uma mera alteração de classe gramatical. O fenômeno linguístico da nominalização, que consiste na substituição de um verbo como "to develop" por um substantivo como "the development," por exemplo, impacta diretamente na densidade do texto e na dificuldade de leitura para aprendizes de língua estrangeira (LE), afetando, portanto, a leiturabilidade (traduzido do inglês "readability") (SINAR et al., 2024).

Esse fenômeno torna o texto mais difícil para aprendizes de LE porque a ação se torna abstrata e menos direta. Isso aumenta a densidade do texto, pois mais informações são "compactadas" em poucas palavras, exigindo maior esforço cognitivo para decifrar.

Basicamente, a densidade lexical, ao quantificar a proporção de palavras lexicais em um texto, revela a sua complexidade semântica e sintática. Em artigos acadêmicos, por exemplo, a alta densidade lexical é frequentemente associada a um maior grau de abstração e formalidade, exigindo do aprendiz um esforço cognitivo mais elevado para a compreensão.

Saragih (2006, p. 9), baseado em Ure (1971), explica que a densidade lexical é uma métrica linguística que quantifica a proporção de palavras lexicais (substantivos, verbos, adjetivos e advérbios) em relação ao total de palavras em uma oração. Essa medida pode ser calculada como a razão entre o número de palavras de conteúdo e o número total de palavras em um texto ou sentença (EGGINS, 2004, p. 97). Um texto caracterizado por uma elevada densidade lexical é frequentemente associado ao registro escrito.

Para viabilizar a análise, os *abstracts* dos periódicos foram separados para constituir o corpus de *abstracts* de pesquisadores experientes (denominado corpus 1) e um novo corpus foi coletado: o de abstracts dos trabalhos de conclusão (TCCs) dos cadetes aviadores da Instituição (denominado corpus 2).

A escolha pelo gênero textual se deu pela facilidade de coleta, visto que os TCCs dos cadetes formados estão disponíveis na biblioteca da Instituição e, também, pela possibilidade de se extrair uma amostra do corpus da pesquisa, composto por artigos acadêmicos que contêm *abstracts*, item obrigatório das publicações. Como hipótese inicial, é esperado que o corpus 1 de textos de pesquisadores experientes seja mais denso lexicalmente devido às experiências linguísticas prévias desses autores. Muitos desses têm vasta experiência acadêmica e escrevem, com frequência, gêneros acadêmicos para periódicos e eventos da área. Enquanto os autores dos textos corpus 2 estão realizando a primeira experiência com o gênero artigo acadêmico, tendo contato na graduação com outros gêneros (resumo, resenha, fichamento, entre outros), porém não com a complexidade lexical de um artigo.

As análises apresentadas neste artigo consideram ambos os pontos de vista, estatístico e linguístico. A primeira é restrita às estatísticas descritivas, gráficos e teste de hipótese para as densidades lexicais. Por sua vez, a segunda procura elucidar escolhas feitas nos corpora, procurando entendê-las com base na teoria Sistêmico-Funcional (HALLIDAY, 1994 e HALLIDAY & MATTHIESSEN, 2004, 2014), refletindo sobre questões de ensino de LE, especialmente o ensino de leitura e escrita em contextos acadêmicos para contribuir com o desenvolvimento de recursos pedagógicos adaptados às necessidades dos aprendizes. Dessa forma, as análises são complementares, pois uma fornece os dados numéricos, enquanto a segunda a interpretação linguística.

A seguir, conceitos essenciais sobre a densidade lexical são discutidos para, em seguida, discorrermos sobre as suas implicações para o ensino de línguas. Posteriormente, os procedimentos de coleta e tratamento dos dados são descritos, assim como os recursos utilizados para o cálculo da densidade e de outras características linguísticas dos corpora, seguidos das análises dos resultados obtidos. Por fim, algumas considerações são tecidas com base nos achados, ventilando sobre as suas implicações para o ensino de línguas.

2 Densidade lexical

Na pesquisa científica, as ferramentas de busca desempenham um papel fundamental, permitindo a localização de artigos a partir de palavras-chave relevantes. Diante da vasta quantidade de publicações disponíveis na academia, os pesquisadores frequentemente iniciam sua análise lendo o *abstract* — um gênero textual conciso e estruturado, que, em grande parte dos casos, apresenta objetivos, metodologia e os principais resultados do artigo. O *abstract* é um gênero específico, geralmente escrito em inglês, sendo acompanhado de um resumo na língua de seu local de circulação. Fornece uma visão geral e rápida do estudo, incluindo o objetivo, a metodologia, os principais resultados e conclusões. Sua leitura permite avaliar a pertinência do artigo conforme interesse acadêmico-investigativo do leitor, de forma eficiente, economizando tempo e direcionando os esforços para os trabalhos mais relevantes.

Para produzir um abstract bem estruturado, é requerido dos cadetes proficiência em organizar ideias e fatos dentro de sua escrita, bem como uma compreensão abrangente das características da linguagem acadêmica. Halliday e Matthiessen (2004, 2014) postulam que a complexidade da língua escrita muitas vezes decorre de sua densidade lexical, caracterizada pela eficiente forma de "empacotar" (em inglês *package*) numerosos itens lexicais em uma oração (HALLIDAY, 2004). Em

geral, os textos escritos apresentam maior densidade lexical em comparação com os textos falados devido à sua maior proporção de palavras de conteúdo (STUBBS, 2002; JOHANSSON, 2008).

Intimamente entrelaçada com o conceito de concisão da informação, a densidade lexical é influenciada principalmente pela prevalência das palavras de conteúdo dentro de um texto. Consequentemente, os textos com maior proporção de palavras de conteúdo (substantivos, adjetivos, verbos e advérbios) são considerados mais densos, pois transmitem mais informações por unidade de texto do que aqueles dominados por palavras-função, que conectam as palavras de conteúdo, como preposições, artigos, conjunções e pronomes (JOHANSSON, 2009).

Dada a natureza informacional e lexicalmente densa do discurso acadêmico, pode-se inferir que os textos acadêmicos de alta qualidade escritos pelos cadetes podem exibir uma porcentagem correspondentemente elevada de palavras de conteúdo. Uma maior densidade lexical sugere a capacidade do aprendiz de LE de empregar a linguagem de forma mais sofisticada, um sinal de maior domínio e de habilidades avançadas de escrita.

