Revista GTLex

DOI: 10.14393/Lex-v10a2025-8
e010008

Artigos

Elaboração de *corpora* para estudo terminológico

Creating *corpora* for terminology studies

Ana Luiza Noventa Dallapicula [*]

RESUMO: Este artigo é um recorte sobre a metodologia de construção e elaboração dos *corpora* do estudo de mestrado da harmonização terminológica entre as variantes faladas no Brasil e Portugal como membros da Comunidade de Países de Língua Portuguesa (CPLP) em termos usados no ingresso à educação superior. O objetivo é apresentar a metodologia desenvolvida com o uso da Linguística de *Corpus* e a ferramenta de *software* usada na construção dos *corpora* do estudo, o Sketch Engine. Ao final, o artigo aborda alguns resultados alcançados, como o uso de termos diferentes para nomear etapas de ensino nos países e a indicação de termos equivocados pelo *software*, o que reforça a indispensabilidade do trabalho humano com a filtragem dos dados fornecidos pela máquina.
PALAVRAS-CHAVE: Terminologia. *Corpora*. Sketch Engine.

ABSTRACT: This article is an excerpt from the methodology used to build and prepare the corpora for a master's study of terminological harmonization between the variants spoken in Brazil and Portugal as members of the Community of Portuguese Language Countries (CPLP) in terms used to enter higher education. The aim is to present the methodology developed using Corpus Linguistics and the software tool used to build the study's corpora, the Sketch Engine. At the end, the article discusses some of the results achieved, such as the use of different terms to name stages of education in the countries and the indication of erroneous terms by the software, which reinforces the indispensability of human work in filtering the data provided by the machine.
KEYWORDS: Terminology. *Corpora*. Sketch Engine.

1. Introdução

Este artigo apresenta a aplicação da Linguística de *Corpus* na metodologia desenvolvida para uma pesquisa em desenvolvimento de mestrado em Linguística na Universidade de Brasília inserida nos estudos de Terminologia. O estudo busca a harmonização terminológica por meio de um estudo das variações do português usado em documentos que dissertam sobre o acesso e processo de ingresso em

[*] Mestre em Linguística - Universidade de Brasília (UnB) / Brasília. analuiza90.unb@gmail.com.

instituições de ensino superior de Estados-membros da Comunidade de Países de Língua Portuguesa - CPLP.

A CPLP foi criada em 17 de julho de 1996 e é composta por nove Estados-Membros, sendo eles Angola, Brasil, Cabo Verde, Guiné-Bissau, Guiné Equatorial, Moçambique, Portugal, São Tomé e Príncipe, e Timor-Leste. Além desses países, possui outros membros efetivos: Goa, na Índia; e Macau, na China.

A comunidade tem como propósito promover a amizade e cooperação entre os países lusófonos que a integram, preservando sua pluralidade. Dessa forma, a comunidade oferece acordos e documentos que contêm terminologias de diferentes áreas, como a Educação, tornando-se essencial para estudos relacionados à língua portuguesa.

Ressalta-se que a CPLP desempenha um papel crucial na preservação da língua portuguesa, promovendo a diversidade entre os povos que a falam, ao mesmo tempo em que valoriza suas diferenças culturais e regionais, além de fomentar a cooperação por meio de acordos. Portanto, a comunidade possui papel crucial no âmbito do Acordo sobre a Mobilidade entre os Estados-Membros da CPLP, uma vez que a partir desse acordo busca criar um modelo que facilite a transição e deslocamento dos cidadãos residentes nos países membros do grupo, para reforçar e preservar a ideia de que a "mobilidade dos cidadãos nos territórios que a compõem deve ser tão livre quanto possível" (Brasil, 2022).

Esse acordo e ideias reforçam a importância da harmonização terminológica, pois essa ajuda na transição e adaptação dos cidadãos em um novo país. Isso porque a harmonização funciona como uma facilitadora na integração dos migrantes através da redução de barreiras de variações linguísticas em setores regulamentados como a saúde, a educação e o direito, já que regulamentação e consistência terminológica são essenciais para esses setores. Além disso, também colabora com a integração acadêmica, social e profissional, pois torna os migrantes capacitados para participar ativamente da vida estudantil, acadêmica, social e profissional.

No campo da terminologia, as variações linguísticas são percebidas em contextos específicos, nesse caso, o contexto da educação superior. Nesse sentido, destaca-se que as linguagens de especialidade refletem o ambiente e a sociedade em que estão inseridas. Essa questão demanda a distinção dos padrões linguísticos presentes nessas variações, analisando se são semelhantes, divergentes ou se cada variação segue suas próprias normas linguísticas.

A partir dessa problemática, se faz importante o uso da Linguística de *Corpus* (LC) como principal auxiliar de análises linguísticas, uma vez que é utilizada para a criação dos *corpora* necessários para a realização do estudo de Harmonização Terminológica. A LC neste estudo é utilizada para identificar os termos mais relevantes (termos-chave e/ou mais frequentes), oferecendo uma visão abrangente e representativa do vocabulário empregado pelos especialistas na área em questão. Assim, é feita a análise do uso desses termos com base na observação dos contextos em que são aplicados, permitindo compreender como são utilizados em diferentes situações e contextos comunicativos. Também é verificada a existência de termos com significados semelhantes, bem como suas estruturas gramaticais. Portanto, serão elaborados os *corpora*, que, conforme mencionado, servirão como banco de dados para o estudo.

Dessa maneira, a metodologia deste estudo foi elaborada em quatro etapas, sendo estas: i) estudar e compreender os sistemas de educação; ii) selecionar os textos e documentos que irão compor os *corpora*; iii) nomear os documentos através de um padrão criado e compilar os *corpora*; e iv) identificar os candidatos a termos em cada *corpus* com uso de ferramentas internas do *software* de LC escolhido.

