Revista da UNIFA
revistaeletronica.fab.mil.br
https://doi.org/10.22480/revunifa.2024.37.617
Artigo original
e-ISSN 2175-2567

A construção do corpus de artigos científicos de aviação: um estudo interdisciplinar

Fernanda Beatriz Caricari de Morais 0000-0001-6075-4101
Divisão de Ensino, Academia da Força Aérea, AFA, Pirassununga, SP, Brasil.

João Paulo Martins dos Santos 0000-0002-0957-7119
Divisão de Ensino, Academia da Força Aérea, AFA, Pirassununga, SP, Brasil.

RESUMO

Este artigo relata a experiência da construção de um corpus de artigos científicos da área de aviação, escritos em Língua Inglesa, e o tratamento linguístico-computacional dado pela Linguística de Corpus. A coleta foi realizada por meio de técnicas de programação computacional de raspagem de dados, o que permitiu coletar artigos de duas revistas eletrônicas: Air & Space Power Journal e Journal of Aviation/Aerospace Education and Research. O corpus é utilizado para pesquisas linguísticas, tendo como base a Linguística Sistêmico-Funcional (Halliday, 1994 e Halliday & Matthissen, 2004, 2014), que vê a língua como um sistema potencial de significados, em que o conceito de escolha é importante por permitir o estudo de regularidades lexicais, com implicações para a descrição linguística e para o ensino. Com o uso de ferramentas computacionais da Linguística de Corpus (Berber-Sardinha, 2000, 2004), é possível trabalhar com uma grande quantidade de textos, obtendo elementos que auxiliam na análise qualitativa dessas regularidades. Como resultado, tem-se um corpus de estudo que pode ser considerado de médio-grande porte (Berber-Sardinha, 2004), com mais de três milhões de palavras. Espera-se que a construção desse corpus fomente novas pesquisas linguísticas e estatísticas na área de aviação, especialmente de cadetes que participam de programas de iniciação científica e que redigem seus trabalhos de conclusão de curso.

Palavras-chave: *Corpus; Linguística de Corpus; Linguística Sistêmico-Funcional.*

The construction of a corpus of aviation scientific articles: an interdisciplinary study

ABSTRACT

This article presents the experience of building a corpus of scientific articles written in English in the field of aviation, and the linguistic-computational treatment given by the [... and the linguistic-computational treatment given by Corpus Linguistics. Data collection was performed using computer programming techniques for data scraping, which allowed the collection of articles from two electronic

*journals: Air & Space Power Journal and Journal of Aviation/Aerospace Education and Research. The corpus is used for linguistic research, based on Systemic-Functional Linguistics (Halliday, 1994 e Halliday & Matthiessen, 2004, 2014), that sees language as a potential system of meanings, in which the concept of choice is essential for allowing the study of lexical regularities, and has implications for both language description and language teaching. With the use of Corpus Linguistics computational tools (Berber-Sardinha, 2000, 2004), it is possible to work with a large number of texts, obtaining quantitative data that help in the qualitative analysis of these regularities. As a result, we have a study corpus that can be considered "[...] medium-large (Berber-Sardinha, 2004), with more than three million words. It is expected that the construction of this corpus will encourage new linguistic and statistical research in aviation, especially involving cadets who participate in scientific initiation programs and who draft their course completion papers.***Keywords:** knowledge management; Brazilian Air Force; SECI model; OKA method.*

Keywords: *Corpus; Corpus Linguistics; Systemic-Functional Linguistics.*

La construcción de un corpus de artículos científicos de aviación: un estudio interdisciplinario

RESUMEN

Este artículo relata la experiencia de construcción de un corpus de artículos científicos en el campo de la aviación escritos en Inglés y el tratamiento lingüístico-computacional proporcionado por la Lingüística de Corpus. La recolección se realizó mediante técnicas de programación computacional de extracción de datos, lo que permitió recopilar artículos de dos revistas electrónicas: Air & Space Power Journal y Journal of Aviation/Aerospace Education and Research. El corpus se utiliza para investigaciones lingüísticas, basándose en la Lingüística Sistémico-Funcional (Halliday, 1994 e Halliday & Matthiessen, 2004, 2014), que ve el lenguaje como un sistema potencial de significados, en el que el concepto de elección es importante ya que permite el estudio de regularidades léxicas, con implicaciones tanto en la descripción lingüística como en la enseñanza. Mediante el uso de herramientas computacionales de la Lingüística de Corpus (Berber-Sardinha, 2000, 2004), es posible trabajar con una gran cantidad de textos, obteniendo elementos que ayudan en el análisis cualitativo de estas regularidades. Como resultado, se obtiene un corpus de estudio que puede considerarse de tamaño mediano-grande (Berber-Sardinha, 2004), con más de tres millones de palabras. Se espera que la construcción de este corpus fomente nuevas investigaciones lingüísticas y estadísticas en el campo de la aviación, especialmente involucrando a estudiantes estudiantes que participan en programas de iniciación científica y que redactan su trabajo de finalización de cursos.

Palabras clave: *Corpus; Lingüística de Corpus; Lingüística Sistémico-Funcional.*

1 INTRODUÇÃO

Este estudo é parte do projeto intitulado "Análise de artigos acadêmicos de aviação escritos em língua inglesa com o suporte teórico-metodológico da Linguística Sistêmico-Funcional: descrição linguística subsidiando a elaboração de materiais didáticos"[1], que tem por objetivo analisar as características léxico-gramaticais de artigos de aviação, escritos em língua inglesa, de dois importantes periódicos da área: *Air & Space Power Journal* e *Journal of Aviation/Aerospace Education and Research*[2]. Esses periódicos foram escolhidos por reunirem em seus *websites* versões eletrônicas de artigos publicados nos últimos dez anos, o que facilita a construção do corpus de estudo, além de serem muito acessados e reconhecidos pela comunidade acadêmica como referências na área.

A escolha de estudo por essa temática, justifica-se pela experiência anterior com a análise da linguagem acadêmica de outras áreas (Morais, 2013, 2014, 2015 e 2016), bem como a experiência profissional como docentes da Academia da Força Aérea, o que motiva o melhor entendimento das práticas escritas, possibilitando a elaboração de materiais didáticos para os cadetes, subsidiando a criação de um curso de extensão para cadetes aviadores.

Inicialmente, o estudo previa a coleta de uma amostra de 40 textos de cada revista, selecionados aleatoriamente de seus *sites*, para a busca de padrões de uso. Esse número já permite fazer generalizações sobre padrões das escolhas léxico-gramaticais dessa área de estudo, porém seria necessário aumentar o corpus para verificar se esses padrões ocorrem em um corpus maior, obtendo, assim, generalizações mais confiáveis.

O quantitativo de textos coletados é importante para trabalhos que estão em conformidade com a visão da língua como um sistema de escolhas, sendo este probabilístico (Halliday, 1994 e Halliday & Matthiessen, 2004, 2014). O autor escolhe determinado item lexical, em detrimento de outros possíveis, combinando-o com outros itens, construindo significados de acordo com o contexto situacional e cultural de produção. No que se refere a esses contextos, pode-se dizer que os autores, pesquisadores da área de aviação, com formações diversas, moldam suas escolhas com base no gênero textual que escrevem - artigos científicos - para relatarem suas pesquisas para seus pares, a comunidade científica.

O contexto influencia nossas escolhas, pois se faz um uso funcional da linguagem. O conceito de escolha é fundamental para a compreensão de que a língua é um sistema potencial de significados. Para a Linguística Sistêmico-Funcional, cada escolha fornece uma série de novas opções que se especificam em redes de possibilidades e, por sua vez, criam significados. Essa teoria busca compreender, portanto, como os textos conseguem ou não expressar seus significados por meio das potencialidades da língua.

A Linguística Sistêmico-Funcional (LSF) dialoga com a Linguística de Corpus (LC) por também trabalhar dentro de uma visão de linguagem enquanto sistema probabilístico, possibilitando estudos sistemáticos de regularidades lexicais, descrevendo a linguagem em uso (Berber-Sardinha, 2004, p. 34).

Segundo Biber *et al. (*1998, p. 9), a abordagem via linguística corpus é útil, pois "quase todas as áreas da linguística podem ser estudadas a partir da perspectiva do uso, e a abordagem baseada em corpus fornece um conjunto de instrumentos particularmente eficaz para tais investigações".