A densidade lexical pode, assim, servir como um critério valioso para avaliar a proficiência linguística dos cadetes, pois pode sinalizar um comando robusto do vocabulário e da estrutura da língua escrita, ambos indispensáveis para uma escrita acadêmica eficaz.

Eggins (2004, p. 97) afirma que a densidade lexical pode ser quantificada calculando-se a proporção de palavras de conteúdo para o total de palavras dentro de um texto. Este cálculo, foi originalmente proposto por Ure (1971), que sugere que a densidade lexical deve ser interpretada como a proporção de palavras lexicais em relação ao número total de palavras do texto.

Para Ure (1971), a densidade lexical (DL), do tipo Tipo-Token (DTT), é uma medida da riqueza vocabular de um texto, ou seja, a quantidade de palavras lexicais (substantivos, verbos, adjetivos, advérbios) em relação ao total de palavras. Pode ser expressa pela razão entre a frequência de palavras lexicais e o total de palavras, ou seja, $DL = DTT = \frac{PL}{TP} \cdot 100$, em que PLdenota Palavras Lexicais e TP denota Total de Palavras.

De acordo com esse autor, a densidade lexical é um indicador da complexidade e sofisticação do texto, bem como da variedade de ideias e informações apresentadas. Um texto com alta densidade lexical pode ser mais difícil de se ler e entender, especialmente para leitores com um vocabulário limitado. Um texto com densidade lexical baixa, porém, pode ser considerado pobre em conteúdo e pouco informativo (HALLIDAY & MATTHISSEN, 2014, p. 726-727).

Ure (1971) ainda explora a relação entre densidade lexical e registro, mostrando que diferentes tipos de texto tendem a ter diferentes níveis de densidade lexical. Dessa forma, é importante encontrar um equilíbrio entre a riqueza vocabular no sentido da DLe a clareza do texto, de acordo com o público-alvo e o objetivo do texto.

A maioria dos textos falados apresenta uma densidade lexical abaixo de 40%, enquanto uma parcela substancial de textos escritos demonstra uma densidade lexical de aproximadamente 40% ou mais, tendo uma maior densidade em relação aos textos falados. Além disso, um texto com uma densidade lexical notavelmente alta provavelmente conterá mais informações e, consequentemente, mais informações "empacotadas", escritas de forma concisa e que, potencialmente, influenciam a compreensão dos leitores de LE.

Sabe-se que existem diversas propostas para o cálculo da densidade lexical, como a de Halliday e Matthiessen (2014, p. 727), que a mensuram pela frequência média de palavras lexicais por oração, excluindo-se as orações encaixadas. No entanto, no presente estudo se concentra na abordagem clássica proposta por Ure (1971) e nas densidades específicas (densidades por especialidade, isto é, nominal, verbal, adjetival e adverbial), uma vez que essas medidas se mostram suficientes para analisar a complexidades linguísticas dos *abstracts* produzidos por pesquisadores experientes da área da aviação e compará-los com aqueles elaborados por pesquisadores em formação (cadetes) para lançar luzes sobre o desenvolvimento da escrita em LE.

3 A densidade lexical no ensino de línguas

O conhecimento lexical de um aprendiz de língua estrangeira é um processo que envolve a extensão, ou seja, a quantidade de palavras da língua alvo conhecida pelo aprendiz, profundidade, que se refere à variedade do vocabulário, e a fluência, que é a velocidade de produção e compreensão de palavras. Mais especificamente sobre o conhecimento lexical, pode-se afirmar que este é o responsável pelo processo de comunicação. Para Biderman (1998), a referência à realidade extralinguística nos discursos humanos é feita por meio dos signos linguísticos, ou unidades lexicais, que designam os elementos desse universo segundo o recorte feito pela língua e pela cultura correlatas. Para a autora, "o léxico é o lugar da estocagem da significação e dos conteúdos significantes da linguagem humana" (BIDERMAN, 1998, p. 73).

É esperado que a densidade lexical aumente conforme os anos escolares e os níveis de ensino, sendo indicativa da crescente complexidade textual e do aprofundamento dos conhecimentos

específicos de cada área. Pensando no ensino-aprendizagem de LE, tem-se no Quadro Comum Europeu (QECR) a divisão de níveis e como eles diferem no léxico, com diferenças graduais e ascendentes entre os níveis.

Por exemplo, no nível A1 é esperado um vocabulário fundamental para a compreensão de textos simples e diretos, enquanto no nível B1 o aprendiz já consegue lidar com textos mais complexos, com maior variedade de vocabulário e estruturas gramaticais. Dessa forma, nos últimos níveis, o aprendiz é capaz de compreender e produzir textos altamente especializados, com densidade lexical elevada e uso preciso de termos técnicos.

Stubbs (1996), em consonância com Ure (1971), também corrobora a definição de densidade como a razão entre a frequência de palavras lexicais e a frequência total de palavras, expressa em valores percentuais. A densidade lexical é considerada um dos principais indicadores de desenvolvimento para uma escrita mais acadêmica, pois mostra a habilidade que o escritor tem de fazer proposições de forma condensada (COLOMBI, 2002, SCHLEPPEGRELL, 2004). Essa densidade se manifesta por meio de diferentes modos de incorporação do léxico à estrutura gramatical da frase, evidenciando como a escolha de palavras influencia a compressão do conteúdo.

Para ilustrar, um texto informativo tende a ser mais denso porque seus itens lexicais são os principais veículos de conteúdo referencial, carregando a maior parte da informação. Em contraste, um texto narrativo geralmente apresenta menor densidade, sendo mais dêitico e, portanto, mais próximo da linguagem falada, que prioriza a progressão da história sobre a compactação conceitual.

Em relação à linguagem acadêmica, sabe-se que ela é densa, complexa e abstrata. Schleppegrell (2004, 2012) discute a dificuldade de alunos graduandos de entender e produzir textos utilizando níveis de abstração mais complexos, isto é, utilizando vocabulário técnico e nominalizações. Devido a essa complexidade, há uma propensão de não entendimento completo (ISMAIL *et al*., 2023).