Optou-se, neste estudo, pelo uso do *software* pago Sketch Engine para a elaboração de dois *princípios* do estudo: o *corpus* do ingresso ao ensino superior do Brasil e o *corpus* do ingresso ao ensino superior de Portugal. Observa-se que esses países foram inicialmente escolhidos como teste para elaborar os *corpora*, visto a

facilidade de encontrar documentos sobre os mesmos em comparação com os outros países lusófonos.

Entretanto, a pesquisa de mestrado visa construir outros *corpora* relativos aos outros países da CPLP, após a criação, elaboração e aplicação dessa metodologia discutida neste artigo. A construção desses *corpora* parte desde os critérios de seleção dos textos que compõem os *corpora* à criação de mapas conceituais que auxiliam na visualização dos sistemas de educação dos países e organização de dados fundamentais para a pesquisa.

2. Pressupostos teóricos

Define-se a LC como a que se ocupa "com um conjunto de textos legíveis por máquina que é considerado como base apropriada para estudar um conjunto específico de questões de pesquisa" (McEnery & Hardie, 2012, p. 1). Sardinha (2000) também define a LC como aquela que:

> (...) ocupa-se da coleta e da exploração de *corpora*, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador (Sardinha, 2000, p. 325).

Pode-se dizer, então, que a LC estuda e analisa grandes conjuntos de textos, com o objetivo de identificar padrões, uso de vocabulário, estruturas gramaticais e outros fenômenos linguísticos. Essa metodologia se apoia na coleta, organização e exame de dados linguísticos reais, utilizando um *corpus* ou um conjunto de *corpora*, o que possibilita explorar como a língua é empregada em contextos de uso real, ou seja, em uma situação autêntica de comunicação e uso da língua com um propósito genuíno, fora de ambientes controlados de aprendizado, como uma sala de aula, por exemplo.

Com isso, a LC está ligada a "um método de investigação de inúmeras linhas de concordância e listas de palavras geradas por um programa de computador, com o objetivo de entender fenômenos que ocorrem em textos grandes ou em compilações de textos pequenos" (McCarthy e O'Keeffe 2010, p. 3 *apud* Shepherd, 2012, p. 15).

Ainda, na Terminologia, "a busca e o tratamento dos dados passaram a se fazer dentro do texto, ou melhor, dentro de um *corpus*. Em outras palavras, dentro de um 'conjunto de fontes relativas a uma área'" (ISO 1087, 1990 *apud* Barros, 2004, p. 262). Dessa forma, a LC oferece evidências para a análise e compreensão de termos usados em contextos especializados, colaborando com a pesquisa terminológica na seleção e definição de termos, além de tornar o processo de criação de recursos terminológicos mais sólido e orientado por dados reais de uso linguístico em contextos especializados.

A pesquisa terminológica atual busca certas análises textuais que dependem de extrair informações dos textos e documentos usados dentro da área de especialidade que se estuda. Dessa maneira, o trabalho com um grande e diversificado número de informações implica a necessidade de extrair informações, como contextos e frequência de uso e conteúdos semânticos e gramaticais dos termos, que justificam o uso de *corpora* eletrônico especializado.

A LC oferece vantagens e disponibiliza ferramentas úteis para a elaboração de *corpus*, identificação e análise de termos, utilizando recursos digitais e tecnológicos que simplificam o processo, desde a compilação dos *corpora* até a identificação de possíveis termos. No contexto do estudo do processo de ingresso em instituições de ensino superior de países da CPLP, ela também auxilia na verificação das semelhanças e diferenças semânticas e gramaticais.

As ferramentas tecnológicas oferecem recursos que tornam mais fácil a coleta, preparação, análise e interpretação de grandes volumes de dados linguísticos. Elas automatizam o processo de coleta de textos em grande escala e utilizam programas de concordância, além de ferramentas computacionais especializadas, que ajudam na análise detalhada e precisa dos dados, facilitando a identificação de padrões

linguísticos e frequência de palavras. Os programas de concordância possibilitam que os pesquisadores examinem palavras ou termos em um *corpus*, gerando listas extensas de ocorrências em diversos contextos de uso. Essas listas permitem analisar as colocações e a frequência de uso dos termos (Biber, Conrad & Reppen, 1998, p. 15).

Nesse contexto, para este estudo foi escolhido o Sketch Engine, uma ferramenta paga disponível online, que é muito efetiva na construção de e tratamento de *corpus*. O Sketch Engine é utilizado para a análise e exploração de grandes conjuntos de dados textuais em diversas línguas e foi criado para atender pesquisadores, tradutores, lexicógrafos, linguistas e profissionais de diversas áreas que trabalham com linguagem. O *software* oferece uma plataforma que permite estudar o uso da língua em diferentes contextos e identificar padrões desde os lexicais aos gramaticais.

Com esse software, os usuários podem acessar *corpora* prontos de várias línguas ou criar seus próprios *corpora* a partir de coleções de textos. A ferramenta oferece funcionalidades como a geração de listas de frequência, análise de concordância, extração automática de colocações e criação de word sketches, que são "resumos" das palavras com suas principais combinações semânticas e sintáticas. Esses recursos permitem uma análise profunda e detalhada do comportamento das palavras, auxiliando na pesquisa terminológica, no estudo de variações linguísticas e na construção de dicionários.

Dessa forma, o Sketch Engine (Kilgarriff *et al*, 2004) foi escolhido para uso nesta pesquisa, pois o software permite criar *corpora* personalizados a partir de textos que podem ser escolhidos pelo pesquisador, como é o caso desse estudo. É, então, uma ferramenta versátil na elaboração de *corpus*, pois podem ser usados fontes textuais como documentos formais, artigos acadêmicos, sites, e até mídias sociais.