[1] Submetido e aprovado pela SPPC em dezembro de 2022.

[2] Ambos possuem acesso livre.

A língua é vista como um sistema que se realiza pelos padrões semânticos criados com base na necessidade de os falantes interpretarem a experiência humana e interagirem com os outros. A língua não pode ser desconectada do uso, pois ela é social e só ocorre por meio de interações sociais reais. Dessa forma, para se estudar essas escolhas, é necessária a realização de uma análise linguística, para que se obtenham padrões linguísticos recorrentes no gênero textual estudado, sendo imprescindível a ampliação do corpus, procurando entendê-los em seus contextos de uso.

A realização dessa ampliação de forma manual seria um grande desafio, pois um tempo demasiado longo seria empregado apenas no processo de *download* de cada artigo de interesse. Em seguida, há o tempo de conversão em arquivo de texto (*.txt*) com exclusão manual de elementos não relevantes para que possam ser tratados por meio de ferramentas computacionais utilizadas na Linguística de Corpus. No sentido de reduzir o tempo empregado em tarefas mecânicas, o emprego de técnicas de programação computacional de raspagem de dados (em inglês *Web Scraping*) foi empregado. Ainda, o processo de conversão de textos pode ser automatizado por meio da utilização de pacotes computacionais específicos. Por fim, a análise pode ser potencializada por meio da utilização dos pacotes e bibliotecas computacionais com fins de análise de corpus. Pormenores adicionais e referências sobre algumas bibliotecas podem ser encontradas na seção 5.

Com o advento do computador, mais especificamente do computador pessoal, a concepção de corpus, bem como seu armazenamento e exploração, transformou-se, tendo em vista que os recursos atualmente oferecidos permitem o tratamento de uma quantidade grande de textos, possibilitando que muitas hipóteses de fenômenos linguísticos possam ser apuradas de forma mais simples e efetiva.

Pretende-se relatar, neste artigo, como o corpus de artigos científicos da área de aviação foi construído. Para isso, questões linguísticas são discutidas, como a concepção de corpus, a visão de língua adotada, como se deu a coleta e processamento inicial por meio do emprego da linguagem computacional *Python*. Para a coleta dos artigos, o processo de raspagem de dados foi utilizado; bibliotecas de computação científica foram utilizadas para o processamento inicial, ou seja, a conversão dos arquivos em formato *.pdf* para arquivos em formatos de texto *.txt*. A eliminação de elementos sem interesse linguístico, tais como *links*, números das linhas, espaços em branco, não foi realizada neste caso, pois o foco foi o processo de construção. Por fim, dados quantitativos foram obtidos por meio do programa *WordSmith Tools*, v. 5 (Scott, 2018), importante ferramenta utilizada para a análise linguística.

Espera-se que a construção desse corpus fomente o envolvimento de mais cadetes nas pesquisas linguísticas e estatísticas, tanto nos programas de Iniciação Científica da Academia quanto nos Trabalhos de Conclusão de Curso (TCC), pois há inúmeras possibilidades de estudos de diversos aspectos linguísticos, como, por exemplo, representações feitas nos artigos, modalidade (relação estabelecida entre autor-leitor) e organização textual. Almeja-se, futuramente, a ampliação do corpus para outros gêneros acadêmicos, presentes em periódicos de aviação, como resumo e resenha de livros. Por fim, é esperado que o artigo evidencie a interdisciplinaridade das áreas de Computação Científica, Linguística e Estatística como um elemento de destaque do processo.

2 A CONCEPÇÃO DE CORPUS PARA A PESQUISA

Na Linguística, há várias definições de corpus, muitas delas ligadas à noção de corpus como uma coletânea de textos naturais, isto é, escolhidos para representar uma variedade de linguagem, podendo ser utilizado na pesquisa linguística.

Para Sanchez (1995, p. 8-9), corpus é um conjunto de dados linguísticos sistematizados por critérios, sendo representativos na totalidade do uso linguístico, tratados por computador para serem descritos e analisados.

Berber-Sardinha (2008, p. 17) destaca a importância de o corpus ser planejado e concretizado de acordo com critérios linguísticos de seleção. O autor define corpus como

> um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (Berber-Sardinha, 2008, p. 19.)

Segundo McEnery & Wilson (1996), o conceito de corpus possui 4 características importantes: a) amostragem e representatividade: um corpus deve ser uma amostragem suficiente da língua ou variedade de língua que se quer pesquisar, para se obter o máximo de representatividade; b) tamanho finito: todo corpus tem um tamanho finito de palavras, isto é, 500 mil palavras, 1 milhão de palavras, etc.; c) formato eletrônico com vantagens consideráveis, como a facilidade de manipulação de dados, bem como a inserção de dados extras; d) referência padrão: um corpus possui uma referência padrão para a variedade de língua que ele representa, sendo disponível para outras pesquisas.

No que diz respeito ao formato eletrônico do corpus, é importante destacar que o trabalho com corpus mudou com a tecnologia, assim como a sua forma de armazenamento e exploração. Os recursos atuais com o auxílio computacional permitem que uma quantidade imensa de textos possa ser processada rapidamente, o que suscita de muitas hipóteses sobre determinados fenômenos linguísticos uma rápida e eficiente checagem. Essa nova maneira de pesquisar permite a observação e a descrição pormenorizada de particularidades da língua, o que antes não era possível com apenas recursos manuais.

Para Trask (2004), a pesquisa com corpus permite que o pesquisador realize checagens precisas sobre o real uso linguístico, com base em textos (orais ou escritos) autênticos de falantes/escritores reais, fornecendo generalizações confiáveis e isentas de comentários, julgamentos e manipulações, pois é baseado em dados quantitativos e qualitativos. Somente com o uso de corpus é possível ter uma descrição dos usos da língua de forma clara e objetiva. Parte-se do pressuposto de que a língua é um sistema probabilístico que permite explicar, em contextos reais de comunicação, por que uma escolha é preferida em detrimento de outra.

Para a criação do corpus de artigos acadêmicos de aviação, foi necessário ter em mente um conjunto de requisitos que impactam na confiabilidade da pesquisa baseada em corpus,

como autenticidade (uso de textos reais), representatividade (ser representativo para o que se quer analisar), balanceamento (escolhas adequadas ao propósito), amostragem, diversidade e tamanho. Esses requisitos estão contemplados na construção do corpus, pois representam uma amostra real de língua em uso, de um mesmo gênero textual e seu tamanho condiz com a representatividade desse gênero nessa área de conhecimento, tendo em vista que os artigos foram coletados de dois dos mais importantes periódicos da área, reconhecidos mundialmente.

No que se refere à representatividade de um corpus, pode-se dizer que a característica diretamente relacionada é a extensão do corpus, ou seja, para um corpus ser representativo, ele deve ser o maior possível. A representatividade está intimamente associada à probabilidade.

Conforme mencionado anteriormente, a língua é um sistema probabilístico, em que alguns usos são mais frequentes que outros. Em relação ao léxico, é possível diferenciar as palavras entre as de maior ocorrência e as de menor ocorrência. Caso se estude uma palavra com ocorrência rara, é necessário que se coletem muitos textos, para que se tenha uma quantidade grande de palavras. Dessa forma, quanto maior a quantidade de palavras, maior a probabilidade de ocorrerem palavras de frequência baixa (Berber-Sardinha, 2008, p.23).

Em suma, ao tornar o corpus maior possível, este irá aproximar-se, ao máximo, da comunidade linguística que o produziu, sendo, portanto, uma amostra mais representativa da área de estudo de aviação, no caso deste estudo. Sendo a língua um sistema probabilístico, podem ser estabelecidos padrões mais comuns e menos comuns em determinado contexto de uso. A Linguística de Corpus permite conhecer as probabilidades de ocorrências de traços lexicais, estruturais, pragmáticos e discursivos.

Além do número de palavras, outro fator importante no que se refere à representatividade é o número de textos (aplica-se a corpora de textos específicos). No caso deste estudo, que investiga um único gênero, o artigo científico, de uma área específica, aviação, um número de artigos maior possibilita que esse gênero seja representado de forma satisfatória.