A densidade está relacionada também ao conteúdo do léxico, sendo considerados linguisticamente conectados. Dependendo do propósito, o léxico pode ser dividido em: termos de conteúdo e palavras de funções. Os termos têm significado e referente, enquanto as palavras de função têm função gramatical (THORNBURY & SLADE, 2006).

Na perspectiva do léxico, a LSF é uma teoria e um método de análise efetivo para analisar a léxico-gramática e, principalmente, a relação entre densidade lexical e complexidade gramatical, que são as principais características da chamada complexidade linguística.

Para Halliday & Matthiessen (2004, 2014), a complexidade linguística de amostras escritas é definida pela complexidade lexical, enquanto a complexidade das amostras orais é medida pela

complexidade gramatical. Nichols (2009) postula que, no discurso escrito, ao contrário da oralidade que contém muitas orações, há um empacotamento de informações, ou seja, as orações são mais enxutas, pois estão compactadas densamente.

Essencialmente, a densidade lexical mede o quanto um texto é informativo e compreensível. Como abordado anteriormente são itens lexicais os substantivos, verbos, adjetivos e advérbios, enquanto os pronomes, determinantes, verbos finitos e algumas classes de advérbios são itens gramaticais. Assim, a densidade lexical pode ser medida pela proporção do total de itens lexicais pelo total de palavras. Abaixo um exemplo de densidade retirado do corpus de TCC desta pesquisa:

Body language is *one of the* ***oldest forms*** *of* ***communication*** *on the* ***planet****...*

No exemplo, há um total de 13 palavras e itens lexicais, com densidade de 53,8%. Observa-se que nesse trecho, primeiro período do texto, o cadete inicia a oração apresentando o conceito que pesquisa. Autores como Bhatia (1993) e Swales (1990) já consideravam orações como a exemplificada acima como prototípica do movimento "apresentar o propósito" em *abstracts*, sendo parte da estratégia nomeada "indicando o escopo da pesquisa".

É consenso nos estudos linguísticos que a língua escrita é mais explanatória e contém naturalmente mais informação, termos lexicais e palavras para detalhar um conceito ou um objeto. Como abordado anteriormente, o *abstract* de artigo de pesquisa tem propósito comunicativo bem definido e compreendido independentemente da disciplina a que pertence (BHATIA, 1993, p. 147). O American National Standards Institute (2023) define *abstract* como "um resumo conciso e preciso de um documento que oferece uma visão geral rápida para o leitor. Ajuda os leitores a entender o conteúdo principal do documento sem que precisem lê-lo na íntegra."[3]. Essa característica de concisão e representatividade é crucial no ambiente acadêmico, especialmente considerando o status da Língua Inglesa (LI) como língua franca (CRYSTAL, 2006) e sua preferência pela comunidade científica. Compreender as densidades dos *abstracts* exige uma abordagem sistemática. A fim de conduzir a investigação dos corpora, a próxima seção descreve a metodologia aplicada neste estudo.

4 Metodologia

A metodologia adotada neste artigo foi estruturada para organizar de forma sistemática a seleção, o processamento e a análise dos corpora compostos pelos *abstracts* de artigos científicos. A Figura 1 fornece uma visão geral dos elementos estruturais desenvolvidos.

[3] No original: "It is a concise, accurate summary of a document that provides a quick overview for the reader. It helps readers understand the core content of the document without having to read it in full".

