

DISCUTINDO O VOCABULÁRIO ACADÊMICO EM LÍNGUA INGLESA

DISCUSSING ACADEMIC VOCABULARY IN THE ENGLISH LANGUAGE

SILVA, Eduardo Batista da (UEG)¹

RESUMO: A leitura e a redação de artigos científicos e obras de referência em língua inglesa dependem do conhecimento do vocabulário acadêmico. O presente trabalho tem o objetivo geral de descrever e analisar o vocabulário acadêmico na língua inglesa, baseando-se na Lexicologia e na Linguística de Córpus. Com relação à metodologia, constituímos um c3rpus acad3mico em língua inglesa com 113.337.773 de *tokens*. No tocante ao aux3lio de *software*, a ferramenta lingu3stico-estatística utilizada para processar os *corpora* e coletar os termos foi o *WordSmith Tools*, vers3o 5. Para identificar os termos, utilizamos o coeficiente de uso (resultado da multiplicação do coeficiente de dispers3o de Juilland pela frequ3ncia normalizada) e a posterior validaç3o dos candidatos a termo por especialistas. Os resultados permitem destacar um n3cleo composto por 505 voc3bulos/termos, predominantemente substantivos, verbos e adjetivos. De forma geral, o uso do vocabulário acad3mico nas oito grandes 3reas do conhecimento revela ser mais intenso que na língua geral, conforme resultados obtidos pelo teste de soma de postos de Wilcoxon, com um n3vel de confiança de 95%. O vocabulário acad3mico identificado em nosso trabalho apresenta, al3m de uma alta ocorr3ncia, uma distribuiç3o marcante nos diferentes *corpora*, n3o sendo identificado, por estas raz3es como pertencente a um dom3nio de especialidade espec3fico.

Palavras-chave: Vocabulário Acad3mico. Linguística de Córpus. Língua Inglesa.

Abstract: Reading and writing of scientific articles and reference works in the English language rely on the knowledge of the academic vocabulary. The present study has the general objective of describing and analyzing the academic vocabulary in the English language, occurring in eight fields of knowledge, based on Lexicology and Corpus Linguistics. Concerning the methodology, we constituted an academic corpus in the English language with 113.337.773 tokens. As for the aid of software, WordSmith Tools, version 5, was the linguistic-statistical tool used to process the corpora and retrieve the terms. In order to identify the terms, we used the use coefficient (result of the multiplication of the Juilland's dispersion coefficient by the normalized frequency) and posterior validation of the eligible terms by experts. The results allow to highlight a core composed of 505 vocabulary/term, mostly nouns, verbs and adjectives. Overall, the academic vocabulary use in the eight knowledge areas proves to be more intense than in the general language, according to results obtained by the Wilcoxon Rank Sum Test, with a 95% confidence level. The academic vocabulary identified in our investigation shows, besides a high occurrence, a marked distribution among the different *corpora*, not being identified, for these reasons as belonging to a specific specialized domain.

Keywords: Academic Vocabulary. Corpus Linguistics. English Language.

¹Professor da Licenciatura em Letras da UEG/Câmpus Morrinhose professor permanente do Programa de Pós-Graduação *Stricto Sensu* em Língua, Literatura e Interculturalidade (POSLLI) na UEG/Câmpus Cora Coralina.

1 INTRODUÇÃO

Frente à escassez de trabalhos escritos em língua portuguesa acerca do vocabulário acadêmico em uma perspectiva quantitativa e computacional, um trabalho que explore essa temática pode servir para que pesquisadores, estudantes e professores – incluindo aqueles que não leem em inglês com proficiência – possam se inteirar a respeito do estado da arte. O resultado da barreira linguística é duplamente negativo: impede o acesso a novas ideias (no caso da habilidade de leitura de artigos científicos) e restringe a produção científica brasileira apenas aos domínios da língua portuguesa (no que se refere à redação de artigos científicos).

O objetivo geral deste trabalho é descrever e analisar o vocabulário acadêmico na língua inglesa, presente em oito grandes áreas do conhecimento, divididas em 69 subáreas. A tabela das áreas do conhecimento elaborada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) serviu como ponto de partida para que a pesquisa pudesse ser estruturada. Com relação aos objetivos específicos, a pesquisa pretende: 1) constituir um corpus de especialidade em língua inglesa; 2) propor uma metodologia de cunho quantiquantitativo para a identificação e coleta de termos e 3) identificar o vocabulário acadêmico fundamental.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, apresentaremos diferentes contribuições referentes à conceituação do vocabulário acadêmico (DUDLEY-EVANS; SAINT JOHN, 1998; COXHEAD, 2000; NATION, 2001; HYLAND; TSE, 2007; BARBOSA, 2009; SILVA; BABINI, 2012). Destacamos que o vocabulário acadêmico recebe diferentes denominações a depender da perspectiva teórica. Para Coxhead (2000), trata-se de palavras acadêmicas. Hyland e Tse (2007) indicam que o termo em questão também é conhecido como vocabulário subtécnico, semitécnico, ou léxico especializado não técnico. Para Barbosa (2009), vocábulo-termo. Nos textos acadêmicos, além dos termos propriamente ditos, existe também um vocabulário compartilhado, que caracteriza-se por ocorrer em domínios de especialidade variados. Adotamos a conceituação de Silva e Babini (2012) no que se refere ao entendimento do que é o vocabulário acadêmico: um conjunto que não é “pertencente” a um domínio de especialidade específico, ou seja, são itens relativamente frequentes em diversos gêneros acadêmicos, mostrando-se, no entanto, relativamente incomum em outros gêneros textuais.

Sob a mesma perspectiva, Nation (2001, p. 17, tradução nossa) categoriza o vocabulário em três grupos: 1) palavras de alta frequência que constituem um grupo de 2.000 famílias de palavras, cobrindo cerca de 80% da maioria dos textos em língua inglesa; 2) Um vocabulário acadêmico de palavras que é relativamente frequente na redação acadêmica que compreende entre 8 e 10% de palavras correntes de um texto acadêmico e 3) Um vocabulário técnico que difere em função da área e abrange até 5% dos textos.

Baker (1988, p. 92, tradução nossa) lista seis categorias de vocabulário, que podem ser relacionadas ao inglês para finalidades acadêmicas (*English for Academic Purposes*, EAP), a saber:

- 1 itens que expressam noções gerais em todas as disciplinas de especialidade;
- 2 itens da língua geral que tem um significado especializado em uma ou mais disciplinas;
- 3 itens especializados que possuem diferentes significados em diferentes disciplinas;
- 4 itens da língua geral que possuem sentidos restritos em diferentes disciplinas;
- 5 itens da língua geral que são usados para descrever ou referir-se a procedimentos técnicos ou funções em detrimento de outros itens com o mesmo significado, por exemplo, *occur* ao invés de *happen*;
- 6 itens utilizados para indicar intenções ou avaliação do autor sobre o material apresentado².