O corpus precisa, portanto, ser representativo e adequado aos interesses da pesquisa. Mais especificamente, é necessário também ser singular para o tipo de pesquisa, podendo ajudar a fornecer respostas para as perguntas de pesquisa. É importante dizer que, mesmo que se tenha corpora específicos, eles podem ser socializados à comunidade acadêmica-científica para que sejam vistos como dados verificáveis, podendo servir para outras pesquisas ou, ainda, integrar corpora maiores.

3 A LINGUÍSTICA DE CORPUS

A Linguística de Corpus (doravante LC) é uma abordagem empirista, por isso vê a linguagem como um sistema probabilístico. Pode-se dizer, ainda, que a LC se preocupa com a criação e análise de corpora, ou seja, com textos armazenados de forma eletrônica que permitem o trabalho com ferramentas computacionais. Essa abordagem tem inovado a forma como se investiga a linguagem, por fornecer maneiras inovadoras de análise e permitir análise de quantidade grande de dados.

Sendo a LC empirista, pode-se dizer que ela observa dados da linguagem organizados na forma de um corpus, para serem analisados com base na visão probabilística da linguagem,

sob a observação dos dados de contextos concretos, produzidos por usuários em situações reais de comunicação.

Berber-Sardinha (2008) argumenta que, embora muitos traços linguísticos sejam possíveis teoricamente, não ocorrem com a mesma frequência. As possibilidades da estrutura não se realizam todas com a mesma frequência. Essa diferença na frequência entre os traços não é aleatória, ou seja, há uma correlação entre características linguísticas e situacionais, condicionadas pelos contextos de uso. Biber (1993) com base em dados, teoriza que conjuntos de dados linguísticos variam, sistematicamente, com relação a textos típicos de contextos comunicativos específicos, provando que a variação não é aleatória, mas condicionada ao contexto.

O conceito de aleatoriedade se contrapõe ao de padronização, pois que, evidenciada esta pela alta ocorrência, estruturas ou itens lexicais se repetem de forma sistemática, mostrando que se tem, na verdade, um padrão léxico-gramatical. Há regularidade nos padrões da linguagem, tendo probabilidade de ocorrências nos traços linguísticos. É a verificação empírica que revela a frequência do uso, materializado por usuários da língua em determinados contextos. Dessa forma, é a partir da observação da frequência do uso e da descrição linguística que se pode estimar a probabilidade teórica.

A busca por padrões regulares é descrita como uma preocupação das pesquisas baseadas em corpus. É fundamental observar as regularidades lexicais e como elas se associam, sendo que, segundo Huston & Francis (2000, p. 37), "os padrões da palavra podem ser definidos como todas as palavras e estruturas que são regularmente associadas à palavra e que contribuem para o seu significado. Um padrão pode ser identificado se uma combinação de palavras ocorre com relativa frequência"[3].

Essas padronizações, vistas como regularidades que ocorrem com frequência sistemática, podem ser descritas por meio dos conceitos de colocação, coligação e prosódia semântica. Para a pesquisa proposta, o interesse nos padrões recorrentes se concentra na associação de itens lexicais (colocação) e associação entre itens lexicais e gramaticais (coligação), o que justifica a importância de analisar-se um corpus representativo para a obtenção de generalizações linguísticas confiáveis sobre as características linguísticas dos artigos da aviação.

A teoria da linguagem que se relaciona com a perspectiva da LC é a LSF, por terem em comum, como abordado anteriormente, a visão de linguagem e a inclinação empirista, com foco na descrição linguística baseada nos usos. Para a LSF, o conceito de escolha é fundamental, pois se escolhe dentro de um sistema que é potencialmente gerador de sentidos, fazendo proposições para a construção de significados específicos. Os conceitos básicos da LSF são apresentados na seção seguinte.

4 A VISÃO DA LINGUAGEM DA LINGUÍSTICA SISTÊMICO-FUNCIONAL

Conforme mencionado, a visão probabilística, assim como o conceito de escolha, é fundamental para a LSF. É o sistema de escolhas semânticas que permite a construção de significados na linguagem. Nessa abordagem teórica, a linguagem é social e

[3] Tradução nossa de "*the patterns of word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of word occurs relatively frequently*."

indissociável da experiência humana, pois permite o estudo das maneiras pelas quais as pessoas usam a linguagem para atingir certos objetivos em situações específicas dentro da sociedade. A língua é um sistema de possibilidades que permite a seus usuários a construção de significados.

A preocupação da LSF é descrever como a língua está estruturada para o uso em diferentes contextos (Eggins, 1994, p. 23). Como se mostrou, as escolhas léxico-gramaticais não são aleatórias, mas condicionadas pelo contexto, motivadas pelas relações sociais estabelecidas. Para Thompson (1996), uma escolha pode determinar ou ser determinada por outra, conforme os elementos ao seu redor.

Mais especificamente, a LSF é uma abordagem semântico-funcional, sendo que o uso da língua é visto como sendo funcional; a função da linguagem é realizar significados; esses significados são influenciados pelo contexto sociocultural e o processo de uso da linguagem é semiótico, ou seja, é um processo de realização de significado por meio de escolhas. É uma teoria linguística que se insere no contexto interdisciplinar da Linguística Aplicada, pelo seu interesse em compreender como a linguagem é usada na vida social diária das pessoas. É interdisciplinar, ou ainda, mestiça (Moita-Lopes, 2006, p. 14), por estar centrada no estudo das questões sociais, em que a linguagem tem um papel central no diálogo com outras áreas para melhor compreender os fenômenos linguísticos.

Tendo em mente que a linguagem é uma prática social, ao estudar a linguagem, estudam-se a sociedade e a sua cultura. As práticas discursivas não são neutras, envolvem escolhas, atravessadas por questões ideológicas e políticas, que provocam diferentes efeitos no mundo social. "A pesquisa é um modo de construir a vida social ao tentar entendê-la" (Moita-Lopes, 2006, p. 97).

O diálogo com outras áreas do conhecimento, assim como ocorre neste trabalho, permite o aprofundamento das questões relativas ao uso, compreendendo como estudiosos da área de aviação fazem suas escolhas linguísticas para estruturarem seus textos e divulgarem suas pesquisas.

Pensando nesse contexto de estudo e divulgação científica, a LSF permite um olhar para além das estruturas linguísticas, compreendendo as escolhas para o desempenho de funções determinadas na vida social. O texto, entendido desde a unidade mínima até a mais ampla, encontra-se inserido em determinado contexto de cultura e determinado contexto de situação.

O contexto de cultura é caracterizado como aquele que dá suporte à compreensão da forma específica em que as diferentes culturas utilizam a linguagem, com base nos interesses comunicativos e por meio das suas escolhas lexicais para estruturação do discurso nas interações. Esse contexto de cultura pode ser identificado, também, por questões mais amplas, assim como fatores histórico-sociais que demarcam o tempo e permeiam o texto, atribuindo um significado mais específico.