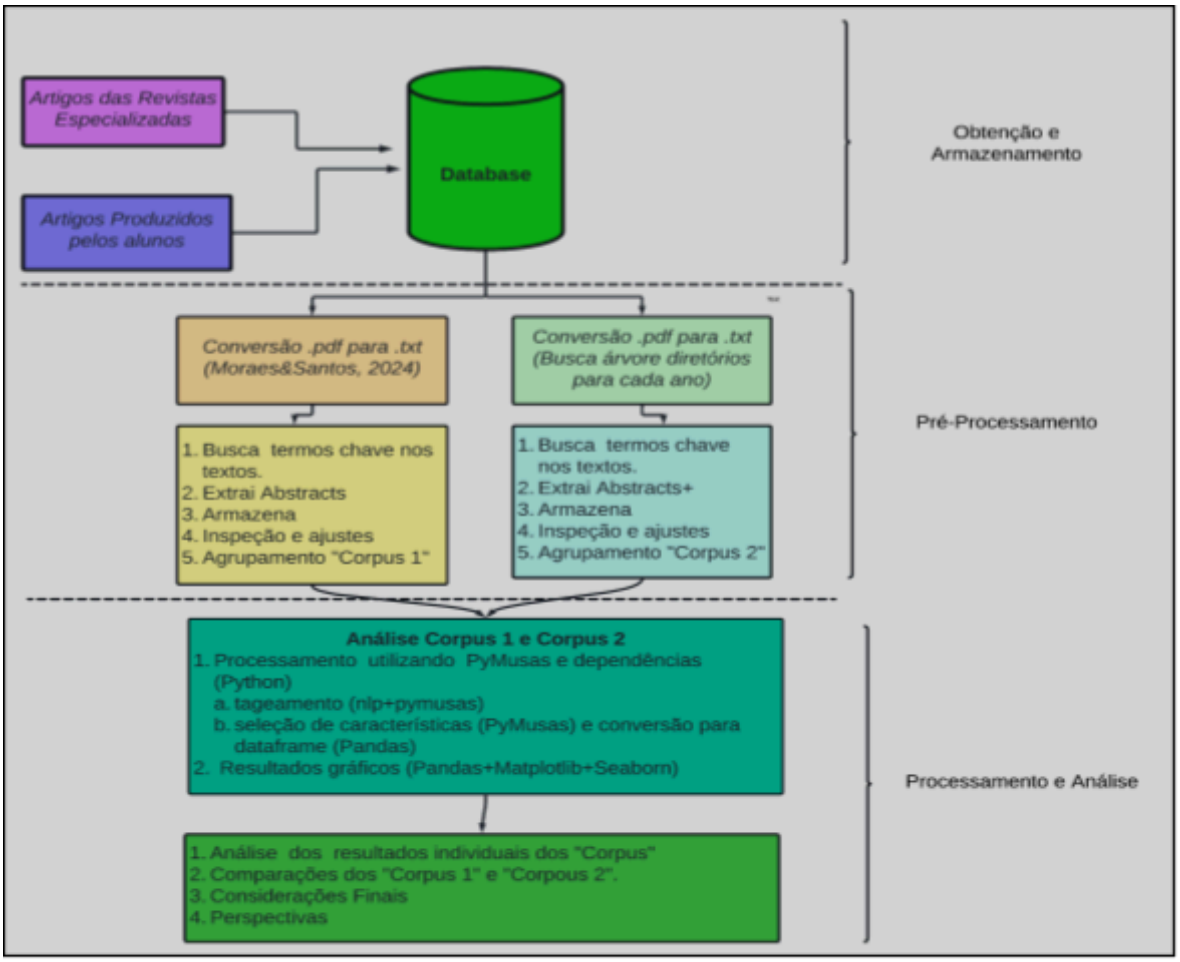

Figura 1: Processo geral de obtenção, pré-processamento, processamento e análise dos corpora.
Fonte: Os autores, 2024

O delineamento metodológico apontado na Figura 1 contempla as etapas de coleta dos textos completos, extração dos *abstracts*, etiquetagem linguística e implementação computacional das rotinas de análise e visualização. Os pormenores das três fases distintas da metodologia, com respectivos pormenores podem ser colocados como segue:

1. Obtenção e armazenamento: fase de busca de artigos, motivada pelos resultados de Morais & Santos (2024);
2. Pré-processamento: fase de extração dos textos a partir do arquivo em *.pdf* e à definição da estrutura de subdiretórios responsável pelo armazenamento dos arquivos de interesse;

3. Processamento e Análise: utilização de bibliotecas computacionais em *Python*™ para etiquetagem dos corpora, extração das características linguísticas relevantes, conversão dos dados para *data frames* e geração dos resultados.

4. 1 Obtenção e armazenamento

Os textos completos dos artigos são provenientes de duas fontes distintas e não correlacionadas: abstracts de artigos das revistas *Air & Space Power Journal e Journal of Aviation/Aerospace Education and Research* e *abstracts* de trabalhos de conclusão de curso de cadetes graduados pela Academia da Força Aérea nos anos de 2021, 2022 e 2023. O primeiro é baseado em Morais & Santos (2024). Ambos foram armazenados para a composição da base de dados inicial. É importante lembrar que os artigos de *Aviation/Aerospace Education and Research* foram obtidos até a data de outubro de 2024. Enquanto o segundo conjunto de dados foi obtido na biblioteca da Instituição.

4. 2 Pré-processamento

Conforme apontado em Figura 1, o pré-processamento consistiu em duas etapas: a conversão dos textos completos em arquivo de texto (*.txt*) e detecção e captura dos *abstracts*. Os textos foram obtidos em formato *Portable Document Format* (*.pdf*) e demandaram a utilização da biblioteca pdfminer.six (PDFMINER.SIX, 2023) para viabilizar a extração do conteúdo textual integral. Ainda durante a conversão, devido às diferenças na organização dos diretórios componentes da base de dados (*Database*), foi necessário mapear os subdiretórios com arquivos tipo *.pdf* para permitir a automatização da leitura e conversão para *.txt*.

Após a obtenção dos textos completos em formato *.txt*, armazenados em diretórios organizados de acordo com a respectiva origem, a etapa de extração dos *abstracts* foi iniciada. Este processo foi precedido por uma análise exploratória individual de cada conjunto textual, com o objetivo de identificar padrões recorrentes que possibilitassem o desenvolvimento de um código (*script)* computacional em Python para a localização automatizada dos trechos de interesse.

No caso dos *abstracts* componentes do corpus 1, a análise direcionou a criação de uma função computacional para a extração dos trechos de interesse que englobasse diferentes formatações estruturais. Por exemplo, alguns artigos com a palavra "*Abstract*" e "*keywords*" delimitando o texto de interesse; outros com a palavra "*Abstract*" e "*Introduction*" delimitando o texto de interesse. Há, ainda,

alguns artigos, os quais não possuem o termo "*Introduction*", mas o texto de interesse é composto de único parágrafo; por fim, alguns artigos possuem um *abstract* que é composto de mais de um parágrafo.

Os elementos-chave identificados previamente foram integrados a um serviço de processamento textual (*Text Processing Services*) por meio de expressões regulares (*Regular Expression Operations – ReGex*) foi utilizado[4]. Por fim, após armazenamento dos *abstracts* em arquivo, foi realizada uma inspeção visual com seleção manual dos *abstracts* sem os títulos em inglês e as *keywords*.

Por sua vez, a extração dos *abstracts* componentes do corpus 2 foi realizada em duas etapas sequenciais após a análise. A primeira consistiu na seleção dos trechos de textos do início da primeira página dos textos completos até a palavra *keywords*[5], e composição de arquivo do tipo *.txt* único para cada ano. Cada arquivo é, portanto, composto de trechos de textos que englobam título, resumo, título em inglês e resumo em língua inglesa (*abstract*). A segunda parte, executada em sequência, foi uma inspeção visual e seleção manual dos trechos de interesse, os *abstracts*, sem os títulos em inglês e sem as "*keywords*". A adoção de tal metodologia para obtenção desses *abstracts* foi devido à variabilidade na formação observada e a quantidade razoavelmente pequena de textos a serem inspecionados.

4. 2 Etiquetagem dos Corpora

A etiquetagem foi baseada na biblioteca PyMusas, acrônimo de *Python Multilingual Ucrel Semantic Analysis System*[6] (PyMUSAS DEV TEAM, 2024), desenvolvida em linguagem Python, especializada em etiquetagem de textos em diferentes línguas conforme o sistema Ucrel Semantic Analysis System (USAS).

O USAS[7] é um sistema desenvolvido pelo UCREL (University Centre for Computer Corpus Research on Language) da Lancaster University, utilizado para análise semântica automática de texto (PIAO *et al.*, 2015; PIAO *et al.*, 2016).

---

[4] Mais informações em (PYTHON SOFTWARE FOUNDATION, 2025).