A referida autora sugere que as seis categorias se sobrepõem consideravelmente e que parece haver duas grandes áreas: vocabulário que é usado na língua geral, mas que tem uma alta frequência nas descrições científicas e técnicas (especialmente as categorias 1 e 5); e o vocabulário que possui significado especializado e restrito em certas disciplinas e que podem sofrer variações de sentido. O quadro a seguir foi traduzido, conforme proposto por Dudley-Evans e Saint John (1998, p. 82):

² 1 items which express notions general to all specialized disciplines; 2 general language items that have a specialized meaning in one or more disciplines; 3 specialized items that have different meanings in different disciplines; 4 general language items that have restricted meanings in different disciplines; 5 general language items that are used to describe or comment on technical processes or functions in preference to other items with the same meaning, for example occur rather than happen; 6 items used to signal the writer's intentions or evaluation of material presented.

Quadro 1 – Tipos de vocabulário³

Tipo de vocabulário	Exemplos
Vocabulário geral com alta frequência em uma área específica	acadêmico: <i>fator, método, função, ocorrer, ciclo</i> ; adjetivos avaliativos como <i>relevante, importante, interessante</i> ; turismo: verbos como <i>aceitar, aconselhar, concordar, confirmar</i> ; colocações, como <i>fazer uma reserva, lançar uma campanha</i> .
Palavras do inglês geral que possuem um significado específico em certas disciplinas	<i>bug</i> na ciência da computação; <i>força, aceleração e energia</i> na física; <i>stress e pressão</i> na mecânica e na engenharia.

Fonte: Dudley-Evans e Saint John (1998, p. 82, tradução nossa).

Dudley-Evans e Saint John (1998, p. 82) acreditam que, no ensino de inglês na modalidade inglês para fins específicos (ESP), é a primeira categoria (com vocabulário semitécnico) que deveria ter prioridade. Nesse sentido, o presente trabalho procura destacar o repertório constituído pelo vocabulário típico da comunicação acadêmica, com a clara intenção de também servir como uma referência no contexto da sala de aula de língua inglesa.

Inicialmente, interessa neste trabalho, recorrendo à classificação de Baker (1988), identificar e avaliar o vocabulário que expressa noções gerais nas disciplinas de especialidade (item 1) e, em segundo lugar, itens preferencialmente utilizados na comunicação acadêmica quando existe um item equivalente na língua geral (item 5).

Uma vez que as diferentes áreas do saber recorrem as estruturas gramaticais e lexicais, além da terminologia propriamente dita, parece claro o entendimento de que um vocabulário acadêmico é apenas um dos passos necessários para que o aprendiz ganhe autonomia na leitura/escrita acadêmica, conforme indicado por Silva e Babini (2011).

Cada área do saber possui seus verbos preferenciais ou funcionais e um vocabulário com o vocabulário acadêmico geral, apesar de útil, não é capaz de servir como referência mais especializada. Ao apresentar os cinco verbos mais comuns em 11 disciplinas, Swales e Feak (2004, p. 163), mostram que na física o verbo expandir (*expand*) possui uma alta ocorrência, ao passo que na área de linguística aplicada o verbo destacar (*point out*) encontra-se entre os mais

³ Type of vocabulary: General vocabulary that has a higher frequency in a specific field. Examples: *academic: factor, method, function, occur, cycle*; evaluative adjectives such as *relevant, important, interesting*; tourism: verbs such as *accept, advise, agree, confirm*; collocations, such as *make a booking, launch a campaign*. Type of vocabulary: General English words that have a specific meaning in certain disciplines. Examples: *bug* in computer science; *force, acceleration and energy* in physics; *stress and strain* in mechanics and engineering.

usados. Por outro lado, verbos como encontrar (*find*), apresentar (*show*) e demonstrar (*demonstrate*) ocorrem em várias disciplinas, com o mesmo sentido.

Frente ao exposto até o momento, acreditamos que existe um conjunto de unidades lexicais funcionais atuando entre a língua comum e os textos acadêmicos (em todas as áreas e subáreas).

Conforme visto, as discussões acima são oriundas de estudos inseridos na área de ESP e EAP. O ensino de inglês na modalidade ESP preocupa-se com o ensino da língua inglesa e inclui, por exemplo, o inglês comercial, o inglês para profissionais da saúde, inglês para o turismo, e inglês para a aviação. O ensino de inglês na modalidade EAP é geralmente definido como aquele com o propósito de auxiliar o estudo ou pesquisa do aprendiz, especialmente no ensino superior (HYLAND, 2006). Um ou mais elementos/habilidades da língua podem ser explorados no ensino, como o léxico, estruturas gramaticais, leitura, redação, pronúncia ou compreensão auditiva.

Nesse sentido, Dayrell (2010), investiga diferenças entre resumos de artigos científicos escritos em inglês por pós-graduandos brasileiros e resumos já publicados e descobre diferenças relevantes entre escolhas lexicais e gramaticais de estudantes não nativos e falantes nativos e/ou experientes – o que indica que a tomada de consciência e o emprego adequado dos padrões léxico-gramaticais mais tipicamente usados por suas comunidades acadêmicas irá, certamente, auxiliar os aprendizes a escrever de forma mais eficaz.

Diversos trabalhos foram e vêm sendo realizados no sentido de estudar o texto acadêmico e seus itens léxico-terminológicos. No que diz respeito à língua inglesa, ressaltamos que, a partir do ano de sua publicação, em 2000, a *Academic Word List* (Lista de Palavras Acadêmicas), AWL, elaborada por Coxhead (2000), vem servindo como uma referência no âmbito dos estudos dos textos acadêmicos.

Outra lista que vem ganhando atenção por tratar do léxico presente nos textos acadêmicos, é a *Academic Vocabulary List*, AVL, desenvolvida por Mark Davies, e lançada no ano de 2012, pretende ser a última referência em termos de lista acadêmica (mais detalhes são apresentados na seção 1.2.2).

Dessa forma, a utilização do vocabulário acadêmico nos textos técnico-científicos em língua inglesa configura-se como um suporte na produção e divulgação de conhecimento nas mais diversas áreas de especialidade. Podemos perceber a relevância da lista de vocabulário

acadêmico recorrente – mais especificamente da AWL – ao atestar sua inclusão no repertório do *Oxford AdvancedLearner'sDictionary*, inclusive na versão online.

Contextualizando estas ideias para o contexto acadêmico, Swales (2004) também indica que o emprego do vocabulário utilizado nos textos de pesquisa acadêmica demanda mais atenção, uma vez que possuem especificidades de uso. Biber (2006, p. 33) apresenta questionamentos que julgamos essenciais:

A descrição do uso de vocabulário no contexto universitário é um pré-requisito essencial para o desenvolvimento de materiais de ensino e abordagens eficientes. Existem muitas perguntas de pesquisa importantes relacionadas ao uso das palavras na linguagem universitária. Por exemplo, quantas palavras um estudante precisaria saber para ler um livro universitário típico? Quantas para entender uma palestra universitária típica? Diferentes disciplinas acadêmicas usam as mesmas palavras? Se não, qual a sobreposição existente entre as disciplinas? Algumas disciplinas utilizam uma variedade maior de palavras diferentes? Algumas palavras são comuns no uso do dia a dia e também comuns na linguagem acadêmica? Que proporção de um texto acadêmico típico é constituído dessas palavras comuns?⁴ (BIBER, 2006, p. 33)

A partir da argumentação de Biber, procuraremos responder à seguinte pergunta: quantas palavras um estudante precisaria saber para ler um texto universitário típico, uma vez que constitui escopo da presente pesquisa identificar quais as palavras que as diferentes disciplinas acadêmicas utilizam. Paralelamente, recorreremos aos estudos de cunho estatístico-lexical que indicam que o domínio das 2.000 palavras mais comuns da língua geral permite a compreensão de mais de 70% de um texto. Estudos anteriores, e mais recentemente, Cobb (2007) e Nation (2001) trabalham com o mesmo recorte de palavras. Nation salienta que as palavras de alta frequência merecem atenção do professor, do aprendiz e do livro-texto. Sob a mesma orientação em termos de frequência, Biderman (2001), em uma pesquisa com a língua portuguesa, descobriu que 2.000 palavras perfaziam 84% de seu corpus.