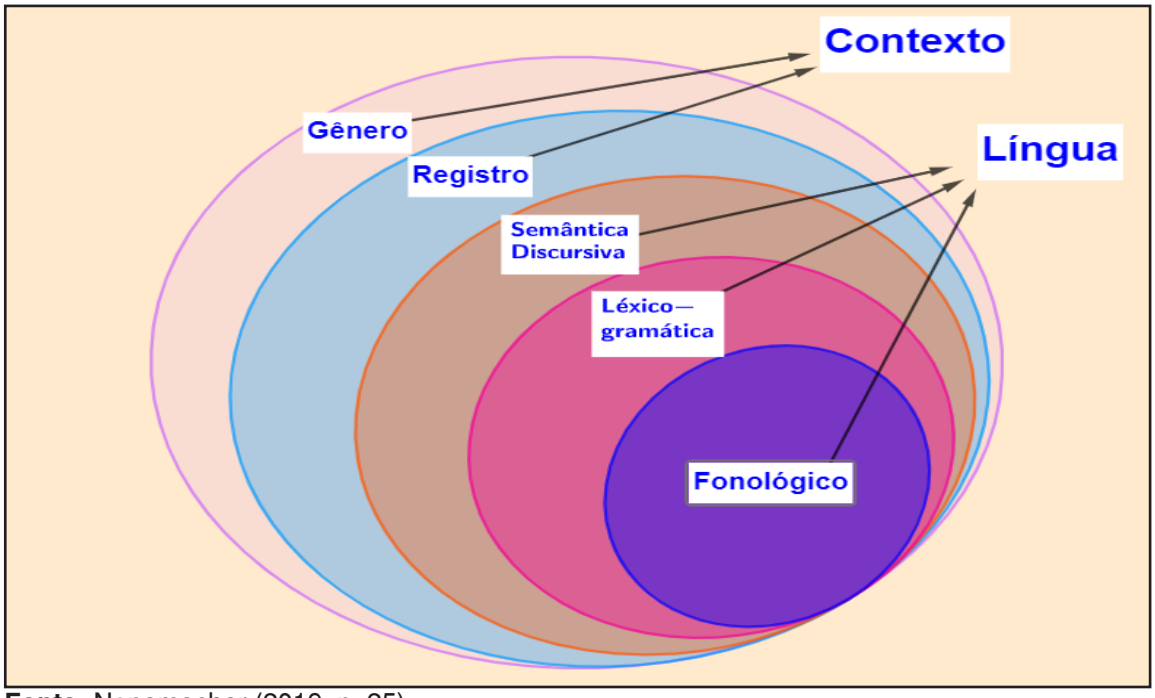
O contexto de situação, em contraste com o contexto de cultura, pode ser caracterizado nos planos real e abstrato, dada a capacidade de atender, ante o campo linguístico, as variações desenvolvidas entre os diferentes contextos culturais. É

necessário ter em mente o tempo em que ocorrem, apresentando, dessa forma, as variações de registro, chamadas de variáveis de campo (que dizem respeito ao assunto do texto), relações (caracterizadas como as interações e representações linguísticas entre os participantes do texto) e modo (o formato em que o texto é apresentado). Essas variáveis se ligam diretamente às três funções fundamentais que a linguagem é capaz de desempenhar (Gouveia, 2009), correspondendo, assim, às três metafunções da LSF, à metafunção ideacional, à metafunção interpessoal e à metafunção textual.

É possível estabelecer relação entre os dois tipos de contexto com os construtos de registro e de gênero. O primeiro, de registro, é a maneira como se usam as estruturas linguísticas apropriadas a certos tipos de contexto, de modo que textos com características linguísticas comuns pertencem ao mesmo registro e, assim, possuem o mesmo contexto de situação; o segundo, de gênero, representa não apenas essa relação entre sistema e contexto, mas também o objetivo social e comunicativo que os usuários da língua pretendem atingir por meio da linguagem, também relacionado ao contexto de cultura.

Esses contextos são fundamentais para a teoria, por permitir a visualização de um sistema em dois níveis: linguístico e extralinguístico. Os contextos de situação e cultura estão no segundo nível, ainda não materializado em forma de linguagem. No nível linguístico está a lexicogramática, que é, basicamente, um sistema de fraseados, composto pela sintaxe, pelo léxico e pela morfologia. A Figura 1 ilustra esses conceitos descritos.

Figura 1 - Estratos de um sistema linguístico.



Fonte: Nonemacher (2019, p. 25).

Os estratos são formas de olhar para o sistema da língua que, integrados, funcionam em conjunto no ato discursivo. Ter os estratos, separadamente, permite ao pesquisador analisar as diferentes características linguísticas. No que diz respeito ao contexto desta pesquisa, o contexto de cultura (acadêmico-científico) e a área de estudos vão influenciar as escolhas linguísticas do gênero em questão. Para Martin (1992, p. 505), o gênero é "um processo social organizado em etapas e orientado para um objetivo, realizado através do registro". Isso mostra que ele é um fenômeno social que ocorre no domínio do contexto de cultura, sendo realizado no contexto de situação por meio de um tipo de registro, realizando-se em texto.

5 REALIZAÇÃO DA COLETA - RASPAGEM DE DADOS

O processo de captura de textos para a construção do corpus de artigos científicos de aviação utilizou a linguagem *Python* por meio do Ambiente colaborativo *Google Collaboratory* ou, simplesmente, Google Colab. Colab é, resumidamente, descrito da seguinte maneira:

> *Google Colaboratory more commonly referred to as "Google Colab" or just simply "Colab" is a research project for prototyping machine learning models on powerful hardware options such as GPUs and TPUs. It provides a serverless Jupyter notebook environment for interactive development. Google Colab is free to use like other G Suite products. (Bisong, 2019, p. 57)*[4].

Foi feita a opção pela linguagem *Python* devido à disponibilidade de bibliotecas especializadas, tais como: *nltk (*em inglês Natural Language Toolkit (Bird; Loper; Klein, 2009)), para processamento de linguagem natural; *pdfminer six (*Pdfminer.six, 2023*)*, para processamento de arquivos em *Portable Document Files* (*.pdf*). Também estão disponíveis bibliotecas de manipulação *NumPy* (Harris *et al.*; 2020), os recursos gráficos de *Matplotlib* (Hunter, 2007), as bibliotecas com finalidade estatística, disponíveis, por exemplo, em *SciPy* (Virtanen *et al.*; 2020) e, por fim, a biblioteca *BeautifulSoup* (Richardson, 2023) para processamento de textos em formato *Hypertext Markup Language* (*.html*) e *eXtensible Markup Language* (*.xml*). Segundo a documentação disponível em (Richardson, 2015*), "Beautiful Soup* é uma biblioteca *Python* para extrair dados de arquivos HTML e XML. Funciona com seu analisador favorito para fornecer formas idiomáticas de navegar, pesquisar e modificar a árvore de análise. Geralmente economiza horas ou dias de trabalho dos programadores"[5].

Python é uma linguagem popular empregada em fins diversos, cuja facilidade de aprendizagem é um dos elementos mais essenciais (Moreira Filho, 2021, p. 27). Por fim, as argumentações relacionadas a seu emprego no artigo podem resumidas no seguinte trecho:

> Atualmente, Python é uma das linguagens de programação mais populares, servindo a diversos propósitos. É considerada também a língua franca para conectar tecnologias, soluções e aplicações científicas. O seu poder computacional e a disponibilidade de bibliotecas para o carregamento de dados, visualização, estatística, processamento de línguas naturais, aprendizado de máquina e tratamento de imagens, além de sua interatividade com o código, são suas características mais atrativas (Moreira Filho, 2021, p.27).

[4] Tradução nossa: "O *Google Colaboratory*, mais conhecido como "*Google Colab*" ou simplesmente "*Colab*", é um projeto de pesquisa para prototipar modelos de aprendizado de máquina em opções de hardware poderosas, como GPUs e TPUs. Ele fornece um ambiente de *notebook Jupyter* sem servidor para desenvolvimento interativo. O Google Colab é gratuito para uso, como outros produtos do G Suite".

[5] Tradução nossa de "*Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work*".

Conforme mencionado anteriormente, os artigos utilizados na construção do corpus estão disponíveis em *Air and Space Power Journal Archives* e *Journal of Aviation/Aerospace Education & Research*. O processo foi dividido em quatro partes:

i.) análise da página e detecção dos elementos de interesse;
ii.) *download* dos arquivos e aplicação de filtros;
iii.) pré-processamento dos arquivos; iv.) análise linguística.

A análise da página requer a identificação dos elementos de interesse. No caso do *Air and Space Power Journal Archives,* cada volume publicado entre os anos de 1947 e 2011 contém um conjunto de elementos agrupados em um único arquivo em formato *.pdf,* provavelmente devido à utilização da versão impressa como um dos principais meios de divulgação. Dessa forma, foi necessário fazer um recorte, excluindo-se os números mais antigos, pois também foi observado que não havia artigos científicos, mas outros gêneros textuais. A observação desses números permite dizer que a referida publicação se tornou mais acadêmica ao longo dos anos, por meio de pesquisas em formato de artigos científicos e críticas de livros da área em formato de resenhas acadêmicas. No caso do *Journal of Aviation/Aerospace Education & Research*, não foi necessário estabelecer um recorte dos artigos.