[5] keywords é um item obrigatório dos *abstracts deste conjunto.*

[6] Python Multilingual Ucrel Semantic Analysis System, is a rule based token and Multi Word Expression (MWE) semantic tagger. The tagger can support any semantic tagset, however the tagset we have concentrated on and released pre-configured spaCy components for is the Ucrel Semantic Analysis System (USAS).

[7] Mais informações e referências sobre o USAS podem ser encontradas em (LANCASTER UNIVERSITY, 2025).https://ucrel.lancs.ac.uk/usas/. Acesso em: 5 fev. 2025.

4.3.1 Aspectos Computacionais da Análise Estatística

Tendo em vista a importância da densidade lexical como indicador de complexidade linguística, decidiu-se analisar a densidade lexical dos textos de pesquisadores experientes, publicados nos periódicos mencionados, e compará-la com a densidade de textos dos cadetes, tidos como pesquisadores em formação. Para tal, recorreu-se à inferência estatística, que se mostra apropriada à tarefa de generalização a partir de dados amostrais. Segundo Devore (2006, p. 221), "a inferência estatística é quase sempre direcionada à obtenção de algum tipo de conclusão sobre um ou mais parâmetros (características da população, por exemplo)". Assim, a análise foi conduzida com base em valores extraídos de amostras representativas de cada grupo, permitindo a formulação de conclusões sob hipóteses bem definidas.

O primeiro passo foi obter uma estimativa dos parâmetros de densidade lexical de cada caso. Essas estimativas, denominadas estatísticas, são valores denominados estimativas pontuais, pois os textos em questão são considerados amostras significativas dos textos de pesquisadores experientes e pesquisadores em formação. Estas estimativas estão sujeitas à variabilidade inerentes ao acaso, ou seja, amostras diferentes (considerando, por exemplo, os *abstracts* de outros *journals* especializados em aviação) irão resultar em diferentes estimativas pontuais para a densidade lexical dos pesquisadores experientes. Da mesma forma, uma estimativa pontual distinta seria obtida ao incluir novas produções dos pesquisadores em formação. É nesse contexto que se empregou a inferência estatística, a qual permite obter as probabilidades sob determinadas hipóteses (em geral, denominada hipótese nula) para julgar se o resultado observado deve ser atribuído ao acaso ou derivam de diferentes populações, no caso textos escritos por pesquisadores experientes e textos escritos por pesquisadores em formação.

Um exemplo clássico é o seguinte: suponha o lançamento de uma moeda 100 vezes com a observação de 60 caras e 40 coroas. Sob a hipótese de que a moeda é honesta (em geral denominada hipótese nula H_0: a probabilidade de obter cara é $p_{cara} = 0.5$)[8], a probabilidade de observar um evento tão extremo (60 caras ou mais em 100 lançamentos (teste unilateral) é, para efeito de cálculo, nula ($p = P(X \geq 60) = 0.0284$). Neste caso, como tal probabilidade p (valor p) é um evento muito raro, rejeitamos a hipótese nula H_0 de que cara e coroa sejam igualmente prováveis, em

[8] os autores adotaram a notação internacional para representação de números decimais.

detrimento de uma hipótese alternativa, em geral denominada H_1: a probabilidade de obter cara é maior que a probabilidade de obter coroa ($p_{cara} \geq 0.5$). Por outro lado, um experimento que resulte em 55 caras e 45 coroas, ocasiona ($p = P(X \geq 55) = 0.184$) e, portanto, não há evidências para rejeitar a hipótese de que a moeda é honesta. O fator de decisão, denominado nível de significância $\alpha = 5\%$ é, em geral, adotado em análises estatísticas. Se a probabilidade excede α, conclui-se pela não rejeição da hipótese nula e pela rejeição em caso contrário.

Do ponto de vista da análise da densidade lexical, a não rejeição da hipótese nula sugere que não há evidências estatísticas suficientes, com base nos dados analisados, para afirmar que as diferenças observadas entre os grupos devem ser atribuídas a distintas competências de escrita. Em contrapartida, a rejeição da hipótese nula indica que há evidências de que pesquisadores experientes e pesquisadores em formação diferem significativamente quanto à densidade lexical em seus textos. Por fim, cabe destacar que, em qualquer teste de hipótese, existe o risco inerente de erro — seja do tipo I (rejeitar uma hipótese verdadeira), seja do tipo II (não rejeitar uma hipótese falsa). Informações adicionais sobre testes de hipóteses podem ser encontradas em Devore (2006), Montgomery e Runger (2012) e Larson e Farber (2015).

Neste artigo a comparação das densidades lexicais foi realizada por meio do teste de hipóteses denominado teste de proporção ou teste z para proporção. O teste estatístico de proporção é, portanto, uma ferramenta utilizada para comparar as proporções de duas ou mais amostras e avalia se há diferença significativa entre a proporção de palavras lexicais nos textos de pesquisadores experientes (corpus 1) e nos textos dos cadetes (corpus 2). Caso a probabilidade calculada (valor p) pelo teste resulte em valor menor que $\alpha = 5\%$, conclui-se pela rejeição da hipótese de que pesquisadores experientes e em formação escrevem com mesma densidade lexical e, portanto, sugere que há evidências em favor da hipótese alternativa de que pesquisadores experientes escrevem com densidade lexical maior. Caso contrário, conclui-se pela não rejeição da hipótese nula, ou seja, o teste não aponta evidências significativas para a rejeição da hipótese nula.

Para fins de análise da magnitude do efeito das diferenças observadas, considerou-se a razão de risco $RR = \frac{p_A}{p_B}$, e as chances de risco (*odd's ratio*) $CR = (\frac{p_A}{(}1 - p_A)\frac{)}{(}\frac{p_B}{(}1 - p_B))$, em que p_A, p_B são as estimativas pontuais das probabilidades.

4.3.2 **Aspectos Computacionais da Análise Estatística e Visualização**

As manipulações das informações textuais e visualizações gráficas foram realizadas por meio da linguagem Python utilizando as bibliotecas Pandas (McKINNEY, 2020), Matplotlib (HUNTER, 2007) e Seaborn (WASKOM, 2021).