A exploração computacional do léxico encontra-se facilitada a partir da observação e análise de usos concretos, fornecendo subsídios teórico-metodológicos para muitas

⁴The description of vocabulary use in university contexts is an essential prerequisite to the development of effective teaching materials and approaches. There are many important research questions about word use in university language. For example, how many words would a student need to know to read a typical university textbook? How many to understand a typical university lecture? Do different academic disciplines use the same words? If not, how much overlap is there across disciplines? Do some disciplines use a greater range of different words? Are some words common in everyday use and also common in academic language? What proportion of a typical academic text is made up of those common words?

áreas dentro da Linguística. A Linguística de Corpus não se dedica a um assunto definido, não constituindo, portanto, uma disciplina tal qual a psicolinguística, sociolinguística ou semântica, pois seu objeto de pesquisa não é delimitado como em outras áreas. Para Leech (apud Berber Sardinha, 2004, p. 37), a Linguística de Corpus é “não somente uma nova metodologia emergente para o estudo da linguagem, mas uma nova empreitada de pesquisa e, na verdade, uma nova abordagem filosófica”.

Sinclair (2004) define um corpus como uma coleção de partes de textos em formato eletrônico, selecionados de acordo com critérios externos para representar, tanto quanto possível, uma língua ou uma variedade de língua como uma fonte de dados para a pesquisa linguística. Berber Sardinha (2004) ressalta que o corpus é uma coleção de textos em linguagem natural, autêntica, que pode ser utilizado como embasamento para pesquisas linguísticas diversas, bem como a observação empírica de dados e padrões encontrados em textos. É neste caso que a disciplina acaba se tornando uma importante aliada, pois é por meio dela que podemos estabelecer padrões advindos da língua alvo e criar materiais diversos, incluindo materiais terminológicos. (SILVA; BABINI, 2012). Por meio dos estudos de investigação da padronização do léxico, que se vale de *corpora*, foram identificadas regularidades na língua, que se definem pela frequência de certos termos nas construções semânticas dos usuários da língua em questão.

A Linguística de Corpus estuda a língua com o auxílio de ferramentas computacionais a fim de processar grandes quantidades de dados linguísticos reais (textos falados e/ou escritos) provenientes da comunicação natural, ou seja, da língua em uso. Trata-se de uma abordagem que privilegia a observação de grande quantidade de dados autênticos coletados a partir de *corpora* eletrônicos, cujo objetivo é o de pesquisar, a partir de ferramentas computacionais, dados linguísticos textuais coletados criteriosamente (BERBER SARDINHA, 2004). Assim, a área em questão trata da coleta e análise de materiais escritos ou falados, que são agrupados de acordo com critérios específicos. O corpus, por sua vez, é formado de textos de diversos gêneros, constituindo amostras autênticas da língua que se almeja compreender e pesquisar. Biber, Conrad e Reppen (2004) afirmam que a abordagem baseada em corpus fornece um meio de manipular grandes quantidades de língua e manter o controle de muitos fatores contextuais ao mesmo tempo. Trata-se de uma tendência do mundo contemporâneo a criação de materiais educativos ou profissionais de qualidade com o subsídio de estudos baseados em corpus.

Acreditamos que a Linguística de Córpus não oferece somente um conjunto de técnicas computadorizadas para verificação de questões tradicionais do léxico, semântica, sintaxe, etc. Pelo contrário, a análise de um cörper pode revelar fatos a respeito de uma língua pelos quais nunca se tinha pensado procurar. Dessa forma, a Linguística de Córpus apresenta-se como tendo um caráter exploratório e ao mesmo tempo contestatório, não sendo somente uma nova metodologia para o estudo da linguagem, mas uma nova forma de pesquisa, uma abordagem.

Berber Sardinha (2004, p. 3) destaca que “foi um *corpus* não-computadorizado que deu feição aos *corpora* atuais, o SEU (*Survey of English Usage*) [...] a partir de 1959”. O SEU serviu como referência para os *corpora* que se seguiram. Frente à modernização da tecnologia, especialmente o advento de computadores dispondo de uma capacidade de armazenamento cada vez maior, a exploração de *corpora* de grandes dimensões passou a ser realizável. O primeiro cörper moderno foi o *the Brown University Standard Present-day American English*, conhecido simplesmente como *Brown Corpus*. Possuía um milhão de palavras e era considerado uma referência para o registro do inglês norte-americano do final dos anos 1960. Os anos seguintes presenciaram o desenvolvimento de inúmeros *corpora* cujas dimensões aumentavam consideravelmente em comparação aos *corpora* anteriores.

A seguir, uma breve descrição de alguns *corpora* constituídos para o estudo dos textos acadêmicos.

A *Academic Word List* (AWL) foi desenvolvida por Coxhead (2000). A AWL é composta de 570 famílias de palavras e contempla as palavras mais comuns dos textos técnico-científicos. Nation (2003, p. 61) explica que a AWL foi elaborada a partir de um cörper com 3.600.000 palavras, contendo textos acadêmicos de quatro seções principais: Artes, Comércio, Ciência e Direito. Cada seção foi subdividida em sete subseções. Para constarem na AWL, as famílias de palavras deveriam ocorrer nas quatro seções principais e, em pelo menos, 15 das 28 subseções. Chegou-se a um número final de 10 listas de palavras acadêmicas. Para o falante de língua portuguesa, diversas palavras da AWL não apresentam dificuldade de compreensão porque são cognatas, de origem latina, fato este que auxilia no estudo do vocabulário acadêmico. As palavras da AWL ocorrem em todos os tipos de textos, mas são muito mais comuns em textos acadêmicos. Elas englobam cerca de 1,9% das palavras correntes na conversação, cerca de 1,9 de ficção, e mais ou menos 4% nos jornais (NATION, 2001). A AWL abrange de 8.5% a 10% das palavras que ocorrem em textos acadêmicos. As palavras nessa lista constituem, portanto, um

objetivo essencial na aprendizagem dos alunos que pretendem dedicar-se à área acadêmica usando o inglês como meio de comunicação e que já conhecem as 2.000 palavras mais frequentes (NATION, 2001). Em outras palavras, Nation (2001) aponta que as palavras acadêmicas devem ser idealmente estudadas após o domínio das primeiras 2.000 palavras.