Dentre as funcionalidades internas do programa, ele possibilita a criação de listas de palavras frequentes, wordlists, e exibe concordâncias. Isso permite ao usuário observar o uso de termos em contextos reais, o que facilita a identificação de padrões linguísticos e semânticos. Ainda, o Sketch Engine oferece recursos de análise

linguística, como o *Word Sketch*, que fornece resumos automáticos de palavras em contextos específicos, para auxiliar no estudo e verificação dos termos em contextos diversos.

O programa também possui uma interface intuitiva que facilita o uso por pessoas com diferentes níveis de conhecimento técnico. Isso porque o seu processamento fornece resultados rápidos e precisos, permitindo a análise de grandes volumes de dados linguísticos de maneira eficiente.

Essas razões tornam o Sketch Engine uma ferramenta ideal para criar, gerenciar e analisar *corpora* de maneira eficiente, oferecendo precisão e profundidade nas análises linguísticas. Por isso foi utilizado no estudo como parte da metodologia na elaboração de *corpora*, identificação e seleção de termos e análise linguística.

Para iniciar o processo de compreensão dos termos usados e suas diferenças, foi necessária a análise da formação dos termos, a depender das diferenças e semelhanças identificadas nos termos em estudo para se propor uma harmonização terminológica. Para isso, baseou-se no constructo de Faulstich (2003) que fundamenta a classificação dos termos como unidades terminológicas simples (UTS) e unidades terminológicas complexas (UTC).

Segundo o constructo, Faulstich (2003), aplicado por Maia-Pires (2009, p. 60), propõem sobre a definição de termos simples e termos complexos, nos quais os complexos são termos compostos formados pela junção de uma base terminológica simples (termos simples), que detém um conceito dentro de uma linguagem de especialidade, e uma "predicação" aplicada ao termo simples pela necessidade de especificar-se ainda mais um conceito dentro da linguagem de especialidade. Dessa forma, "os elementos compostos apenas por uma base lexical correspondem às UTS e os elementos compostos de 'base + predicação' correspondem às UTCs" (Faulstich, 2003 *apud* Maia-Pires, 2009, p. 60).

Apresenta-se a seguir, a metodologia elaborada usada para a criação dos *corpora* e deste estudo.

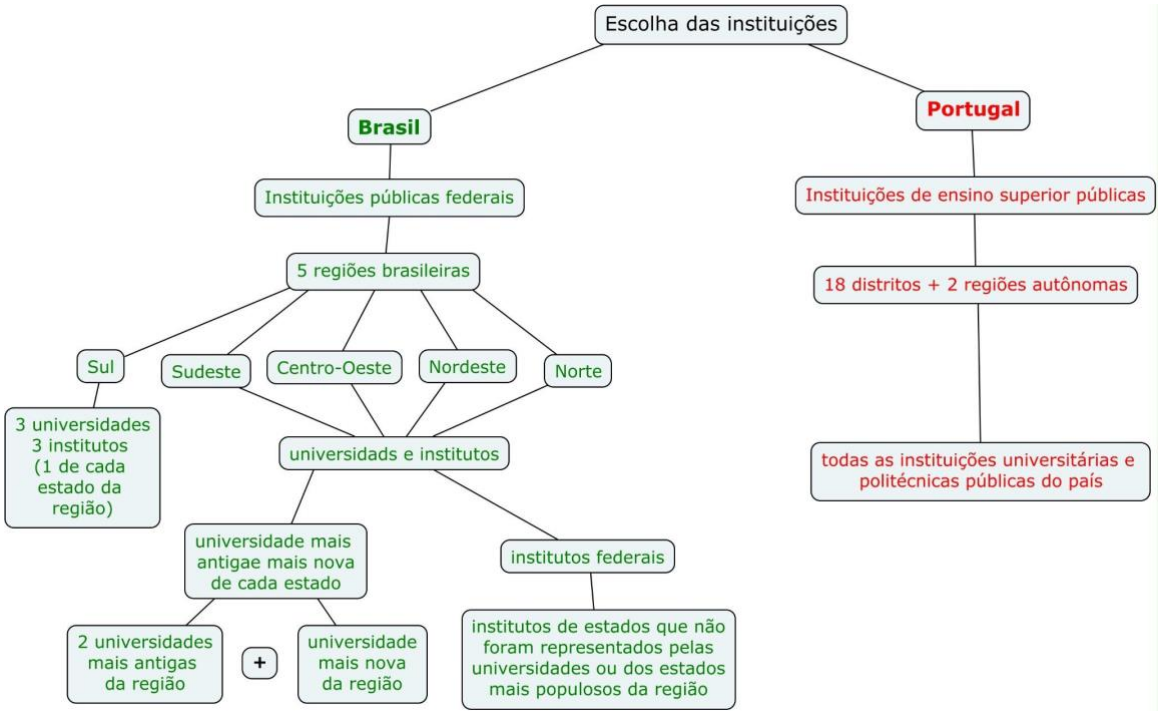
3. Metodologia

A elaboração de *corpora* é uma parte essencial da metodologia da pesquisa e iniciou-se a partir da primeira etapa da metodologia: a criação de mapas conceituais. Para isso, utilizou-se a ferramenta gratuita IHMC CmapTools (Cañas *et al*, 2004), em que se organizou as informações que auxiliaram na compreensão dos sistemas de educação do Brasil e Portugal, e das terminologias usadas nos níveis de educação e ciclos de estudo. Observou-se que os sistemas educacionais de ambos os países são organizados de forma distinta e apresentam variações terminológicas, embora compartilhem semelhanças conceituais.

Após compreender os sistemas de educação a partir de buscas em sites oficiais e legislações, concluiu-se que, para os fins da pesquisa, os níveis de educação secundária (ou ensino médio no Brasil) e educação superior, bem como as instituições de ensino superior, são essenciais para fornecer os documentos relativos aos processos de ingresso que irão compor os *corpora*. Por se tratar de um estudo comparativo entre duas variações da língua portuguesa, foram elaborados dois *corpora* representativos de cada país para detectar os termos usados com mais frequência e relevância em cada localidade.