Para calcular a densidade lexical, primeiramente, a etiquetagem morfossintática (POS tagging) do corpus foi realizada. Em seguida, o corpus etiquetado foi convertido em uma tabela de dados (dataframe). Dessa forma, a DL pôde ser calculada como uma composição das classes de POS (em inglês, *Part of Speech*) que apresentam cargas semânticas Substantivos ('NOUN'), Adjetivos ('ADJ'), Verbos ('VERB') e Advérbios ('ADV'). Nesse contexto, considerou-se que PN, PV, PAdv, PAdj, TP fossem os números de Substantivos, Adjetivos, Verbos, Advérbios e Total de palavras no texto. Então, a densidade lexical pôde ser reescrita como a soma das densidades por especialidades $DL = (PN + PV + PAdv + PAdj\frac{)}{TP}$. Sem perda de generalidade, o valor 100 em equação foi suprimido para obter um valor no intervalo $[0, 1]$, o qual é, imediatamente, relacionado à probabilidade.

Para as análises estatísticas, o módulo *Python Stats Models* (SEABOLD; PERKTOLD, 2010) foi foi utilizado[9]. Os conceitos elementares envolvendo estatísticas descritivas ou testes de hipóteses para diferenças entre as proporções podem ser encontrados em Montgomery (2012) ou em Larson e Farber (2015)[10].

5 Resultados e discussões

A seção apresenta uma análise dos resultados provenientes das comparações dos corpora por meio de uma tabela de frequências das classes com carga semântica. Em sequência, um gráfico com as densidades por especialidade (substantivo, verbo, adjetivo e advérbio) são apresentadas. Por fim, a comparação das proporções das densidades lexicais dos corpora via teste z e o gráfico das entidades nomeadas.

Corpora	DV	DAdj	DAdv	DN	DL
Corpus 1	0.12253	0.08848	0.02263	0.32287	0.55650
Corpus 2	0.10581	0.09005	0.02324	0.25757	0.47666

Tabela 01: Densidades por especialidades e lexical: Densidade Verbal (DV), Densidade Adjetival (DAdje), Densidade Adverbial (DAdve), Densidade Nominal (DN) e Densidade Lexical (DL).
Fonte: Os Autores

[9] Mais informações em https://www.statsmodels.org/stable/index.html. Acesso em: 07 de novembro de 2024.
[10] z teste para proporções.

É importante ter em mente que o Corpus 1 (22.232 *tokens*), composto de *abstracts* de revistas especializadas, é significativamente menor que o Corpus 2 (59.814 *tokens*), formado por *abstracts* de alunos, pois o abstract passou a ser obrigatório apenas nos últimos anos da publicação. Ao se verificar os dados da tabela, é possível dizer que o corpus 1 tem maior incidência de verbos e substantivos, tendo assim maior informação compactada, especialmente possibilitada pelo uso de nominalizações, que são substantivos que permitem o empacotamento de informações, elevando a densidade lexical.

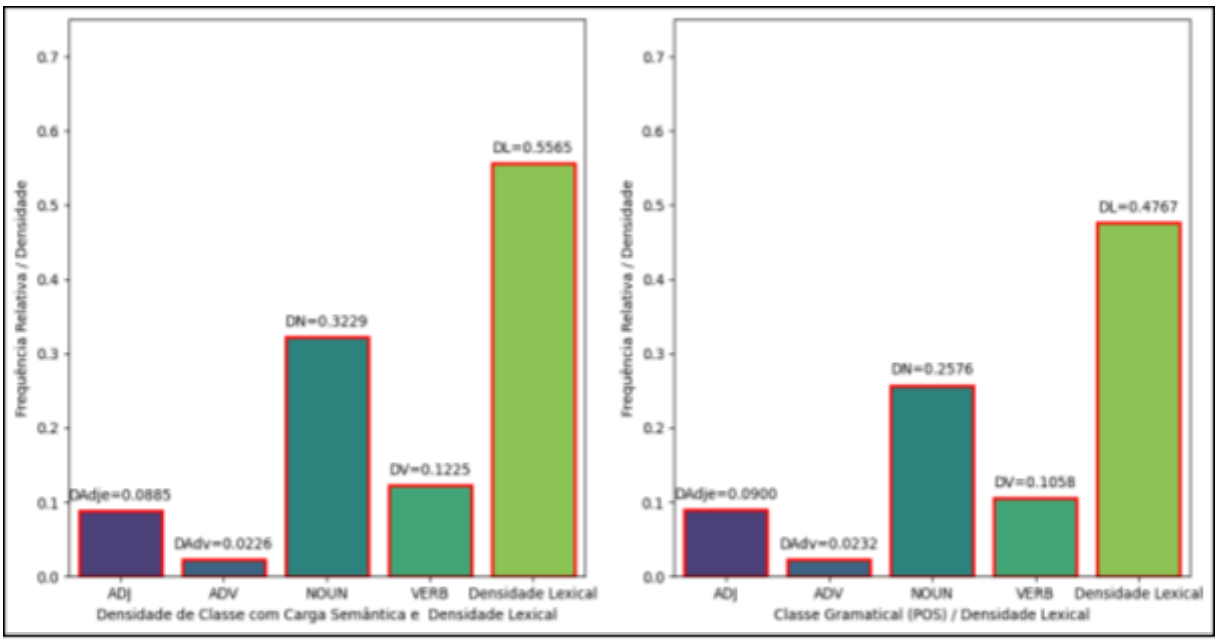

Figura 02: Densidades por especialidade e lexical para Corpus 1 e Corpus 2.
Fonte: Os Autores.

Essas diferenças observadas nas densidades lexicais foram comparadas por meio do teste para proporções. A aplicação do teste z para proporções resulta em $z = 20.37, valorp = 0.0$, apontando para a rejeição da hipótese nula H_0(as densidades lexicais dos corpora são diferentes $\rho_1 = \rho_2$) em detrimento da hipótese alternativa H_1((as densidades lexicais dos corpora são diferentes $\rho_1 \neq \rho_2$). Portanto, o teste sugere de que há evidências estatisticamente significantes de que há diferenças entre as densidades lexicais DL ao nível de significância $\alpha = 0.05$.