O *cópus* norte-americano, da universidade de Michigan, MICUSP (*Michigan Corpus of Upper-Level Student Papers*), que totaliza pouco mais de dois milhões e meio de palavras (2.600.000) possui apenas 830 artigos em seu banco de dados e possui 16 disciplinas, categorizadas em quatro grandes áreas – Ciências Humanas e Artes, Ciências Sociais, Ciências Biológicas e da Saúde, e, por último, Ciências Exatas. Apesar de dispor de algumas funcionalidades interessantes para a pesquisa terminológica em língua inglesa, a página online do projeto não permite a visualização das concordâncias ou o cálculo de palavras-chave.

Mais recentemente, foi desenvolvida por Mark Davies a *Academic Vocabulary List* (AVL), publicada em 2012. A AVL é baseada em um *cópus* acadêmico de aproximadamente 120.000.000 de palavras, possuindo 9 divisões: História (14 milhões de palavras), Ciências Sociais (16.7 milhões), Ciência e Tecnologia (22.8 milhões), Educação (8 milhões), Ciências Humanas (11.1 milhões), Medicina e Saúde (9.7 milhões), Direito e Ciência Política (12.2 milhões), Filosofia, Religião e Psicologia (12.5 milhões) e Negócios e Finanças (5.3 milhões). Gardner e Davies (2012) adotam o coeficiente de dispersão (informado como sendo maior que 0,80) e a frequência em sua seleção.

A título de exemplificação, o *Pearson International Corpus of Academic English* (PICA) não se encontra disponível para consulta. Trata-se de um *cópus* com 37 milhões de palavras cuja divisão das áreas do conhecimento segue um modelo disponível na enciclopédia colaborativa *Wikipedia*.

Para a presente pesquisa, nosso *cópus* foi batizado como Academic Corpus of English (ACE). A seguir, de forma resumida, são apresentadas as características dos *corpora* acadêmicos citados.

Tabela 1 – Comparação entre *corpora* acadêmicos em língua inglesa

cópus acadêmico	grandes áreas	subáreas	tokens
AWL	4	28	3.600.000
MICUSP	4	16	2.600.000
PICA	4	28	37.000.000
AVL	9	-	120.000.000

ACE	8	69	122.464.043
-----	---	----	-------------

Fonte: Dados da presente pesquisa.

Nota 1: O número de subáreas da AVL não é indicado pelos autores.

Nota 2: No ACE, o valor bruto de *tokens* é 122.464.043. No entanto, o valor processado pelo *software* foi 113.337.773, uma vez que símbolos e números não são reconhecidos pelo WST.

O tema do vocabulário acadêmico em língua inglesa discutido aqui mostra-se relevante porque procura trazer uma contribuição à descrição linguística. Acreditando que o aproveitamento de insumos obtidos no diálogo com outros ramos do saber fornece subsídios que enriquecem as discussões e a compreensão do fenômeno linguístico sobre o qual nos debruçamos – o vocabulário acadêmico –, conforme visto, recorreremos principalmente as contribuições da Lexicologia, aliando nossas reflexões à Linguística de Córpus.

3 METODOLOGIA

No que se refere aos recursos materiais (*softwares* da pesquisa) utilizados, foram empregados quatro *softwares*, a saber: MS Word 2013, MS Excel 2013 (incluindo o suplemento *AnalysisToolPak*), *WordSmith Tools* e SPSS.

No que se refere ao processamento do conteúdo textual dos *corpora* de especialidade, optamos pelas ferramentas linguístico-estatísticas do *WordSmith Tools* (WST), versão 5, desenvolvido por Scott (2010). Com relação ao software utilizado para as análises estatísticas mais complexas, optamos pelo pacote estatístico *StatisticalPackage for the Social Sciences*, SPSS, versão 22. Sua utilização deve-se ao fato de o *software* incluir os testes mais comuns juntamente com uma variedade de recursos para análise e visualização de dados pela geração de gráficos.

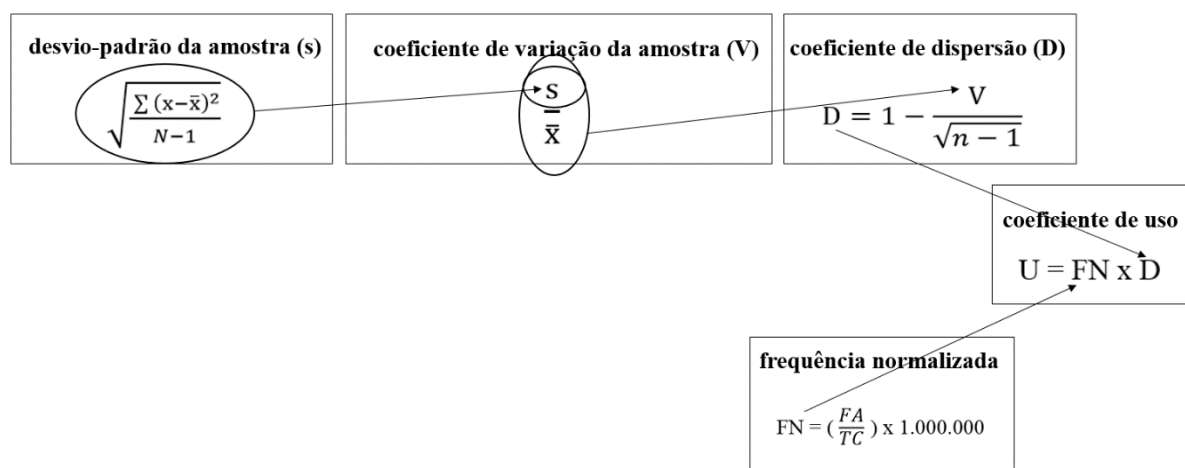
No âmbito dos procedimentos metodológicos, a pesquisa pode ser dividida em oito etapas distintas, sequenciais e complementares. A primeira delas consistiu na constituição e caracterização do *cópus* de pesquisa. A fim de constituir um *cópus* que representasse o vocabulário acadêmico e que apresentasse uma alta frequência e uma distribuição regular em oito grandes áreas do conhecimento (Ciências Exatas e da Terra, Ciências Biológicas, Engenharias, Ciências da Saúde, Ciências Agrárias, Ciências Sociais Aplicadas, Ciências Humanas e Linguística, Letras e Artes), foram compilados *corpora* em 69 subáreas. A etapa subsequente desenvolveu-se pela exploração dos recursos presentes no WST, a saber: lematização, *stop list* e o uso da lista de consistência detalhada. Na terceira etapa, houve a

necessidade de uma normalização dos dados obtidos e a aplicação de fórmulas estatísticas. Na quarta etapa, todos os dados foram tabulados e passaram por outras análises estatísticas em uma planilha eletrônica do MS Excel 2018. Na quinta etapa, desenvolvemos o formulário para que os candidatos a termo fossem validados por especialistas. Na sexta etapa, organizamos o sistema nocional do vocabulário acadêmicos. Na sétima etapa, foi desenvolvida uma ficha terminológica eletrônica na qual os dados foram inseridos. Por fim, na oitava e última etapa, empregamos recursos do MS Excel para organizar o vocabulário acadêmico em função do quartil.

Uma vez que a presente pesquisa reúne características de comparação, houve a preocupação em estabelecer condições propícias e válidas relacionadas aos cálculos dos valores de frequência encontrados.

Nosso cálculo principal é chamado coeficiente de uso, que é o resultado do produto da frequência normalizada multiplicada pelo coeficiente de dispersão ($U = FN \times D$). O coeficiente de uso demandou outros cálculos, conforme demonstrado na Figura 1.