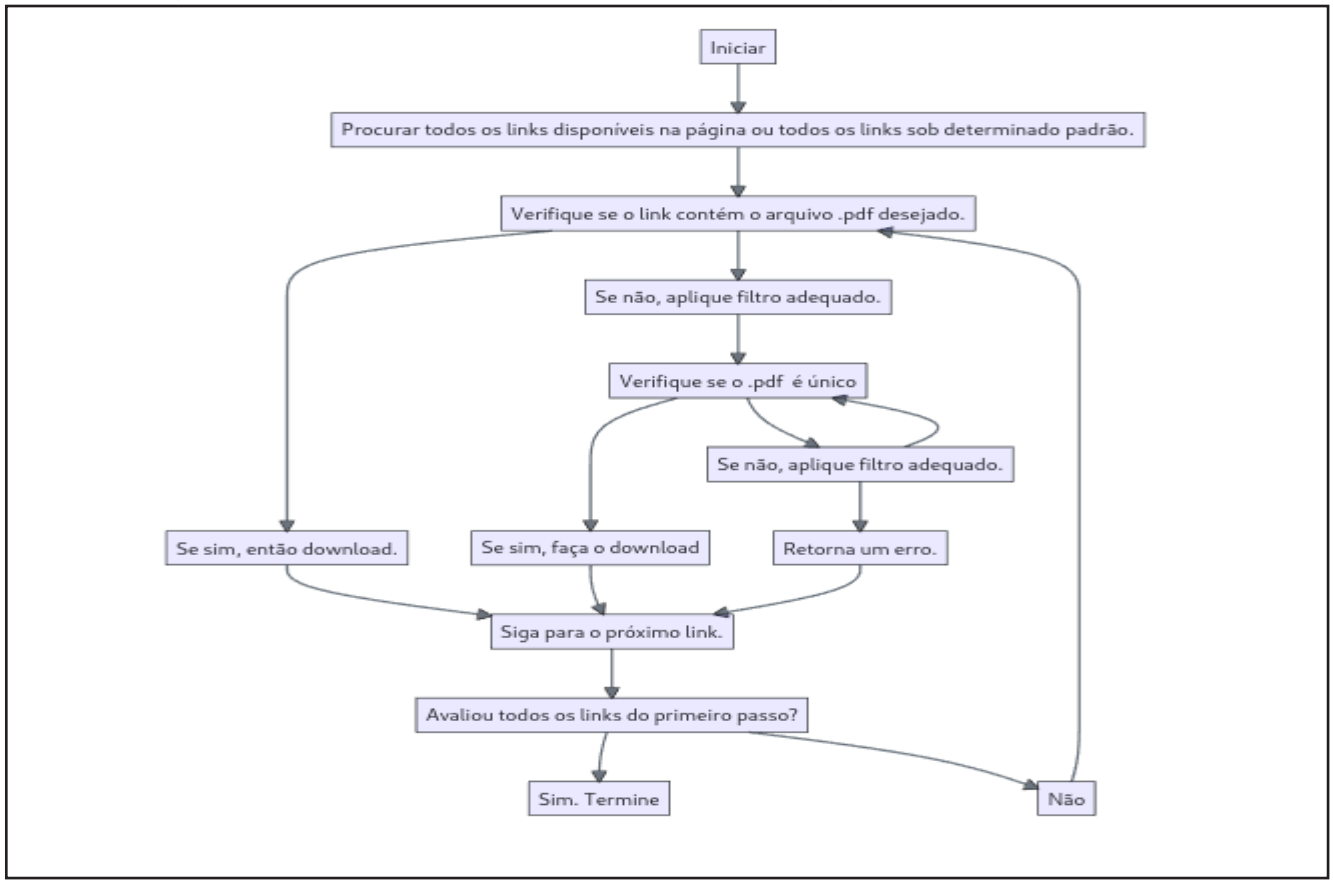
5.1 Procedimento e discussões

O procedimento foi baseado na linguagem *Python* com utilização da biblioteca *Beautiful Soup*. De forma geral, o algoritmo de busca dos artigos consiste nos seguintes passos:

a) Busque todos os *links* disponíveis na página ou todos os *links* sob determinado padrão. Por exemplo, todos os *links* disponíveis em h3.
b) Se cada um dos *links* contém o arquivo *.pdf* desejado, então efetue *download*.
c) Se não, execute um filtro para obter a página que contém o arquivo de interesse.
d) Se o *.pdf* disponível em 3. é único, então faça *download*. Se não, utilize um filtro para capturar o arquivo *.pdf* desejado. Caso o *link* retorne um erro ao executar o *download*, o processo passa para o próximo *link* disponível.

Os itens anteriores são componentes do processo, mas não foram executados de forma algorítmica. Todos os detalhes da execução foram desenvolvidos, utilizando o conceito de funções *Python*. O diagrama da Figura 2 descreve, pormenorizadamente, o processo de busca e detecção dos artigos de interesse.

Figura 2 - Diagrama descritivo dos passos para identificação dos artigos de interesse.



Fonte: Os autores.

Em alguns pontos do diagrama foram necessárias as adaptações específicas ao processo de busca. Essas adaptações foram traduzidas por modificações nas funções *Python* utilizadas.

Principais pontos abordados:

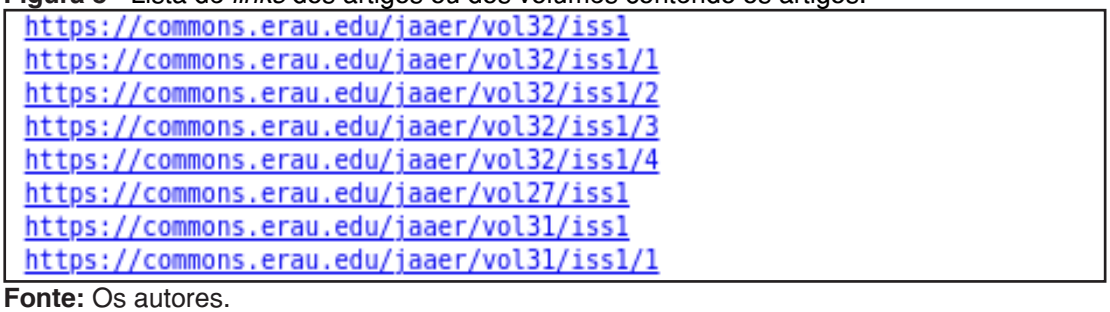
Caso 1: artigos de *Air & Space Power Journal*

1. O conteúdo da página https://www.airuniversity.af.edu/ASPJ/Articles/ foi obtido por meio da função *requests.get(url)*. Em seguida, todos os elementos do tipo *ul* foram detectados por meio da função *soup.find*.
2. Os resultados de 1 foram passados para a função *Beautiful Soup* com o *html.parser*. Os *links* da página foram filtrados apenas para obter os arquivos da forma em *.pdf* disponíveis e os resultados, adicionados a uma lista.
3. O processo em 1 e 2 é repetido para todas as páginas disponíveis por meio de processo iterativo.
4. Foram efetuados os *downloads* de todos os arquivos da lista gerada em 3.
5. O filtro foi aplicado aos arquivos do tipo *feature*.
6. Todos os arquivos foram baixados e salvos em uma pasta; cada um dos arquivos recebeu designação dada pelo respectivo *link*. Os *links* que retornaram erros foram descartados do processo.

Caso 2: artigos de Journal of Aviation/Aerospace Education & Research

1. O conteúdo da página https://commons.erau.edu/jaaer/all_issues.html foi obtido por meio da função *requests.get(url)*.
2. Os resultados de 1 foram passados para a função *BeautifulSoup* com o *html.parser*. Para cada um dos *links* disponíveis na página, verificou-se se a palavra "iss" pertence ao *link*. Os resultados foram adicionados em uma lista.
3. Cada um dos *links* anteriores possui um conjunto de artigos de interesse e *links* que não são de interesse. Basicamente, o processo foi reiniciado para se obter uma lista de *links*, em que cada um contenha o arquivo de interesse. Os resultados são ilustrados na Figura 3.

Figura 3 - Lista de *links* dos artigos ou dos volumes contendo os artigos.



Fonte: Os autores.

4. A seguir, os *links* com as designações de volumes e número foram obtidos por meio de expressão regular *iss\d+/\d+*. Uma nova lista de *links* foi gerada.
5. A lista anterior foi, então, utilizada para *download* dos arquivos em *.pdf*, visto que todos os *links* se referem aos artigos de interesse. Todos os arquivos baixados foram salvos em uma pasta; cada um dos arquivos recebeu a designação dada pelo respectivo *link*. Os *links* que retornaram erros foram descartados do processo.