Assim, a segunda etapa da metodologia consistiu na seleção de textos para compor os *corpora.* Assim, o critério de seleção de textos baseou-se na obrigatoriedade de textos digitalizados e em escrita formal. Foram selecionados textos de legislações atuais sobre os sistemas de educação do Brasil e Portugal, além de editais de ingresso e matrícula nas instituições de ensino superior desses países. Para melhor selecionar os documentos das instituições de ensino superior, com a finalidade de manter os *corpora* equilibrados, criou-se o seguinte mapeamento para a coleta de dados:

Figura 1: Escolha das instituições.



Fonte: Elaborado pela autora.

Para o *corpus* do Brasil, utilizaram-se apenas instituições públicas federais, escolhendo uma universidade federal e um instituto federal de cada um dos 26 estados, além do Distrito Federal. No entanto, para equilibrar o número de instituições entre Brasil e Portugal, outros critérios foram definidos, como a seleção de 3 universidades federais e 3 institutos federais de cada uma das cinco regiões do Brasil.

Para as regiões Sul e Sudeste, foram escolhidos institutos e universidades de cada estado, enquanto nas regiões Centro-Oeste, Norte e Nordeste, priorizou-se a seleção das universidades mais antigas e novas de cada estado. Como os Institutos Federais foram todos fundados na mesma data, a seleção priorizou os estados não incluídos na seleção das universidades e os estados mais populosos.

Para o *corpus* de Portugal, os documentos foram retirados de sites de todas as instituições públicas de ensino superior do país, uma vez que possui uma instituição pública por distrito/região autónoma.

Passou-se, então, para a terceira etapa da metodologia, a nomeação dos documentos e a compilação dos *corpora.* Dessa forma, teve-se a elaboração dos *corpora* da pesquisa com os seguintes conjuntos de dados:

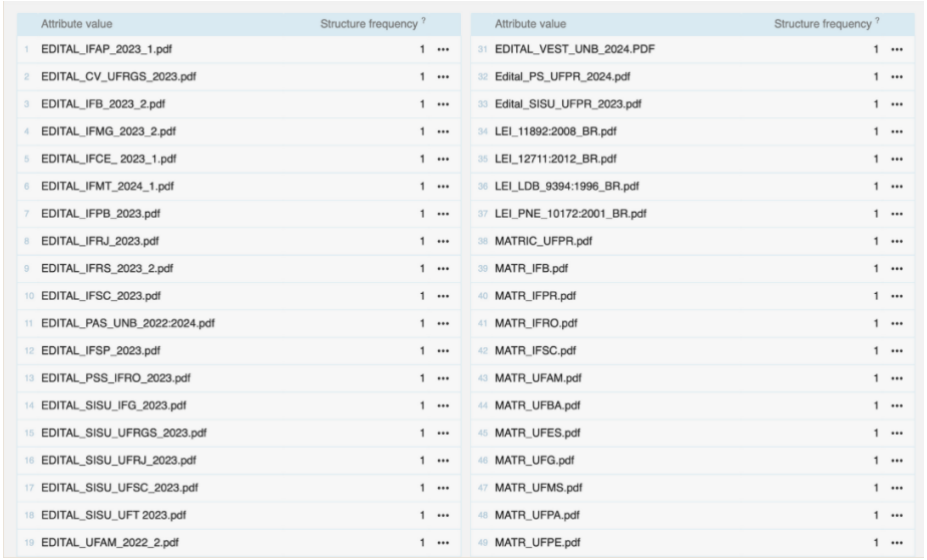
Quadro 1: *Corpora* Brasil e Portugal.

Corpus	Documentos	*Tokens*	Palavras	Tipos de documentos
Ingresso.Ens.Sup.BR	59	830.103	624.322	4 leis, 1 portaria, 16 informativos de matrículas e 38 editais.
Ingresso.Ens.Sup.PT	64	543.626	421.528	19 despachos, 6 decretos-lei, 1 lei, 4 portarias, 1 deliberação, 5 editais, 26 regulamentos e 1 retificação.

Fonte: Elaborado pela autora.

Todos os documentos dos *corpora* são textos digitalizados, escritos e em linguagem formal que seguem um padrão de nomenclatura criado para organizar os documentos e facilitar a identificação dentro do *corpus*, como o exemplificado a seguir no *corpus* do Brasil:

Figura 2: Parte dos documentos do *corpus* "Ingresso.Ens.Sup.BR**".**



	Attribute value	Structure frequency
1	EDITAL_IFAP_2023_1.pdf	1
2	EDITAL_CV_UFRGS_2023.pdf	1
3	EDITAL_IFB_2023_2.pdf	1
4	EDITAL_IFMG_2023_2.pdf	1
5	EDITAL_IFCE_ 2023_1.pdf	1
6	EDITAL_IFMT_2024_1.pdf	1
7	EDITAL_IFPB_2023.pdf	1
8	EDITAL_IFRJ_2023.pdf	1
9	EDITAL_IFRS_2023_2.pdf	1
10	EDITAL_IFSC_2023.pdf	1
11	EDITAL_PAS_UNB_2022:2024.pdf	1
12	EDITAL_IFSP_2023.pdf	1
13	EDITAL_PSS_IFRO_2023.pdf	1
14	EDITAL_SISU_IFG_2023.pdf	1
15	EDITAL_SISU_UFRGS_2023.pdf	1
16	EDITAL_SISU_UFRJ_2023.pdf	1
17	EDITAL_SISU_UFSC_2023.pdf	1
18	EDITAL_SISU_UFT 2023.pdf	1
19	EDITAL_UFAM_2022_2.pdf	1