Figura 1 – Cálculos estatísticos utilizados



Fonte: Dados da presente pesquisa.

Destacamos, nos próximos parágrafos, os cálculos e fórmulas que sustentam e precedem a frequência normalizada e o coeficiente de dispersão.

As dimensões desiguais entre os *corpora* e a possibilidade de interpretações errôneas demandaram um cálculo conhecido como normalização, que fornece valores comparáveis dentro de determinada dimensão. Procedemos à normalização dos valores de frequência por meio da seguinte fórmula:

$$FN = \left(\frac{FA}{TC} \right) \times 1.000.000 \quad (1)$$

A frequência normalizada (FN) é igual à frequência absoluta (FA) dividida pelo tamanho do corpus (TC). O valor obtido é então multiplicado por um milhão. Assim, todas as comparações são realizadas levando em consideração a mesma escala, com as mesmas proporções. Uma palavra que ocorre 5 vezes em um subcorpus contendo meio milhão de palavras é mais frequente que outra – ou ela mesma – que ocorre 7 vezes em outro subcorpus com um milhão de palavras. Portanto, todas as comparações entre as áreas de especialidade podem ser efetuadas em um determinado número de ocorrências a cada milhão de palavras.

No entanto, a simples observação direta das frequências normalizadas nos diferentes *corpora* prova ser insuficiente para afirmar se determinada expressão ocorre uniformemente nas diferentes áreas do conhecimento. Portanto, com a frequência normalizada de cada palavra em cada corpus de pesquisa, o passo seguinte consistiu na descoberta de quais delas apresentavam uma distribuição mais regular. Vale destacar que uma palavra com uma alta frequência pode se concentrar em apenas um, ou ainda, em poucos *corpora*, não constituindo o tipo de expressão acadêmica idealizada neste estudo. Por outro lado, uma outra expressão pode ter uma distribuição equânime entre todos os *corpora*, mas uma baixa frequência. O intuito, portanto, é selecionar as expressões que, além de frequentes, sejam bem distribuídas nos oito *corpora* de pesquisa. Lançamos mão da fórmula do coeficiente de dispersão, tal como apresentada a seguir:

$$D = 1 - \frac{V}{\sqrt{n-1}} \quad (2)$$

Conforme indicado, o coeficiente de variação da amostra (V) é obtido pela divisão do desvio padrão da amostra pela média. O valor de n é substituído pelo número de elementos da amostra, que, em nossa pesquisa, equivale ao número de subáreas em cada grande área do conhecimento.

4 RESULTADOS E ANÁLISE DOS DADOS

Conforme indicado anteriormente, o vocabulário acadêmico da presente pesquisa concentra 505 vocábulos/termos. Seu coeficiente de dispersão varia entre o valor mínimo de 0,30 e o valor máximo de 0,95. Seu coeficiente de uso, por sua vez, varia entre o valor mínimo de 90 e o valor máximo de 4.164.

A lista constando o vocabulário acadêmico a seguir foi obtida após a organização dos *corpora*, dos procedimentos com o auxílio do WST, dos cálculos estatísticos e da validação dos especialistas. Esse processo demandou o corte e a seleção do vocabulário que comporia o núcleo da comunicação acadêmica.

No que se refere à distribuição das categorias gramaticais entre o vocabulário acadêmico, podemos notar a importância dos verbos, substantivos e adjetivos. Levando em consideração apenas duas classes gramaticais, quase 77% dos casos são representados por verbos e substantivos. Os adjetivos perfazem pouco mais de 20% dos casos.

Tabela 1 – Categorias gramaticais do vocabulário acadêmico

categoriagramatical	N (%)
verbos	200 (39,60)
substantivos	184 (36,44)
adjetivos	112 (22,18)
advérbios	5 (0,99)
preposições	3 (0,59)
conjunção	1 (0,20)
total	505 (100)

Fonte: dados da presente pesquisa

Apresentados em ordem decrescente, os substantivos identificados em nossa pesquisa são os seguintes:

FUNCTION, STRUCTURE, PROCESS, CHAPTER, METHOD, DATA, AREA, RESEARCH, FACTOR, PERIOD, APPROACH, RANGE, SOURCE, SECTION, REGION, RESPONSE, THEORY, POTENTIAL, JOURNAL, CULTURE, ELEMENT, ROLE, COMPONENT, SITE, TECHNIQUE, IMAGE, TECHNOLOGY, CONCEPT, PRINCIPLE, PROJECT, CYCLE, FEATURE, PHASE, FOCUS, ISSUE, LINK, PERCENT, SERIES, INDEX, STRATEGY, EVIDENCE, COMMUNITY, IMPACT, RESOURCE, ACCESS, MECHANISM, SEQUENCE, RATIO, ASPECT, TRADITION, TARGET, CONTRAST, SUMMARY, AUTHOR, DIMENSION, CATEGORY, CAPACITY, PROPORTION, CONTACT, ERROR, BENEFIT, LAYER, SHIFT, CONVENTION, PARAMETER, POLICY, SURVEY, GOAL, NETWORK, TASK, CONTEXT, INSTANCE, CODE, MEDIUM, STATUS, OBJECTIVE, TEXT,

AID, OUTPUT, EMPHASIS, ACADEMY, ITEM, MODE, COLUMN, ATTRIBUTE, CRITERION, CONSEQUENCE, DOCUMENT, CONCLUSION, TREND, PORTION, PERSPECTIVE, SEGMENT, INPUT, DOMAIN, SCHEME, TRANSITION, SYMBOL, MATRIX, DEVICE, OUTCOME, VERSION, SYNTHESIS, CORE, OPTION, TOPIC, INTERVAL, COUPLE, FOUNDATION, PHENOMENON, HYPOTHESIS, CHANNEL, GRADE, MOTIVE, DECADE, DIAGRAM, FRACTION, PEAK, COMMENT, CLUSTER, DURATION, BARRIER, FRAGMENT, FRAMEWORK, MARGIN, TEAM, EQUILIBRIUM, ATTITUDE, BIAS, CONVERSION, PANEL, LIKENESS, CIRCUMSTANCE, FEEDBACK, OVERLAP, ZONE, APPENDIX, INSIGHT, MAGNITUDE, PERCEPTION, PATHWAY, SPECTRUM, STYLE, CONFLICT, COMMISSION, HAZARD, REVOLUTION, OVERVIEW, GUIDELINE, GENDER, PROFILE, EXPERT, DEPENDENCE, SCHEDULE, MODULE, DEBATE, SECTOR, CHART, TRAIT, ROUTINE, INTERFACE, THEME, CAREER, THRESHOLD, SPAN, USAGE, SCOPE, PRIORITY, BULK, FORMAT, UTILITY, HETEROGENEITY, NOTION, PROSPECT, ARRAY, NORM, SCENARIO, CONTROVERSY, PARAGRAPH, INCIDENCE, CONVERSE, DISRUPT, SCHOLAR, COLLEAGUE

É importante salientar que parte dos substantivos elencados acima pode se transformar em verbos, mantendo a mesma grafia. É o caso de FUNCTION, STRUCTURE, PROCESS, RESEARCH, APPROACH, PROJECT, FEATURE, FOCUS, ISSUE, LINK, EVIDENCE, ACCESS, TARGET, CONTRAST, CONTACT, BENEFIT, SHIFT, SURVEY, CODE, TEXT, AID, ATTRIBUTE, GRADE, FRACTION, PEAK, COMMENT, CLUSTER, FEEDBACK, OVERLAP, OVERVIEW, DEBATE, SPAN, FORMAT, DISRUPT. A presença do vocábulo/termo como verbo ou substantivo no FAV dependeu, em primeiro lugar da lista organizada por Nation, agrupada em função das famílias de palavras do BNC. Uma vez que o corpus do BNC continha etiquetas com as classes gramaticais, foi possível adotar o critério de frequência no corpus para selecionar a classe determinante do verbete.