A razão de risco ($RR = 1.167, IC95\% = [1.15, 1.18]$) e a razão de chances ($= 1.38$, e $IC\ 95\% = [1.336, 1.421]$) sugerem diferenças significativas em ambos os casos. A razão de risco sugere que é mais provável encontrar palavras com carga semântica no Corpus 1, enquanto há mais chances relativas de encontrar palavras com conteúdo no Corpus 1 em relação ao Corpus 2.

Indicam, portanto, que o Corpus 1 é mais informativo ou formal, usando mais palavras de conteúdo, o que pode ser típico de textos mais complexos ou densos.

Do ponto de vista linguístico esses dados sugerem que a diferença observada na densidade lexical entre os *abstracts* de pesquisadores experientes e de cadetes se alinha com princípios estabelecidos na teoria linguística sobre a relação entre linguagem, contexto e propósito comunicativo. A LSF (HALLIDAY, 1994 & HALLIDAY & MATTHIESSEN, 2004 e 2014) possibilita o entendimento de como as escolhas linguísticas são influenciadas pelo contexto da situação, incluindo o campo (assunto), tenor (relação entre os participantes) e modo (canal de comunicação). Neste contexto acadêmico, pode-se dizer que o campo dos *abstracts* de pesquisa requer um estilo denso e cheio de informações para transmitir eficientemente descobertas complexas a um público mais especializado. O tenor, refletindo a comunicação entre pesquisadores da área, permite um alto grau de conhecimento compartilhado assumido, contribuindo ainda mais para a concisão. O modo, muitas vezes, reflete as normas das publicações acadêmicas, limitando a quantidade de palavras e, portanto, reforçando as escolhas mais densas lexicalmente.

Por outro lado, o campo dos resumos de cadetes parece incluir assuntos menos complexos, tendo na relação entre os participantes um público menos especializado, o que permite que sejam realizadas escolhas menos complexas, priorizando a clareza pedagógica. O tenor, possivelmente envolvendo uma dinâmica docente-cadete, pode incentivar um estilo mais explicativo. O modo, embora enfatize a concisão, pode não estar sujeito aos mesmos limites de palavras rigorosos das publicações de pesquisa analisadas, permitindo uma abordagem mais discursiva. Isso corrobora com o que afirma Swales (1990), em seu trabalho sobre análise de gênero, ao pontuar que diferentes gêneros têm propósitos e convenções comunicativas distintas, as quais moldam suas características linguísticas. A menor densidade lexical nos resumos dos cadetes parece ser um reflexo da função do gênero como ferramenta de aprendizagem, em que a clareza e a acessibilidade são primordiais, contrastando com *abstracts* concisos e complexos publicados por pesquisadores experientes.

Além disso, o trabalho de Bernstein (1990/1996) sobre a teoria do código pode oferecer insights sobre as diferenças observadas. Os *abstracts* de pesquisa muitas vezes refletem um código elaborado, caracterizado por uma linguagem concisa e especializada, refletindo o entendimento compartilhado dentro de uma comunidade de discurso específica de aviação. Contudo, os dos cadetes podem empregar um código mais restrito, que seja menos dependente do contexto e forneça explicações mais explícitas, potencialmente atendendo a um público mais amplo com níveis variados

de especialização. Essa diferença na orientação do código e da comunidade de discurso pode contribuir para as variações na densidade lexical.

Do ponto de vista cognitivo, a escolha de determinado código impacta na leitura dos textos. Para Langacker (2008), o processamento de textos mais complexos requer maior esforço cognitivo, pois envolve abstrações que requerem leitores especialistas. Por isso, assumem um nível mais alto de conhecimento linguístico, permitindo uma maior carga de informação por palavra. Tendo em vista que para muitos a experiência de redação mais acadêmica e elaborada ocorre somente ao elaborarem seus textos de TCC, os resumos dos cadetes, potencialmente direcionados a um público menos especializado, tendem a priorizar a facilidade de processamento, o que resulta em uma menor densidade lexical e a uma maior distribuição da carga informativa.

6 Considerações Finais

Esta análise revela uma diferença estatisticamente significativa na densidade lexical entre os *abstracts* de pesquisadores e de cadetes, apoiando a hipótese de os primeiros apresentarem maior grau de concentração informacional. Os achados se alinham com teorias linguísticas utilizadas, especialmente a LSF (HALLIDAY, 1994 e HALLIDAY & MATTHISSEN, 2004, 2014), a análise desse gênero textual (SWALES, 1990 e HYLAND, 1998) e a teoria do código (BERNSTEIN, 1971/1990), sugerindo que os *abstracts* têm objetivos comunicativos, públicos-alvo e fatores contextuais que influenciam significativamente suas respectivas características linguísticas.

A maior densidade lexical nos textos dos pesquisadores pode ser atribuída à necessidade de concisão da informação dentro de uma comunidade de discurso especializada, enquanto a menor densidade nos dos cadetes provavelmente reflete um foco na clareza pedagógica e acessibilidade para um público mais amplo.

Espera-se que pesquisas futuras explorem a análise de outras características linguísticas, como complexidade sintática, especificidade de termos e o uso de mecanismos coesivos, elucidando ainda mais as distinções estilísticas entre gêneros acadêmicos escritos por autores em diferentes níveis de formação. Dessa maneira, haveria uma compreensão ainda mais abrangente de como a linguagem se adapta para atender às demandas específicas dos diferentes contextos comunicativos.

7 Referências

BHATIA, V. K. *Analyzing genre: language use in professional settings*. London: Longman, 1993.