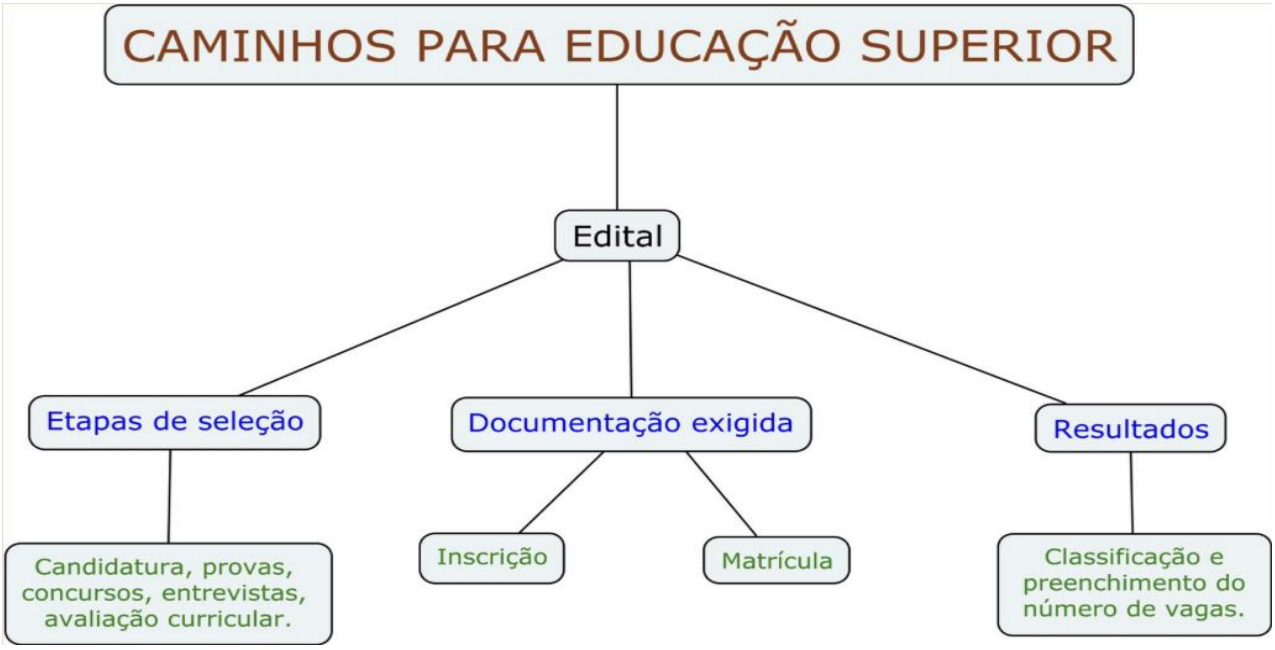
	Attribute value	Structure frequency
31	EDITAL_VEST_UNB_2024.PDF	1
32	Edital_PS_UFPR_2024.pdf	1
33	Edital_SISU_UFPR_2023.pdf	1
34	LEI_11892:2008_BR.pdf	1
35	LEI_12711:2012_BR.pdf	1
36	LEI_LDB_9394:1996_BR.pdf	1
37	LEI_PNE_10172:2001_BR.pdf	1
38	MATRIC_UFPR.pdf	1
39	MATR_IFB.pdf	1
40	MATR_IFPR.pdf	1
41	MATR_IFRO.pdf	1
42	MATR_IFSC.pdf	1
43	MATR_UFAM.pdf	1
44	MATR_UFBA.pdf	1
45	MATR_UFES.pdf	1
46	MATR_UFG.pdf	1
47	MATR_UFMS.pdf	1
48	MATR_UFPA.pdf	1
49	MATR_UFPE.pdf	1

Fonte: Sketch Engine, 2024.

Após a finalização da elaboração dos *corpora* do Brasil e de Portugal, com o intuito aplicar a metodologia da pesquisa, iniciou-se a quarta e última etapa da

metodologia, o processo de identificação de candidatos a termos a serem analisados dos dois *corpora* finalizados. Para isso, foi feito um novo mapeamento para compreender os domínios em que os termos deveriam ser procurados:

Figura 3: Organização de domínios.

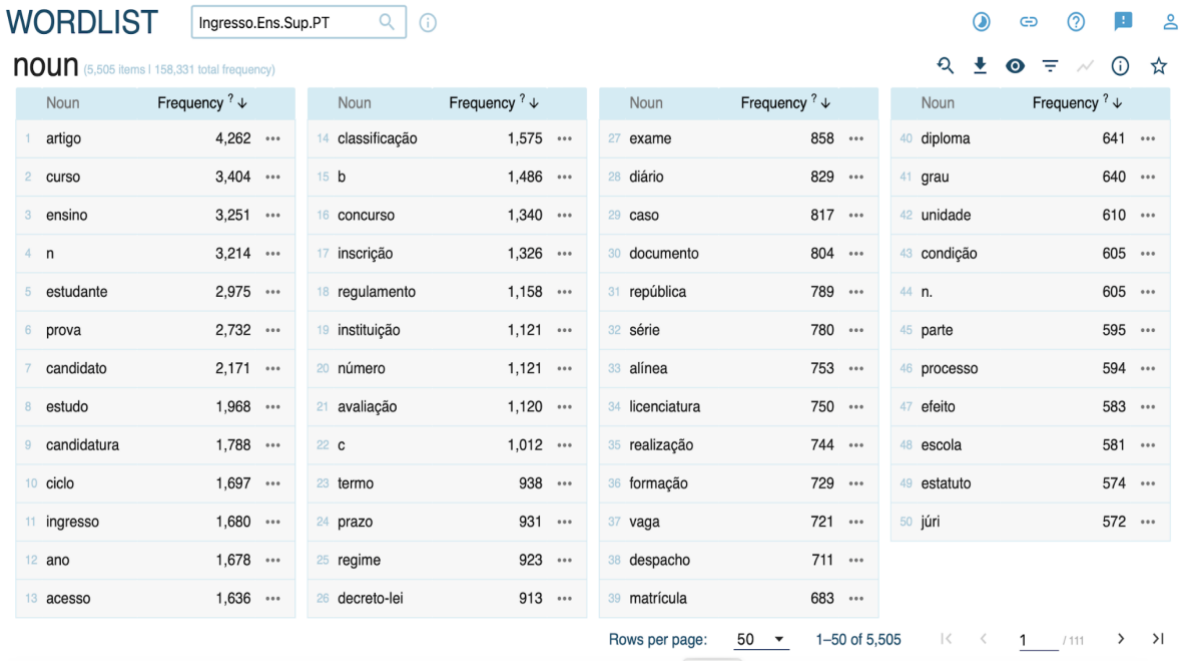


Fonte: Elaborado pela autora.