5.2 Pré-processamento dos arquivos

O pré-processamento consistiu nos seguintes passos:

1. Conversão dos arquivos em formato *.pdf* para o formato *.txt* utilizando a biblioteca *pdfminer.six (Pdfminer.six, 2023)*.

   a. Após a obtenção das listas de artigos, cada um dos textos foi convertido para o formato *txt*.
2. Cada um dos arquivos *.txt* foi salvo em uma pasta com o nome do *link* de *download*. Um total de 132 arquivos foram obtidos para *Air & Space Power Journal* e 677 arquivos foram obtidos para *Journal of Aviation/Aerospace Education and Research*.

A análise dos elementos linguísticos é descrita a seguir. Nesse caso, optou-se por utilizar o *software WordSmith* v. 8 (Scott, 2018), ferramenta reconhecida no meio acadêmico para análise linguística. A seguir, estão alguns aspectos preliminares da análise da construção efetuada.

5.3 Uso do programa *WordSmith Tools*

Após a coleta do corpus de estudo, o tratamento computacional foi realizado pelo programa *WordSmith* v. 8 (Scott, 2018) para a extração de dados quantitativos sobre os padrões linguísticos mais frequentes, possibilitando fazer generalizações sobre como os artigos científicos da área de aviação são escritos por análise qualitativa, tendo como base o contexto de uso desses padrões.

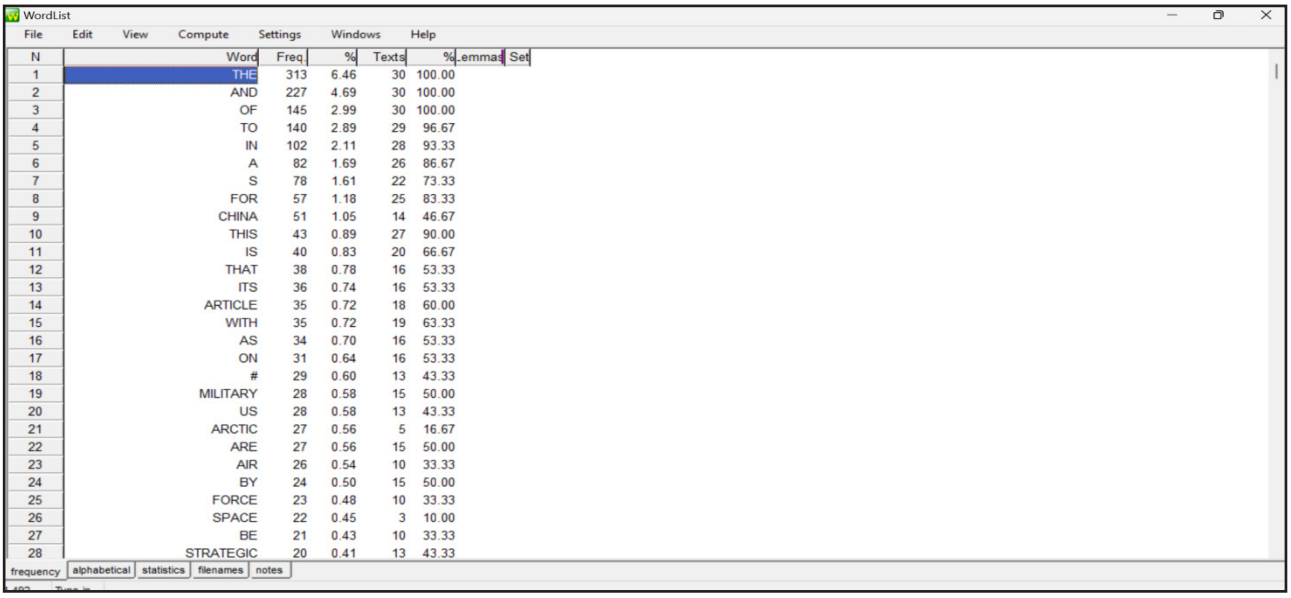
Basicamente, pode-se dizer que o programa é um conjunto de ferramentas desenvolvidas para a análise linguística. Foi idealizado pelo linguista Mike Scott e vendido pela Oxford University Press. É utilizado para estudar diversos fenômenos da linguagem, por diferentes profissionais em escala mundial.

De acordo com Berber-Sardinha (2004, p. 86), o programa permite ao analista uma série de recursos que são muito úteis e poderosos para a análise de vários aspectos da linguagem, tais como: composição lexical, a temática dos textos selecionados e a organização retórica e composicional de gêneros discursivos. Neste artigo, uma visão geral dos principais recursos é explorada, com especificação dos usos estabelecidos nas pesquisas linguísticas *Wordlist* e *Concord*.

5.3.1 *WordList*

Esta ferramenta possibilita a criação de listas de palavras, sendo ordenadas alfabeticamente, por ordem de frequência das palavras, com a palavra mais frequente encabeçando a lista, como se pode observar na Figura 3.

Figura 3 - WordList obtida com base no corpus de artigos da aviação.



N	Word	Freq	%	Texts	%	Lemmas	Set
1	THE	313	6.46	30	100.00		
2	AND	227	4.69	30	100.00		
3	OF	145	2.99	30	100.00		
4	TO	140	2.89	29	96.67		
5	IN	102	2.11	28	93.33		
6	A	82	1.69	26	86.67		
7	S	78	1.61	22	73.33		
8	FOR	57	1.18	25	83.33		
9	CHINA	51	1.05	14	46.67		
10	THIS	43	0.89	27	90.00		
11	IS	40	0.83	20	66.67		
12	THAT	38	0.78	16	53.33		
13	ITS	36	0.74	16	53.33		
14	ARTICLE	35	0.72	18	60.00		
15	WITH	35	0.72	19	63.33		
16	AS	34	0.70	16	53.33		
17	ON	31	0.64	16	53.33		
18	#	29	0.60	13	43.33		
19	MILITARY	28	0.58	15	50.00		
20	US	28	0.58	13	43.33		
21	ARCTIC	27	0.56	5	16.67		
22	ARE	27	0.56	15	50.00		
23	AIR	26	0.54	10	33.33		
24	BY	24	0.50	15	50.00		
25	FORCE	23	0.48	10	33.33		
26	SPACE	22	0.45	3	10.00		
27	BE	21	0.43	10	33.33		
28	STRATEGIC	20	0.41	13	43.33		

Fonte: Os autores.

Em cada janela diferente, no campo inferior esquerdo, uma dessas listas é apresentada em ordem alfabética e de frequência. É oferecida, ainda, uma lista estatística relativa aos dados usados para produção das listas. Na lista estatística, podem ser obtidos dados importantes, como número de itens (ou ocorrências), chamados de *Tokens*, número de formas ou vocábulos (*Types*), extensão de palavras (em número de letras); relação tipos/ocorrências; número e extensão de sentenças e parágrafos, para textos individuais e para todo o corpus.

A *Wordlist* permite que o pesquisador verifique os dados estatísticos, construa hipóteses iniciais sobre as preferências de uso e constate fatos interessantes sobre as escolhas lexicais, além de ser um quadro geral da forma como as palavras se portam na área de aviação.

Com base nessa ferramenta, foi possível extrair os dados quantitativos do corpus de estudo, conforme o seguinte quadro:

Quadro 1 - Dados quantitativos dos periódicos.

Periódico/ Dados extraídos da *WordList*	*Air & Space Power Journal*	*Journal of Aviation/Aerospace Education and Research*
Número de artigos coletados	132	617
Total de palavras (*tokens*)	1.012.872	2.468.761
Palavras diferentes (*types*)	39.462	71.103
Número de orações	35.200	139.008

Fonte: Os autores.

