CAPÍTULO 3

COLETA E TRATAMENTO DE DADOS NO *CORPUS* DO PORTUGUÊS

Ravena Beatriz de Sousa Teixeira
Jeane Nunes da Penha
(Universidade Federal do Rio de Janeiro)

3.1 INTRODUÇÃO

Com a popularização da internet, assim como dos aparatos tecnológicos, encontramos disponíveis, atualmente, diversas ferramentas capazes de facilitar o pesquisador no que concerne à coleta e ao tratamento de dados linguísticos. Nesse sentido, cabe-nos destacar aqui o papel da Linguística de *Corpus*, subárea da Linguística assim nomeada por ser utilizada, essencialmente, para a coleta, tratamento e análise de *corpora* linguísticos do âmbito escrito produzidos por falantes reais. Na Linguística de *Corpus*, todos os dados são selecionados prévia e criteriosamente por um ou mais estudiosos da área, a fim de facilitar na busca por dados de uma ou mais línguas e/ou variedades (cf. SARDINHA, 2000).

No Projeto PREDICAR (Formação e expressão de predicados complexos e predicações: estabilidade, variação e mudança construcional) algumas investigações já vêm adotando gerenciadores de dados *online*, como o *Corpus* do Português (https://www.

corpusdoportugues.org/),[1] que consiste em um banco de dados de *corpora* textuais em língua portuguesa, criado pelos professores Mark Davis e Michael J. Ferreira, das Universidades Brigham Young (BYU) e Georgetown, respectivamente. Desenvolvido para estudos de cunho linguístico, é constituído por uma interface que propicia ao pesquisador buscar por expressões exatas, lemas, trechos de textos e, inclusive, classes gramaticais.

O *Corpus* do Português (DAVIES, 2016) possibilita aos pesquisadores a realização de investigações em um número diversificado de sincronias, desse modo, mostra-se como uma ferramenta relevante para os linguistas que tencionam debruçar-se sobre a análise da língua portuguesa e/ou suas variedades. Com isso, a partir dessa breve apresentação do *Corpus* do Português, objetivamos: (i) mostrar os aspectos positivos e negativos de se trabalhar com bancos de dados online; (ii) ilustrar os diferentes *corpora* que compõem o gerenciador de dados *Corpus* do Português; (iii) explorar os passos metodológicos para a coleta no gerenciador de dados aqui focalizado; e (iv) auxiliar o pesquisador no tratamento dos dados linguísticos.

Dividimos este capítulo da seguinte maneira: após esta seção introdutória, apresentamos uma seção com os aspectos positivos e negativos de se utilizar bancos de dados *online*. Logo após, demonstramos os quatro *corpora* que constituem o *Corpus* do Português (Género/Histórico, Web/Dialetos, NOW e WordAndPhrase). Em seguida, detalhamos os passos metodológicos para a coleta de dados (Como se dá a coleta?; Como funciona o menu de busca?; Quais opções o pesquisador encontra na ferramenta de busca?). Por fim, expomos as considerações finais e apresentamos as referências que nortearam este trabalho.

3.2 QUAIS OS ASPECTOS POSITIVOS E NEGATIVOS DOS BANCOS DE DADOS ONLINE?

Reunimos nesta seção alguns pontos positivos e negativos – mais positivos – de se utilizar gerenciadores de *corpora* do meio digital para a coleta de dados linguísticos. Como bônus, o investigador encontra:

1) *Fácil acessibilidade* – os bancos de dados *online* estão disponíveis para pesquisadores do mundo inteiro. Mesmo cobrando uma taxa de cadastro, oferecem a opção de utilizarmos o gerenciador gratuitamente, durante um período de 30 dias, para teste/experiência;

2) *Rápido rastreamento de dados de diversas línguas e variedades* – os bancos de dados digitais processam, selecionam e armazenam milhares de dados linguísticos. Para o linguista que busca analisar o Português, o *Corpus* do Português é uma boa opção, pois disponibiliza dados de diferentes países que possuem a língua portuguesa como idioma oficial. Em contrapartida, se procura investigar outros idiomas, há ou-

1 Teixeira (2020) ilustra complexos verbo-nominais de carga semântica passiva coletados no *Corpus* do Português NOW, como *levar uma pancada, tomar uma pancada, sofrer uma pancada, receber uma pancada* e *ganhar uma pancada*.

tros gerenciadores de *corpora*, como o *Sketch Engine,* que, além do Português, ainda armazena dados do Inglês, Espanhol, Francês, Italiano, Chinês, dentre outros;

3) *Mais agilidade e menos esforço para a obtenção de dados linguísticos* – diferentemente de outras maneiras usadas por investigadores para a coleta de dados (leituras físicas ou digitais em demasiados textos, procuras incessantes por diferentes fontes, gêneros etc.), os bancos de dados permitem que o pesquisador se depare com milhares de resultados a partir de um *click*;

4) *Fácil seleção das fontes e/ou gêneros dos dados* – por armazenar dados de diferentes fontes e gêneros textuais, os bancos de dados *online* são úteis para o linguista que objetiva compor um *corpus* apenas com dados acionados em um determinado gênero textual, por exemplo, o gênero notícia. Nesse caso, ao invés de pesquisar múltiplos jornais, o linguista tem a opção de selecionar, na página de resultados, somente os dados oriundos de jornais e, assim, buscar os dados mobilizados no gênero em questão; e

5) *Informações sobre os dados* – os gerenciadores de *corpora* digitais oferecem ao investigador informações importantes sobre os dados, como a data de rastreamento, a variedade pertencente, endereço eletrônico da fonte em que o dado foi extraído etc.