Com base no mapa acima, os passos essenciais que delineiam os subdomínios dos termos incluem: as etapas de seleção, a documentação necessária e os resultados das etapas de seleção.

Para selecionar candidatos a termos, utilizou-se a ferramenta interna chamada *Wordlist* para gerar uma lista dos substantivos mais frequentes de cada *corpus*. Ressalta-se que a classe dos substantivos é a mais predominante no *corpora*, pois é a classe gramatical responsável por nomear etapas, documentos e procedimentos necessários em uso no contexto dessa linguagem de especialidade. Como exemplo, segue abaixo a lista de 50 "substantivos" (*nouns*) gerada pela aplicação da ferramenta *Wordlist* ao corpus de Portugal:

Figura 4: *Wordlist corpus* "Ingresso.Ens.Sup.PT".



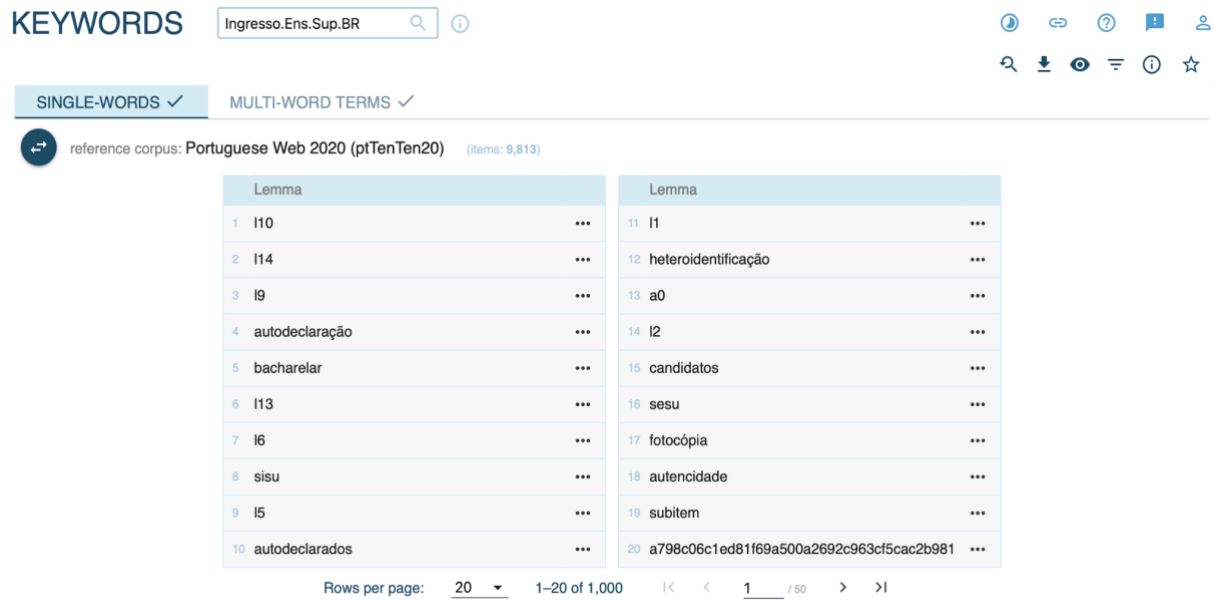
	Noun	Frequency
1	artigo	4,262
2	curso	3,404
3	ensino	3,251
4	n	3,214
5	estudante	2,975
6	prova	2,732
7	candidato	2,171
8	estudo	1,968
9	candidatura	1,788
10	ciclo	1,697
11	ingresso	1,680
12	ano	1,678
13	acesso	1,636
14	classificação	1,575
15	b	1,486
16	concurso	1,340
17	inscrição	1,326
18	regulamento	1,158
19	instituição	1,121
20	número	1,121
21	avaliação	1,120
22	c	1,012
23	termo	938
24	prazo	931
25	regime	923
26	decreto-lei	913
27	exame	858
28	diário	829
29	caso	817
30	documento	804
31	república	789
32	série	780
33	alínea	753
34	licenciatura	750
35	realização	744
36	formação	729
37	vaga	721
38	despacho	711
39	matrícula	683
40	diploma	641
41	grau	640
42	unidade	610
43	condição	605
44	n.	605
45	parte	595
46	processo	594
47	efeito	583
48	escola	581
49	estatuto	574
50	júri	572

Fonte: Sketch Engine, 2024.

No entanto, as listas geradas pelo sistema podem levar a erros ou à identificação de caracteres como termos, além de incluir substantivos que não são pertinentes ao contexto da pesquisa. Por isso, é fundamental uma filtragem humana das listas para identificar os verdadeiros candidatos a termos, e, a frente, uma verificação dos contextos de uso nos *corpora*, a fim de excluir aqueles que não são significativos ou que apresentam estruturas e significados iguais (ou que não requer harmonização).