A seguir, os verbos mais comuns na comunicação acadêmica, identificados em nossa pesquisa:

VARY, ANALYSE, REQUIRE, OCCUR, INVOLVE, INDICATE, REACT, SELECT, IDENTIFY, DISTRIBUTE, CREATE, DEFINE, CONCENTRATE, GENERATE, PROCEED, OBTAIN, INTERACT, ESTABLISH, AFFECT, LOCATE, ESTIMATE, ASSUME, CONTRIBUTE, MAINTAIN, CONSTRUCT, REMOVE, ACHIEVE, EVALUATE, COMPUTE, ASSESS, DERIVE, CONDUCT, CONSIST, PREDICT, REGULATE, ILLUSTRATE, TRANSFER, DETECT, PUBLISH, INCORPORATE, INVESTIGATE, ADAPT, DEMONSTRATE, CORRESPOND, EXPOSE, RELEASE, RELY, TRANSFORM, INTERPRET, INTEGRATE, PROMOTE, EVOLVE, RESTRICT, COMMUNICATE, CHALLENGE, ENHANCE, REVEAL, ISOLATE, EMERGE, CORRELATE, INDUCE, ASSIST, MONITOR, EXCLUDE, ENSURE, ORIENT, ACQUIRE, ATTACH, TRANSMIT, MODIFY, ADJUST, DISPLAY, CONSTITUTE, ELIMINATE, LABEL, PARTICIPATE, COORDINATE, ALTER, RECOVER, SEEK, ENABLE, INHIBIT, EXPAND, EXTRACT, IMPLEMENT, EQUIP, EXHIBIT, INSTRUCT, CONVERT, IMPLICATE, ACTIVATE, ASSIGN, IMPLY, RESIDE, SUM, SIMULATE, UTILISE, UNDERLIE, SUSTAIN, CAPTURE, FACILITATE, RETAIN, ENCOUNTER, CONFER, CONSTRAIN, ELEVATE, STIMULATE, SUBSTITUTE, CIRCULATE, DECLINE, TRACE, SUPPLEMENT, CONTRACT, ACCUMULATE, DIFFERENTIATE, EXCEED, OCCUPY, ACKNOWLEDGE, ACCOMPANY, CONSERVE, PRECEDE, ENGAGE, INSERT, GRANT, UNDERGO, REJECT, REVISE, CITE, DEVIATE, MANIPULATE, RESTORE, COMPLEMENT, MEDIATE, INTERVENE, CONSULT, ACCOMPLISH, AGGREGATE, COLLABORATE, CONFIRM, ATTAIN, PERSIST, COMPENSATE, COMMIT, COOPERATE, REGISTER, RECALL, DISPLACE, DEPICT, COMPRISE, EXPLOIT, DENOTE,

DISCHARGE, ALIGN, DESIGNATE, HIGHLIGHT, RESOLVE, DISPERSE, RESEMBLE, MANIFEST, GRADUATE, JUSTIFY, SWITCH, SUSPEND, RENDER, POSE, CONFORM, MAXIMISE, CAST, ACCOMMODATE, REINFORCE, UNDERTAKE, IMPOSE, INNOVATE, INFER, CLARIFY, CONDENSE, EXERT, CONVEY, WITHDRAW, TRIGGER, REFORM, CONVERGE, VERIFY, ALLOCATE, DIMINISH, PRESUME, EXECUTE, INSPECT, IMPRESS, ASSURE, COMPILE, RETRIEVE, EMBED, IMPAIR, TERMINATE, ANTICIPATE, REPLICATE, QUOTE, DISTORT, ENCODE

Da mesma forma que alguns substantivos podem ocorrer como verbos, mantendo a mesma grafia, alguns verbos podem ocorrer como substantivos, com a mesma grafia, como por exemplo ESTIMATE, CONSTRUCT, CONDUCT, TRANSFER, CHALLENGE, MONITOR, LABEL, EXTRACT, SUBSTITUTE, TRACE, SUPPLEMENT, CONTRACT, COMPLEMENT, REGISTER, HIGHLIGHT, MANIFEST, GRADUATE, SWITCH.

Os adjetivos identificados em nossa pesquisa foram os seguintes:

INDIVIDUAL, SIMILAR, MAJOR, AVAILABLE, SPECIFIC, COMPLEX, NORMAL, SIGNIFICANT, PRIMARY, POSITIVE, FINAL, INITIAL, STABLE, PREVIOUS, APPROXIMATE, APPROPRIATE, NEGATIVE, CONSTANT, ALTERNATIVE, MAXIMUM, INTENSE, INTERNAL, DISTINCT, ACCURATE, SUBSEQUENT, LINEAR, EXTERNAL, GLOBAL, CONSISTENT, RANDOM, CONSIDERABLE, ADULT, APPARENT, REVERSE, PRECISE, SUFFICIENT, PRIOR, VALID, RELEVANT, ADEQUATE, DIVERSE, UNIQUE, ANNUAL, EQUIVALENT, FUNDAMENTAL, SECURE, MINIMUM, PARALLEL, OBVIOUS, PRINCIPAL, DIFFUSE, PROFESSIONAL, CLASSIC, EVENTUAL, DYNAMIC, ULTIMATE, SIMULTANEOUS, AWARE, VISIBLE, BRIEF, UNIFORM, ABSTRACT, VIRTUAL, INTERMEDIATE, DOMINANT, COMPREHENSIVE, RADICAL, OPTIMAL, NEUTRAL, IDENTICAL, EMPIRICAL, FLEXIBLE, NOVEL, CRUCIAL, ADJACENT, TEMPORARY, CONTEMPORARY, ADVERSE, SUPERIOR, GRAPHIC, EXPLICIT, TEMPORAL, DEFINITE, PASSIVE, PREDOMINANT, VAST, DISCRETE, INTRINSIC, RATIONAL, INTEGRAL, QUALITATIVE, WIDESPREAD, HYBRID, VITAL, INHERENT, MUTUAL, ABNORMAL, PROMINENT, CONCRETE, LONGITUDINAL, WORLDWIDE, RIGID, MASSIVE, VIABLE, SOLE, PRELIMINARY, PREVALENT, SPONTANEOUS, REMOTE, INTERIOR, COMPACT, OVERALL

Conforme o exposto, identificamos 505 vocábulos/termos acadêmicos, divididos em seis categorias gramaticais (adjetivos, advérbios, conjunções, substantivos, preposições e verbos). Trata-se, portanto, do vocabulário mais frequente e melhor distribuído entre as áreas do conhecimento acadêmico-científico (Ciências Agrárias, Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias e Linguística, Letras e Artes).