Como nem tudo na vida é viver apenas de bônus, com os bancos de dados digitais não é diferente. Como ônus, o linguista defronta-se com:

1) *Páginas indisponíveis* – alguns dados encontrados nos resultados de busca podem não estar mais com o endereço eletrônico disponível. Cabe lembrar que se um dado provém de um *corpus* que foi rastreado três anos antes da ação da coleta é importante se ter em vista que algumas páginas tenham saído do ar/não existam mais; e

2) *Política de privacidade de determinadas fontes* – como mencionado na introdução, os bancos de dados digitais reúnem *corpora* linguísticos do meio escrito produzidos por falantes reais. Sendo assim, o investigador pode notar que algumas páginas apresentam uma política de privacidade seguida de bloqueio do conteúdo, a fim de garantir a não disseminação ou plágio do teor ali apresentado. Os gerenciadores de dados *online*, geralmente, demonstram os dados no contexto discursivo de aplicação, por isso, cabe ao pesquisador decidir se coleta o dado no pequeno trecho apresentado mas não o expõe, ou contabiliza no *corpus*, ou se o ignora.

Dado o exposto, notamos que, embora o procedimento de coleta de dados em bancos digitais apresente alguns pontos negativos, os aspectos positivos são predominantes, o que facilita o investigador na hora de optar por um método mais ágil.

3.3 *CORPUS* DO PORTUGUÊS: CONHECENDO OS *CORPORA*

Ao acessar o site do *Corpus* do Português, é possível notar que o banco de dados é composto por duas partes distintas. De acordo com a página de apresentação, encontramos:

> *um corpus (original e menor) que permite ver as mudanças históricas assim como variações de gênero; um corpus (novo e muito maior) que permite verificar as variações dialéticas (e tem 50 vezes mais dados do português moderno).*

Como observado na Figura 1, o *Corpus* do Português conta com quatro abas de pesquisas dos *corpora*: Género/Histórico, Web/Dialetos, NOW e WordAndPhrase, descritos a seguir.

Figura 1: *Corpora* disponíveis no banco de dados online *Corpus* do Português.

		Corpus	Tamanho	Criado
1	Info	Género / Histórico	45 milhões de palavras	2006
2	**Info**	**Web / Dialetos ***	**1 mil milhão de palavas**	**2016**
3	Info	NOW (2012 - 2019)	1,1 mil milhão de palavras	2018
4	Info	WordAndPhrase (agora parte do #2)	40.000 palavras principais	2017

Género / Histórico

Apresentado como o "original" do *Corpus* do Português, possui duas interfaces (2006 e 2016). É constituído por uma base de dados com 45 milhões de palavras entre os séculos XIII e XX – o que o torna importante para o pesquisador que deseja investigar a Língua Portuguesa desde uma perspectiva diacrônica. No que diz respeito aos dados do século XX, estes se dividem, igualmente, entre os gêneros de estilo falado, ficção, textos acadêmicos e provenientes de jornais digitais. A versão mais recente (2016) permite-nos criar *corpora* virtuais, por exemplo, por um determinado conjunto de fontes, tópicos etc.

Web / Dialetos

Interface desenvolvida em 2016, composta por uma base de dados com cerca de 1 milhão de palavras de páginas digitais oriundas de quatro países de língua portuguesa: Brasil, Portugal, Angola e Moçambique. Os textos foram rastreados entre os anos de 2013 e 2014, por isso permite que o pesquisador realize uma análise sincrônica do *corpus*, além possibilitar a comparação entre as diferentes variedades da língua portuguesa.

News on the Web (NOW)

A aba NOW é a interface mais recente do *Corpus* do Português (agosto de 2018) e está constituída por mais de 1,1 milhão de palavras das distintas variedades do português (Brasileira, Europeia, Africana). De acordo com a página de apresentação, todos os meses são adicionadas ao *corpus* cerca de 35 milhões de palavras oriundas de revistas e jornais digitais.