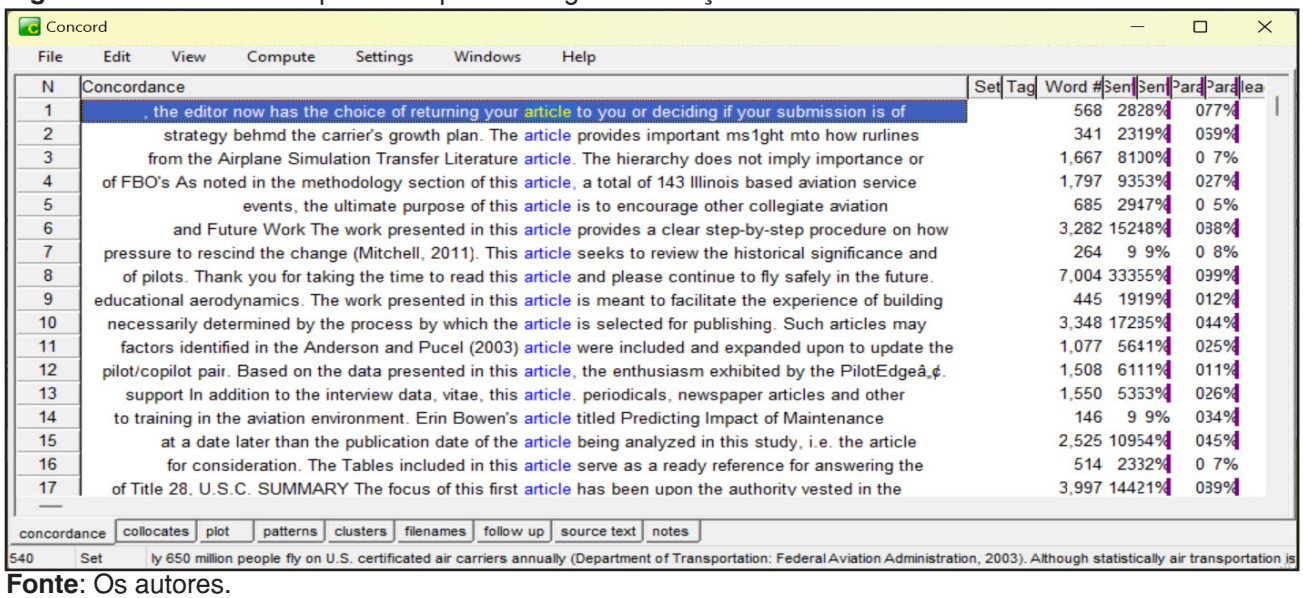
De acordo com as definições de tamanho de corpus (Berber-Sardinha, 2004, p. 26), pode-se constatar que o corpus construído é médio-grande, em total de 3.481.639 de palavras. No que se refere à representatividade, tem-se uma amostra considerável da comunidade acadêmica da área de aviação.

Por meio desses dados estatísticos, são geradas listas de concordância, tendo como palavra de busca as de maior frequência no corpus de estudo. Essas listas são geradas a partir da ferramenta *Concord*, assim descrita.

5.3.2 *Concord*

Nessa ferramenta, listas de ocorrências são geradas por um item de busca ou nódulo (pode ser formado por uma ou mais palavras) acompanhado do seu contexto de ocorrência. O item de busca aparece centralizado e seu contexto pode ser ampliado, o que facilita a verificação de palavras ao redor. As listas de concordâncias são recursos importantes para o estudo de colocação e da padronização lexical. Na Figura 4, é possível verificar a palavra de busca "*article*", centralizada e em destaque, bem como os contextos de ocorrência.

Figura 4 - Concordâncias para o corpus de artigos da aviação.



Fonte: Os autores.

Entre as principais ferramentas do *Concord* que podem ser visualizadas no canto inferior esquerdo, podem ser destacadas *collocates, patterns* e *clusters*. O primeiro, *collocates*, colocados (em português), diz respeito às palavras que ocorrem ao redor da palavra de busca em posições que podem ser determinadas pelo pesquisador, ou seja, pode-se verificar, por exemplo, o quantitativo das palavras que mais ocorrem com a palavra de busca, bem como as posições que ocupam na oração. Nos colocados (*patterns*), obtém-se uma lista apenas de palavras que ocorrem juntas, organizadas pelas posições que ocupam nas orações. Já os *clusters* geram listas de agrupamentos lexicais para a checagem de sequências fixas de palavras recorrentes da concordância, isto é, multipalavras extraídas da concordância.

Ao efetuar-se, por exemplo, a concordância com a palavra de busca "*article*", tem-se, na lista de colocados, "*research*" e "*education*" como as palavras que mais ocorrem com "*article*". "*Research*" ocorre 667 vezes na posição primeira, à esquerda de "*article*", formando a combinação "*research article*". Nos colocados, obtêm-se dados quantitativos das palavras mais frequentes que ocorrem com a palavra de busca. Nos *patterns*, são visualizadas as combinações de palavras que ocorrem com "article" em diferentes posições nas orações. Novamente, "research" é apresentada apenas pela posição em que ocupa na oração (primeira à esquerda da palavra de busca). Nos *clusters*, encontram-se os seguintes aglomerados: "*education research article*", "*this article is*", "*writing this article*", entre outros.

No que se refere às implicações no ensino, pode-se afirmar que a descrição dos usos das colocações implica o desenvolvimento de materiais com foco no contexto e no uso por uma dada comunidade científica, ainda não estudada pelos linguistas. Sabe-se que a escolha de uma palavra ou até mesmo um sentido específico de uma palavra acarreta, obrigatoriamente, a escolha de outra, ou seja, há combinações preferenciais de palavras nas línguas.

Na construção do corpus de estudo desta pesquisa, pode-se verificar não somente palavras, mas combinações de palavras identificadas de forma rápida e eficiente, por meio de ferramentas computacionais, que analisam essas combinações em seus contextos reais em um corpus de mais de 3 milhões de palavras. No passado, demandaria uma difícil busca manual. Um trabalho praticamente impossível seria listar todas as colocações e combinações recorrentes de palavras.

Há, ainda, um outro recurso chamado *Keyword*, que poderá ser utilizado, posteriormente, se houver interesse em comparar-se a lista de palavras do corpus de estudo com listas gerais, geradas por meio de corpus maiores, as chamadas listas de corpus de referência, como, por exemplo, o *British National Corpus* (BNC), que possui aproximadamente 100 milhões de palavras.

CONSIDERAÇÕES FINAIS

As técnicas quantitativas são fundamentais para uma pesquisa linguística com base em corpus. Para entender como as escolhas léxico-gramaticais são feitas em um determinado contexto, em uma determinada comunidade, é necessário o uso de métodos quantitativos que verifiquem as palavras mais ou menos usadas em dada língua e os padrões de combinações frequentemente utilizados. A análise quantitativa enriquece a análise linguística, que é essencialmente qualitativa. O conhecimento do que é frequente permite fazer generalizações dos padrões linguísticos recorrentes. Nesse contexto, a exploração da linguística, sob o ponto de vista computacional, possibilita interação e interdisciplinaridade, pois gera necessidade de um esforço conjunto com as áreas de Computação Científica e Estatística.

Do ponto de vista dos autores, com base na interdisciplinaridade, é possível mostrar a ramificação em outras áreas do conhecimento de forma direta, por meio da visualização de dados, probabilidades do ponto de vista clássico e testes de hipóteses. Relacionado aos aspectos do ensino, também há possibilidades envolvendo a programação, uma vez que foi necessário empregar um método de raspagem de dados para a obtenção dos artigos e conversão automatizada dos arquivos em arquivos de texto. Por fim, é necessário observar que essas perspectivas não esgotam o conjunto de possibilidades para o desenvolvimento de pesquisas futuras com integração e participação do corpo discente e docente.

Espera-se que a construção deste corpus e sua análise viabilizem o *design* de materiais didáticos para o ensino de compreensão e produção escrita de textos acadêmicos. Sabe-se que a língua inglesa é a língua franca (*English as Lingua Franca*), ou ainda, a língua global, sendo majoritariamente utilizada na comunidade acadêmico-científica em situações que envolvem falantes de diferentes línguas maternas e não somente nativos (Widdowson, 2014, Crystal, 1997).

O incentivo à pesquisa pode ser estendido a cadetes de outros cursos oferecidos pela instituição - Intendência e Infantaria - para que se torne possível a construção de novos corpora de artigos acadêmicos.