Em sendo o vocabulário acadêmico um subgrupo da língua geral, vale o questionamento acerca dos resultados discutidos aqui. Dito de outra forma, questionamos se o vocabulário com alto coeficiente de uso identificado no ACE, de fato, demonstra um emprego mais expressivo na comunicação acadêmico-científica em comparação à língua geral. Até o

momento, todas as análises foram baseadas em uma exploração utilizando apenas os textos acadêmicos.

Consideramos a possibilidade de que o vocabulário acadêmico fundamental, na verdade, desempenhem um papel de destaque também na língua geral. Se assim for, não podem ser consideradas funcionais apenas na comunicação acadêmico-científica em si, mas, também, na língua de uma forma mais abrangente – perdendo, portanto, o estatuto de vocabulário tipicamente ocorrente na comunicação acadêmica.

Recorremos a um arquivo disponibilizado gratuitamente online, o BNC *World*, com uma lista de palavras, contendo 97.860.872 *tokens*. Uma vez que a lista não se encontrava lematizada, procedemos à lematização, de maneira a considerar todas as formas derivadas. Todos os dados foram tabulados individualmente em uma planilha do MS Excel 2018.

A maneira encontrada para obter um resultado confiável foi a comparação das frequências normalizadas das palavras ocorrentes nos dois *corpora*.

Usando o WST, calculamos as frequências normalizadas do BNC e comparamos com as frequências normalizadas já disponíveis do ACE, para cada grande área do conhecimento. Este procedimento é importante para embasar a afirmação de que os dados obtidos no *córpus* acadêmico são significantes. Até o momento, todas as análises foram realizadas intracórpus e não intercórpus.

Para tanto, realizamos o teste de soma de postos de Wilcoxon, que aponta se existe ou não diferença entre as duas amostras. Elaboramos o par de hipóteses, a saber a hipótese nula (H_0) e a hipótese alternativa (H_1):

H_0 : as frequências normatizadas do ACE são **iguais** frequências normatizadas do BNC;

H_1 : as frequências normatizadas do ACE são **diferentes** das frequências normatizadas do BNC.

Foi escolhido o nível de $\alpha = 0,05$. Isso quer dizer que o teste apresenta um nível de confiança de 95%. A decisão pela hipótese nula ou pela hipótese alternativa depende do valor de p obtido.

Confrontamos as todas as frequências normalizadas de cada grande área do ACE com as frequências também normalizadas do BNC *World*, obtendo o valor indicado na Tabela 2:

Tabela 2 – Frequências normalizadas no ACE vs. frequências normalizadas no BNC World

teste de soma de postos de Wilcoxon	valor de p
Ciências Agrárias vs. BNC <i>World</i>	p=0.038753*
Engenharias vs. BNC <i>World</i>	p=0.00550032*
Ciências Exatas vs. BNC <i>World</i>	p=1.871172e-06*
Ciências da Saúde vs. BNC <i>World</i>	p=0.0470366*
Ciências Sociais Aplicadas vs. BNC <i>World</i>	p=0.000302482*
Ciências Biológicas vs. BNC <i>World</i>	p=0.287958
Ciências Humanas vs. BNC <i>World</i>	p=0.0086954*
Linguística, Letras e Artes vs. BNC <i>World</i>	p=4.29862e-11*

Fonte: Dados da presente pesquisa

Os resultados obtidos nos quais $p < 0,05$ ($\alpha = 0,05$) indicam que os valores referentes as frequências do ACE demonstram ter uma diferença estatisticamente significativa quando comparadas as frequências no BNC *World*, o que leva à rejeição da hipótese nula e consequente aceitação da hipótese alternativa.

O asterisco indica que as amostras são diferentes em sete grandes áreas, nomeadamente nas Ciências Agrárias, Engenharias, Ciências Exatas, Ciências da Saúde, Ciências Sociais Aplicadas, Ciências Humanas e Linguística, Letras e Artes.

Apenas uma grande área não apresenta diferença quanto as suas frequências normalizadas e as frequências normalizadas do BNC *World*: as Ciências Biológicas. O valor de p nas Ciências Biológicas (p=0.287958) leva-nos a aceitar a hipótese nula, de que as amostras não apresentam diferença. A ausência de diferença entre os valores das frequências normalizadas das Ciências Exatas e as frequências normalizadas do BNC *World* não permite a interpretação de que essa área é menos acadêmica, em termos léxico-terminológicos. Na verdade, o que pode ser inferido é que a quantidade de vezes em que o vocabulário acadêmico fundamental é utilizado (que também faz parte da língua geral) nessas duas áreas aproxima-se da quantidade de vezes em que também é empregado em outros contextos comunicativos.

Vale ressaltar que as Ciências da Saúde com um valor de $p < 0,05$, mais especificamente p=0.0470366, em um primeiro momento, leva-nos a tomar a hipótese nula como sendo falsa e consequentemente aceitar a hipótese alternativa de que existe diferença entre os conjuntos. No

entanto, como o valor de p está muito próximo do valor de alfa ($\alpha = 0,05$), existe o risco de incorrermos no erro do tipo 2 – quando H_0 é falsa e aceita-se a hipótese nula, temos o erro do tipo 2. Por isso, preferimos manter as Ciências da Saúde em uma situação de dúvida.

Nosso último cálculo foi realizado no intuito de checar se a comparação entre as médias das frequências normalizadas do ACE (incluindo as 8 grandes áreas) e as frequências normalizadas do BNC World indicaria diferença ou semelhança. A Tabela 3 indica o resultado do teste para as frequências normalizadas no ACE vs. frequências normalizadas no BNC World:

Tabela 3 – Frequências normalizadas no ACE vs. frequências normalizadas no BNC World

teste de soma de postos de Wilcoxon	valor de p
médias das frequências normalizadas do ACE vs. frequências normalizadas do BNC World	$p=0.00295022^*$

Fonte: Dados da presente pesquisa.

Uma vez que o valor de p obtido ($p=0.00295022$) é inferior ao valor de alfa ($\alpha = 0,05$), devemos rejeitar a hipótese nula e aceitar a hipótese alternativa. Dessa forma, entendemos que existe uma diferença estatisticamente significativa entre as médias das frequências normalizadas do ACE (incluindo as 8 grandes áreas) e as frequências normalizadas do BNC World.

5 CONSIDERAÇÕES FINAIS

Entendemos que o vocabulário acadêmico – que desempenha um papel importante por sua presença em domínios de especialidade variados, não “pertencendo” a um domínio específico. Definimos o vocabulário acadêmico como sendo um conjunto de adjetivos, advérbios, conjunções, substantivos, preposições e verbos típicos da comunicação acadêmica. Trata-se de um repertório constituído de palavras pertencente à língua geral, sendo essencial na comunicação técnico-científica. Tal repertório constitui uma alta porcentagem em todos os *corpora* analisados. Ao servir como insumo na elaboração de material de divulgação acadêmica e técnico-científica, o vocabulário acadêmico proporciona traços específicos ao texto científico, fazendo com que seja mais parecido com outros textos do mesmo gênero. Neste sentido, pensamos que o vocabulário acadêmico pode ocorrer no discurso acadêmico em suas diversas subáreas e em diferentes registros.