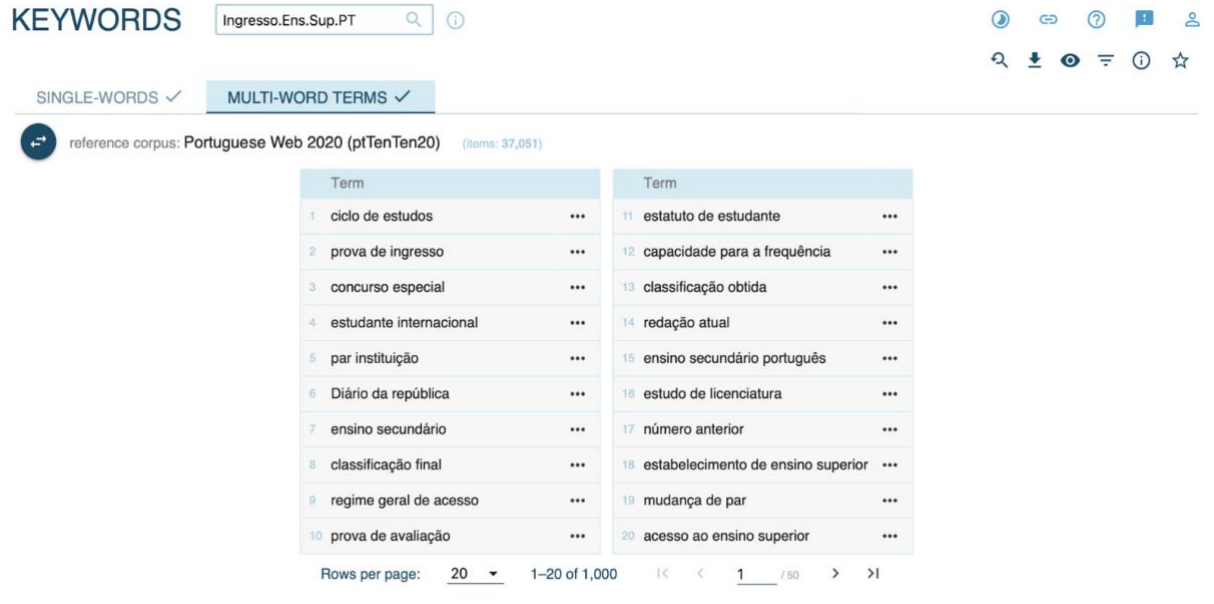
Outra ferramenta que o Sketch Engine fornece é a *Keywords*, palavras-chave, que busca termos de base lexical simples e termos de base lexical complexa. Foram extraídas, então, outras duas listas de palavras-chave: *Single-words* e *Multi-words*. Novamente uma revisão de palavras-chave através de uma filtragem humana foi necessária. Segue abaixo, como exemplos, as listas, respectivamente, de 20 *single-words* do *corpus* do Brasil e 20 *multi-words* do *corpus* de Portugal, geradas automaticamente pelo software:

Figura 5: *Single-words corpus* "Ingresso.Ens.Sup.BR".



Fonte: Sketch Engine, 2024.

Figura 6: *Multi-words corpus* "Ingresso.Ens.Sup.PT".

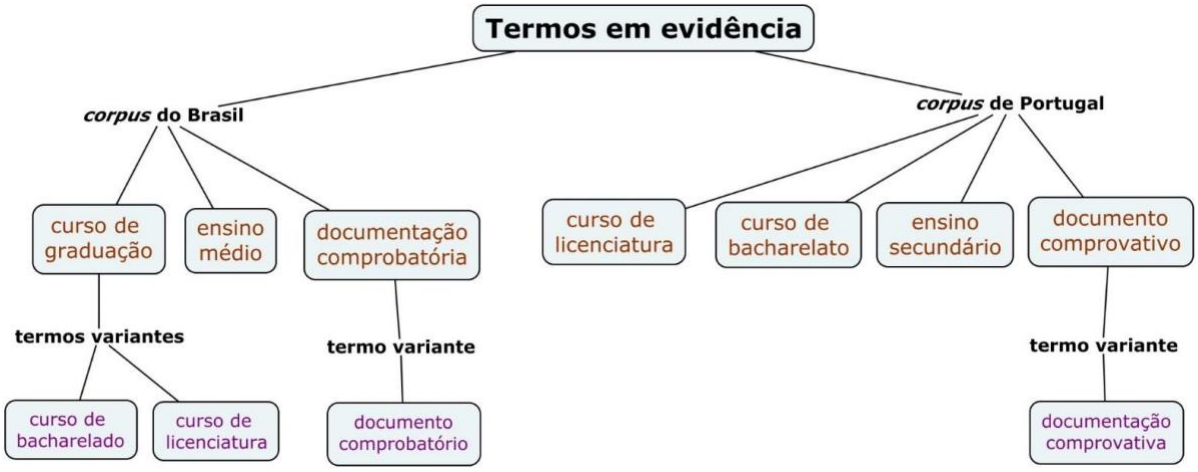


Fonte: Sketch Engine, 2024.

4. Resultados

Diante disso, após as revisões das listas e a aplicação dos procedimentos explicitados e organizados pelo mapeamento mostrado na Figura 3, pode-se identificar os termos mais frequentes nos *corpora* do Brasil e de Portugal que são essenciais para o trabalho de harmonização. Portanto, são estes:

Figura 7: Termos identificados.



Fonte: Elaborado pela autora.

Esses termos foram identificados a partir de palavras-chave e substantivos indicados pelo software, como “licenciatura”, “documento”, “curso”, “ensino secundário” e “ensino médio”. Entretanto, foi necessário aplicar a teoria do constructo de Faulstich (2003) para alcançar esses resultados de identificação de termos.

Ao observar a Figura 6, notam-se termos UTCs, em que a base do termo é tipicamente um substantivo, enquanto a predicação é um adjetivo que especifica o substantivo. Assim, nos termos identificados no *corpus* do Brasil, tem-se 3 UTCs de termos de entrada: curso de graduação, ensino médio e documentação comprobatória; e 3 UTCs como termos variantes dos termos de entrada: curso de bacharelado, curso de licenciatura e documento comprobatório. No *corpus* de Portugal, foram identificadas 4 UTCs de termos de entrada: curso de licenciatura, curso de bacharelato,

ensino secundário e documento comprovativo; e 1 UTC como termo variante: documentação comprovativa.