Informações sobre os autores:

Fernanda Beatriz Caricari de Morais
https://orcid.org/0000-0001-6075-4101
http://lattes.cnpq.br/ 1552381480469415
fernandacaricari@gmail.com; fernandafbcm@fab.mil.brm
É Professora Adjunta III da Academia da Força Aérea. Doutora em Linguística Aplicada e Estudos da Linguagem (PUC-SP), com período no Departamento de Estudos Anglísticos da Universidade de Lisboa. Pós-doutorado na UFU (PNPD/CAPES) e na PUC-SP (PDJ/CNPq). Professora do Mestrado Profissional em Educação Bilíngue do INES/MEC-RJ desde 2014. Membro do grupo de pesquisa internacional SAL (Systemics Across Languages), dialogando também com o Núcleo de Estudos Interdisciplinares em Ciências Aeroespaciais (NEICA/UNIFA). Seus interesses de pesquisa estão relacionados com o uso da Linguística Sistêmico-Funcional e da Linguística de Corpus para análise de diversos aspectos de uso da linguagem. Atualmente, analisa as características léxico-gramaticais de artigos acadêmicos da área da aviação publicados em periódicos americanos.

João Paulo Martins dos Santos
https://orcid.org/0000-0002-0957-7119
http://lattes.cnpq.br/8043024792815551
jpmdossantos@yahoo.com.br
Possui graduação em Licenciatura em Matemática pela Universidade Estadual Paulista Júlio de Mesquita Filho (2006), mestre em Matemática pela Universidade Estadual Paulista Júlio de Mesquita Filho (2009) e Doutor em Ciências pela Escola de Engenharia de São Carlos - EESC-USP. É professor Adjunto na Academia da Força Aérea em Pirassununga/SP. Possui experiência na área de Sistemas Dinâmicos não lineares e não ideais, métodos de perturbação, métodos numéricos para solução de sistemas lineares, método de elementos finitos. Tem experiência nas áreas de Ensino e Matemática com interesse em método numéricos para solução de equações diferenciais ordinárias e parciais, estimador de erro do tipo residual para a equação do transporte de poluentes, linguagem Python de programação, Computação Científica em Python e métodos numéricos para solução de sistemas lineares, ensino de Matemática.

Contribuições dos autores:

Como os autores são de áreas distintas, seus conhecimentos se somaram para a realização deste estudo. A primeira autora ficou responsável pela conceituação linguística, bem como parte da descrição metodológica no que diz respeito ao uso de ferramentas computacionais para tratamento de dados para análise léxico-gramatical do corpus de estudo. Enquanto o segundo autor se preocupou com a raspagem de dados para coleta do corpus. Ambos trabalharam juntos na redação do artigo e na discussão das implicações futuras e no uso do corpus em pesquisas envolvendo cadetes dos cursos oferecidos pela Academia da Força Aérea.

Como citar este artigo:

ABNT

MORAES, F. B. C.; SANTOS, J. P. M. A construção do corpus de artigos científicos de aviação: um estudo interdisciplinar. **Revista da UNIFA**, Rio de Janeiro, v. 37, p. 1-21, 2024.

APA

MORAES, F. B. C.; SANTOS, J. P. M. (2024, Março) A construção do corpus de artigos científicos de aviação: um estudo interdisciplinar. **Revista da UNIFA**, 37(1), P. 1-21.

REFERÊNCIAS

BERBER SARDINHA, T. Computador, corpus e concordância no ensino de léxico-gramática de língua estrangeira. In: V, Leffa (org.) **As palavras e sua companhia: o léxico na aprendizagem**. Pelotas: EDUCAT, UCP, p. 45-72, 2000.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri-SP: Manole, 2004.

BIBER, D. Representativenessin Corpus Design. **Linguist Computing**. v. 8, p. 243-257, 1993.

BIRD, Steven; LOPER, Edward; KLEIN, Ewan. **Natural Language Processing with Python. O'Reilly Media Inc**., 2009. Disponível em: https://www.nltk.org/book/. Acesso em: 24 jul. 2023.

BISONG, E. Google Collaboratory. In: **Building Machine Learning and Deep Learning Models on Google Cloud Platform**. Berkeley, CA: Apress, 2019. Capítulo 7. Disponível em: https://doi.org/10.1007/978-1-4842-4470-8_7. Acesso em: 8 de set. 2023.

CRYSTAL, D. **English as a global Language**. Cambridge. Cambridge University Press, 1997.

EGGINS, S. **An introduction to Systemic Functional Linguistics**. Londres: Pinter Publishers, 1994.

GOUVEIA, C. Texto e gramática: uma introdução a linguística sistêmico-funcional.

Matraga. Rio de Janeiro, v. 16, n. 24, p. 13-47, 2009.

GROSS, A. **The rhetoric of science**. Cambridge, MA: Harvard University Press, 1996.

HALLIDAY, M. A. K. **An introduction to Functional Grammar**. Londres: Edward Arnold, 1994.

HALLIDAY, M. A. K; MATTHIESSEN, C. M.I.M. **An introduction to Functional Grammar**. Londres: Edward Arnold. Third Edition, 2004.

HALLIDAY, M. A. K; MATTHIESSEN, C. M.I.M. **An introduction to Functional Grammar**. Londres: Edward Arnold. Third Edition, 2014.

HARRIS, Charles R. et al. Array programming with NumPy. **Nature**, v. 585, n. 7825, p. 357-362, set. 2020. DOI: 10.1038/s41586-020-2649-2. Disponível em: https://doi.org/10.1038/s41586-020-2649-2. Acesso em: 10 de set. 2023.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90-95, 2007.

pdfminer.six. (2023). pdfminer.six (Version 20221105). [Software de extração de texto de PDF]. Disponível em: https://pypi.org/project/pdfminer.six/. GitHub repository: https://github.com/pdfminer/pdfminer.six. Acesso em: 21 de set. 2023.

MARTIN, J. R. English Text: **System and Structure**. Amsterdam: Benjamins, 1992.

McENERY, T. & WILSON, A. **Corpus Linguistics**. Edinburgh, Edinburgh University Press, 2001.

MOITA LOPES, L. P. (Org.) **Por uma Linguística Aplicada Indisciplinar**. São Paulo: Parábola Editorial, 2006.

MORAIS, F. B. C. **Entre alhos e bugalhos – os usos do clítico SE na escrita acadêmica**. Tese de Doutorado. PUC-SP, 2013.

MORAIS, F. B. C. Os dizentes nos artigos científicos de Linguística - um estudo baseado na Linguística Sistêmico-Funcional e com o auxílio da Linguística de Corpus. **Letras & Letras**, v. 30, p. 46-63, 2014.

MORAIS, F. B. C. O uso do processo existencial 'haver' na escrita acadêmica: um estudo com base em um corpus de artigos científicos de diversas áreas do conhecimento. **Revista (Con) Textos Linguísticos** (UFES), v. 9, p. 142-160, 2015.

MORAIS, F. B. C. O gênero resenha na sala de aula de Língua Portuguesa como L2. **Anais do IV Encontro Mundial de Ensino de Língua Portuguesa**. Washington: Georgetown University, 2016.

MOREIRA FILHO, J. L. **Python para Linguística de Corpus : guia prático**, 1. ed., São Paulo: Ed. do Autor, 2021.

RICHARDSON. L. BeautifulSoup (Version 4.11.2). Pacote Python para análise de documentos HTML e XML. Disponível em: https://pypi.org/project/beautifulsoup4/. Repositório do GitHub: https://github.com/wention/Beautifu lSoup4. Acesso em: 19 out. 2023.

SANCHEZ, A. Definicion e historia de los corpus. In: SANCHEZ, A et al (Org.) **CUMBRE – corpus linguistico de espanol contemporaneo**. Madrid: SGEL, 1995. SCOTT, M. R. **Wordsmith Tools v. 8**. Software for text analysis. Oxford University Press, 2018.

THOMPSON, G. **Introducing Functional Grammar**. New York: Routledge, 1996.

TRASK, R. L. **Dicionário de Linguagem e Linguística**. São Paulo: Contexto, 2004.

VIRTANEN, Pauli et al. SciPy 1.0: Algoritmos fundamentais para computação científica em Python. **Nature Methods**, v. 17, p. 261-272, 2020. DOI: 10.1038/s41592-019-0686-2.

WIDDOWSON, H. ELF and the pragmatics of language variation. **Journal of English as Lingua Franca**. V. 4 (2), pp. 359-372, 2015.

Recebido: 06 Set 2024

Aceito: 21 Nov 2024