As Ciências Exatas, as Ciências Humanas e a Linguística, Letras e Artes foram as três grandes áreas que se mostraram mais peculiares em termos de uso do vocabulário acadêmico. Essas áreas apresentaram peculiaridades em todas as comparações realizadas com as demais grandes áreas. Outras cinco áreas (Ciências Agrárias, Engenharias, Ciências da Saúde, Ciências Sociais Aplicadas e Ciências Biológicas) apresentaram uma diferença significativa quando comparadas a outras áreas. Por outro lado, mantinham um uso do vocabulário acadêmico semelhante com algumas áreas.

O último teste destinou-se a determinar se havia diferença entre as frequências normalizadas do vocabulário acadêmico em cada uma das grandes áreas e as frequências normalizadas do vocabulário acadêmico na língua geral (representada pelo *corpus BNC World*). Com exceção da área das Ciências Biológicas, as demais áreas apresentaram um valor de p inferior a 0,05 – consequentemente mostrando que existe diferença entre as grandes áreas e a língua geral no tocante ao uso do vocabulário acadêmico. Vale ressaltar que no caso das Ciências Sociais Aplicadas, o valor de p obtido ($p=0,047$) encontra-se bastante próximo do limite. Dessa forma, podemos entender que a área das Ciências Sociais Aplicadas encontra-se em uma posição que inicialmente poderia ser categorizada como diferente (em termos estatísticos), mas que, ao mesmo tempo, poderia encaixar-se na categoria sem diferença significativa, desde que haja um procedimento de arredondamento para 0,05 do valor de p . Preferimos incluir essa grande área em uma região de dúvida, na qual duas categorias seriam possíveis. Quando consideradas isoladamente, das 8 grandes áreas, 7 delas (Ciências Agrárias, Engenharias, Ciências Exatas, Ciências da Saúde, Ciências Sociais Aplicadas, Ciências Humanas e Linguística, Letras e Artes) diferenciam-se da língua geral no que diz respeito à intensidade de emprego do vocabulário acadêmico. Uma grande área (as Ciências Biológicas), analisada individualmente, não demonstra diferença no que se refere à intensidade de emprego do vocabulário acadêmico. Quando as 8 grandes áreas são manipuladas como um conjunto único, percebemos que diferenciam-se da língua geral no que diz respeito à intensidade de emprego do vocabulário acadêmico. Conforme o exposto, a análise independente das grandes áreas gera resultados distintos da análise que é executada levando em conta todas elas como um conjunto. Nessa situação, o resultado obtido indica que existe diferença significativa na frequência de emprego do vocabulário acadêmico dentro do conjunto, em comparação com a língua geral.

Os procedimentos descritos em nosso trabalho mostram que identificar um termo como acadêmico implica levar em consideração detalhes fundamentais como a frequência (absoluta e normalizada), a dispersão entre as áreas do conhecimento e o uso – obtido por meio da frequência normalizada multiplicada pelo coeficiente de dispersão. Portanto, o vocabulário acadêmico identificado em nosso trabalho por meio desses critérios apresenta, além de uma ocorrência significativa, uma distribuição homogênea nos diferentes *corpora*, não sendo identificado, por estas razões como pertencente a um domínio de especialidade específico.

A lista de vocabulário acadêmico identificada e analisada aqui pode servir como material de referência, especialmente voltada para o ensino de língua inglesa, uma vez que reflete uso real da língua na comunicação acadêmica.

Diante da lacuna que existia no tratamento do termo acadêmico, sob a ótica específica dos estudos terminológicos, a interface explorada aqui – Lexicologia e Linguística de Corpus – procurou trazer perspectivas positivas para os estudos linguísticos. Esperamos que a pesquisa por ora apresentada seja replicada, questionada e incrementada em pesquisas vindouras, quer por outros colegas pesquisadores ou por nossas próprias investigações futuras.

REFERÊNCIAS

- BAKER, M. Sub-technical vocabulary and the ESP teacher: an analysis of some rhetorical items in medical journal articles. *Reading in a foreign language*, Honolulu, v.4, n. 2, p. 91-105, 1988.
- BERBER SARDINHA, T. *Linguística de Corpus*. Barueri: Manole, 2004.
- BEVILACQUA, C. R. Por que e para que a Linguística de *Corpus* na Terminologia. In: TAGNIN, S.; BEVILACQUA, C. R. *Corpora na terminologia*. São Paulo: Hub Editorial, 2013. p. 11-28.
- BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics: investigating language structure and use*. Cambridge: Cambridge University Press, 2004.
- BIBER, D. Vocabulary use in classroom teaching and textbooks. In: _____. *A corpus-based study of spoken and written registers*. Philadelphia: John Benjamins Publishing Company, 2006.
- BIDERMAN, M. T. C. *Teoria lingüística: teoria lexical e computacional*. 2. ed. São Paulo: Martins Fontes, 2001.

COBB, T. Computing the vocabulary demands of L2 reading. *Language learning & technology*, [S.l], v. 11, n. 3, p. 38-63, 2007.

COXHEAD, A. A new academic word list. *TESOL Quarterly*, [S.l], v. 34, n. 2, p. 213-238, 2000.

DAYRELL, C. Corpora no ensino do inglês acadêmico: padrões léxico-gramaticais em abstracts de pós-graduandos brasileiros. In: VIANA, V.; TAGNIN, S. (Org.). *Corpora no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2010, p. 137-171.

DUDLEY-EVANS, T; SAINT JOHN, M. J. *Developments in English for Specific Purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press, 1998.

HYLAND, K. *English for academic purposes: An advanced resource book*. New York: Routledge, 2006.

_____; TSE, P. Is there an academic vocabulary? *TESOL Quarterly*, [S.l], v. 41, n. 2, p. 235-253, jun. 2007.

NATION, I.S.P. *Learning vocabulary in another language*. Cambridge: Cambridge University Press, 2001.

_____. *Como estruturar o aprendizado de vocabulário*. Tradução Cristiane Arruda. São Paulo: Special Book Services, 2003.

SCOTT, M. *WordSmith Tools: version 5*. Oxford: Oxford University Press, 2010.

SILVA, E. B.; BABINI, M. A preparação de material terminológico em língua inglesa por meio de ferramentas linguístico-computacionais. *Trabalhos em Linguística Aplicada*, Campinas, v. 50, n. 1, p. 119-132, jan./jun. 2011.

_____. A terminologia acadêmica nos textos científicos em língua inglesa: uma abordagem baseada em corpus. In: ISQUERDO, A. N.; SEABRA, M. C. T. C. (Org.). *As ciências do léxico: lexicologia, lexicografia, terminologia*. v. 6. Campo Grande: Ed. UFMS, p. 415-427, 2012.

SINCLAIR, J. *Corpus and text: basic principles*. In: WYNNE, M. (Ed.). *Developing linguistic corpora: a guide to good practice*. 2004. Disponível em: <<http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm>>. Acesso em: 31 ago. 2018.

SWALES, J. M. *Research genres: explorations and applications*. New York: Cambridge University Press, 2004.

SWALES, J. M.; FEAK, C. B. *Academic writing for graduate students: essential tasks and skills*. Ann Arbor: University of Michigan Press, 2004.