Observa-se que a base dos termos identificados é um substantivo comum parte do léxico da língua portuguesa e que pode ser usado em contextos diferentes, sem que necessariamente façam parte de uma linguagem de especialidade. Esse é o caso de "curso", "ensino", "documentação" e "documento", por exemplo. Assim, esses substantivos sozinhos não podem ser classificados como termos, pois não são específicos no contexto da linguagem usada no ingresso em instituições de ensino superior, como neste estudo.

É a partir disso que se entende que esses substantivos estão normalmente acompanhados de adjetivos ou outros substantivos que funcionam como adjetivos dos substantivos base que acompanham, seguindo a forma "base + predicação", formada pelo padrão sintático "substantivo + adjetivo". Isso porque esses substantivos precisam ser caracterizados para se tornarem específicos de um contexto e, consequentemente, um termo. Têm-se nessa pesquisa, então, termos formados por UTCs, como reúne o quadro abaixo:

Quadro 2: Formação das UTCs identificadas.

Base dos termos (substantivos)	**Predicação** (adjetivação dos substantivos)
curso	graduação; bacharelado; licenciatura; bacharelato.
ensino	médio; secundário.
documentação	comprobatório; comprovativo.
documento	comprobatório; comprovativo.

Fonte: Elaborado pela autora.

Diante dos dados apresentados após a aplicação da metodologia, identificação e análise dos termos, pode-se chegar a algumas conclusões que são apresentadas a seguir.

5. Considerações finais

Com a elaboração dos *corpora*, o uso das ferramentas tecnológicas de geração automática de lista de candidatos a termos feita pelo software e os estudos dos possíveis termos, bem como a aplicação da teoria do constructo, pode-se chegar à algumas conclusões.

Com a identificação dos termos presentes na Figura 6, mesmo que sejam apenas dos *corpora* do Brasil e de Portugal, pode-se concluir que nesse estudo, em especial por ser uma pesquisa terminológica, os termos são formados por unidades terminológicas complexas. Isso acontece, pois por se tratar de uma linguagem de especialidade usada em contexto específico de comunicação e que intenciona, também, alcançar fins específicos, há necessidade de caracterizar substantivos para que sejam específicos e próprios da linguagem de especialização em que se encontram em uso.

Nota-se, também, que as bases dos termos podem ser iguais em termos diferentes e o que os diferencia são suas predicações. Ou seja, o adjetivo ou o substantivo adjetivado que torna um termo único. Ainda, percebe-se que não é eficaz a escolha de apenas um termo, visto que em cada país há usos diferentes.

Conclui-se que esse estudo precisa ser aprimorado e que a harmonização terminológica não é viável nesse contexto. Dessa forma, a metodologia de criação de *corpora* é eficaz, mas para a análise dos termos deve-se aplicar outra teoria pertinente aos objetivos do estudo e da CPLP.

Para os fins deste artigo, entende-se que a elaboração de *corpora* é essencial para realizar o recorte terminológico. O uso da ferramenta do Sketch Engine, também e em paralelo com observação humana dos dados, é fundamental para auxiliar na identificação dos termos.

Referências

BARROS, L. A. **Curso Básico de Terminologia**. São Paulo: Edusp, 2004.

BIBER, D.; CONRAD, S.; REPPEN, R. ***Corpus* Linguistics:** investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BRASIL. Decreto n° 11.156, de 29 de julho de 2022. Promulga o Acordo sobre a Mobilidade entre os Estados-Membros da Comunidade dos Países de Língua Portuguesa, firmado em Luanda, em 17 de julho de 2021. **Diário Oficial da União**: seção 1, Brasília – DF, ed. Extra – A, p. 10, 29 jul. 2022. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2022/decreto/d11156.htm. Acesso em: 13 de julho de 2025.

CAÑAS, A. J. *et al*. CmapTools: A Knowledge Modeling and Sharing Environment. In: **Concept Maps**: Theory, Methodology, Technology, Proceedings of the First International Conference on Concept Mapping. Spain: Editorial Universidad Pública de Navarra, 2004. Disponível em: https://cmap.ihmc.us. Acesso em: 13 de julho de 2025.

KILGARRIFF, A. *et al*. **The sketch engine**. Proceedings of the 11th EURALEX International Congress: 105-116, 2004. Disponível em: https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2004.pdf. Acesso em: 13 de julho de 2025.

MAIA-PIRES, F. O. **Brasília em termos***:* um estudo lexical do Plano Piloto. 2009. Dissertação (Mestrado em Linguística) - Instituto de Letras, Universidade de Brasília, Brasília, 2009. Disponível em: http://www.realp.unb.br/jspui/handle/10482/10999. Acesso em: 13 de julho de 2025.

McENERY, T. & HARDIE, A. ***Corpus* linguistics**: method, theory and practice. Cambridge: Cambridge University Press, 2012.

SARDINHA, T. B. **Linguística de *Corpus***: histórico e problemática. **D.E.L.T.A.**, v. 16, n. 2, São Paulo, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005. Acesso em: 13 de julho de 2025.

SHEPHERD, T. M. G. Panorama da Linguística de *Corpus*. In: SHEPHERD, T. M. G.; SARDINHA, T. B.; PINTO, M. V.. (orgs). **Caminhos da Linguística de *Corpus***. Campinas: Mercado das Letras, 2012, p. 15-29.

SKETCH ENGINE. **Learn how languages works**, 2024. Disponível em: https://www.sketchengine.eu/. Acesso em: 13 de julho de 2025.

Artigo recebido em: 15.01.2025 Artigo aprovado em: 13.07.2025