BHATIA, V. K. Analyzing professional genres: From situated practices to global genres. Hong Kong: Hong Kong University Press, 2017.
BERBER SARDINHA, V. G. *Linguística de corpus*. São Paulo: Manole, 2004.
BERNSTEIN, B. *Pedagogy, symbolic control and identity: theory, research, critique*. Londres: Taylor and Francis, 1996.
_____. *Class, codes, and control: The structuring of pedagogic discourse.* Londres: Routledge. 1990. v.4.
BIDERMAN, M. C. R. *Teoria estilística: semiótica e retórica*. São Paulo: Edusp, 1998.
COLOMBI, M. C. Academic discourse socialization: Community practices and individual development. L2 Journal, 1(1), 61-88, 2002.
CRYSTAL, D. *The Cambridge Encyclopedia of the English language*. 2. ed. Cambridge: Cambridge University Press, 2006.
DEVORE, J. L. *Probabilidade e estatística: para engenharia e ciências*. Tradução de Joaquim Pinheiro Nunes da Silva. São Paulo: Cengage Learning, 2006. ISBN 978-85-221-0924-1.
EGGINS, S. *An introduction to Systemic Functional Linguistics.* New York: Continuum International Publishing Group, 2004.
HALLIDAY, M. A. K. *An Introduction to Functional Grammar.* London: Edward Arnold, 1994.
HALLIDAY, M. A. K. *The language of science*. New York: Continuum, 2004.
HALLIDAY, M. A. K. e MATTHIESSEN, C. M.I.M. *An Introduction to Functional Grammar.* London: Edward Arnold. Third Edition, 2004.
HALLIDAY, M. A. K. e MATTHIESSEN, C. M.I.M. *An Introduction to Functional Grammar*. London: Edward Arnold. Third Edition, 2014.
HUNTER, J. D. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, v. 9, n. 3, p. 90-95, 2007. Disponível em: https://matplotlib.org/. Acesso em: 28 out. 2024.
HYLAND, K. *Hedging scientific research articles*. Applied Linguistics, v. 19, n. 2, p. 54-77, 1998.
ISMAIL, N. M. YOESTARA, M. & JAMILAH, S. Comparing lexical density in teacher talks: elementary school and higher educational level. *A Journal on Language and Language Learning*. Vol. 26, n. 1, April 2023.
LANGACKER, R. W. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press, 2008.
LARSON, R. e FARBER, B. *Estatística aplicada.* 6. ed. São Paulo: Pearson, 2015.
MONTGOMERY, D. C. e RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros.* 5. ed. Rio de Janeiro: LTC, 2012.
MCKINNEY, W. *Pandas: powerful Python data analysis toolkit*. Versão 1.0.5, 2020. Disponível em: https://pandas.pydata.org/. Acesso em: 28 out. 2024.
MORAIS, F.B.C. & BARBARA, L. O uso de nominalização como recurso de impessoalização em artigos científicos escritos em Língua Portuguesa: um estudo com base na Linguística Sistêmico-Funcional. *Cadernos de Linguagem e Sociedade*, v. 19, p. 73-91, 2018.
_____. Análise de Abstracts da Área de Aviação: padrões de organização textual e léxico-gramatical. *Revista Agulhas Negras*, 7(10), 172-191, 2023.

_____. & MARTINS, J.P. *A construção do corpus de artigos científicos de aviação: um estudo interdisciplinar. Revista da UNIFA*, v. 37, p. 1-21, 2024.

_____. Análise léxico-gramatical de adjuntos de modo em artigos científicos de aviação: contribuições para o ensino de Língua Inglesa. *Revista E-SCRITA*, v. 15, p. 109-121, 2024a.

_____. O uso de operadores modais em artigos acadêmicos de aviação: um estudo descritivo. *LÍNGUATEC*, v. 9, p. 70-86, 2024b.

NICHOLS, J. Linguistic complexity: A comprehensive definition and survey. In G. Sampson, D. Gil & P. Trudgill (Eds.). *Language complexity as an evolving variable*. New York: Oxford University Press, 2009.

PIAO, S. *et al*. Development of the multilingual semantic annotation system. *In proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*, Denver, Colorado, United States, pp. 1268-1274, 2015. Disponível em https://aclanthology.org/N15-1137.pdf.

PIAO, S. *et al*. Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. In proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016), Portoroz, Slovenia, pp. 2614-2619. Disponível em http://www.lrec-conf.org/proceedings/lrec2016/pdf/257_Paper.pdf

PYMUSAS Development Team. *PyMUSAS: multilingual semantic annotation system for Python*. 2024. Disponível em: https://pypi.org/project/pymusas/ ou https://github.com/UCREL/pymusas. Acesso em: 28 out. 2024.

PYTHON SOFTWARE FOUNDATION. *Python documentation*. [S.l.]: Python Software Foundation, 2001-2025. Disponível em: https://docs.python.org/. Acesso em: 23 fev. 2025.

SEABOLD, S. e PERKTOLD, J. *Statsmodels: Econometric and statistical modeling with Python*. Proceedings of the 9th Python in Science Conference, 2010. Disponível em 6417e5350e29cb6bf04ea5a4785601d5a215.pdf (semanticscholar.org). Acessado em 04 nov. 2024.

SARAGIH, A. Metaphorical representations and scientific texts. Englonesia: *An Indonesian Scientific Journal on Linguistics and Literature*, 2(1), 1-11, 2006.

SCHLEPPEGRELL, M. J. *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, 2004.

SCHLEPPEGRELL, M. J. *Linguistics and literacy development: An educational perspective*. Routledge, 2012.

SCOTT, M. *WordSmith Tools.* Version 6.0. Liverpool: Lexical Analysis Software, 2008.

SWALES, J. M. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press, 1990.

SINAR, T. S., Zein, T. T., Ganie, R., Syarfina, T., Mahriyuni, Yusuf, M., & Rangkuti, R. Content Words and Readability in Students' Thesis Findings. *Journal of Curriculum and Teaching*, 12(6), 347-355, 2024.

SUBBS M. *Text and Corpus Analysis*. Oxford: Blackwell, 1996.

THORNBURY, S., & SLADE, D. *Conversation: From description to pedagogy*. Cambridge University Press, 2006.

URE, J. N. Lexical density and register differentiation. In: D. Crystal & D. Davy (Eds.), *Investigating language style*. London: Longman, 1971.

LANCASTER UNIVERSITY. *UCREL Semantic Analysis System (USAS)*. Disponível em: https://ucrel.lancs.ac.uk/usas/. Acesso em: 5 fev. 2025.

WASKOM, M. L. Seaborn: Statistical data visualization. *Journal of Open-Source Software*, v. 6, n. 60, p. 3021, 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 28 out. 2024.

Data de submissão: 07/03/2025. Data de aprovação: 28/05/2025.