WordAndPhrase

Essa interface, não mais disponível a partir de 2022, proporcionava ao investigador pesquisar e navegar por 40 mil palavras do português com base na frequência do *corpus*. Para cada palavra buscada, o investigador encontrava disponível diversas informações, como sua definição, sinônimos, gênero, país, colocação, concordância, dentre outras. Além disso, tornava possível inserir e analisar textos completos, destacar palavras-chaves ou frases no texto e realizar uma pesquisa com frases relacionadas em todo o *Corpus* do Português. Assim, considerando a importância de suas ferramentas de busca para a análise linguística, suas funções, em janeiro de 2022, passam a ser incorporadas nos recursos da aba Web/Dialetos.

Agora que nos familiarizamos com todas as abas de pesquisa disponibilizadas no *Corpus* do Português, daremos início aos procedimentos metodológicos – busca, tratamento e análise de amostras linguísticas presentes no banco de dados.

3.4 A PLATAFORMA DE BUSCA E O PROCESSO DE COLETA

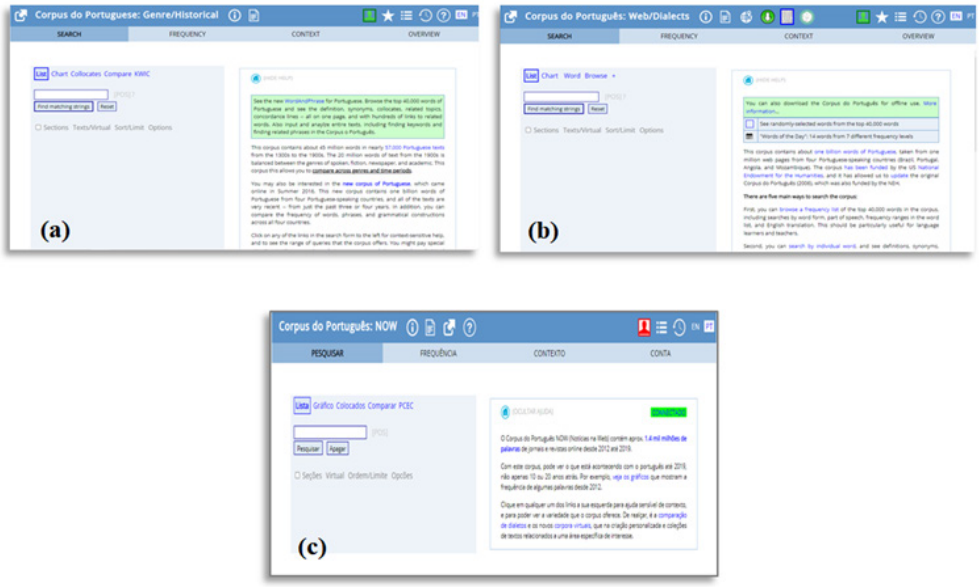
Dentre os aspectos essenciais, ao se tratar do processo de análise linguística, encontra-se a delimitação das características e fonte dos dados a serem considerados pelo pesquisador. A seleção de um (ou mais) banco(s) de dados deve refletir as diretrizes teórico-metodológicas da pesquisa a ser levada a cabo. Entretanto, um dos desafios enfrentados por pesquisadores é a falta de familiaridade com as interfaces de busca propiciadas pelas plataformas associadas aos *corpora online*.

Assim, como indicado anteriormente, nesta seção, buscamos apresentar, de forma sucinta, as ferramentas de busca oferecidas pelas abas do *Corpus* do Português, em especial do *Corpus* do Português NOW (aba mais recente do banco de dados). Inicialmente, apresentaremos os atributos do menu de busca de cada uma das abas, quais são as ferramentas de busca oferecidas e como, de fato, se dá seu funcionamento, para, em seguida, visualizarmos, segundo exemplos, quais são os passos metodológicos a serem considerados no processo de busca.

3.4.1 FAMILIARIZANDO-SE COM A PLATAFORMA DE BUSCA

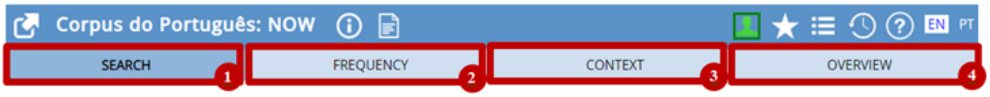
Em relação à sua configuração, as abas *Gênero/Histórico*, *Web/Dialeto* e *NOW* compartilham a mesma interface de busca, conforme observável nas Figuras 2.a, 2.b e 2.c.

Figura 2: Páginas iniciais da interface de busca das abas do banco de dados *online Corpus* do Português.



Como nosso intuito é, primordialmente, apresentar um tutorial prático do processo de coleta de dados, em especial de predicadores complexos compostos por verbo (semi-)suporte, podemos visualizar, a seguir, uma descrição dos elementos que compõem as janelas de pesquisa das três referentes abas, constituídas por quatro janelas principais: a janela inicial de busca (1); a janela de disposição inicial dos resultados (2); a janela de contexto (3), na qual temos acesso à lista de dados referentes ao resultado da busca; e a janela de ajuda (4).

Figura 3: Janelas referentes ao processo de pesquisa.

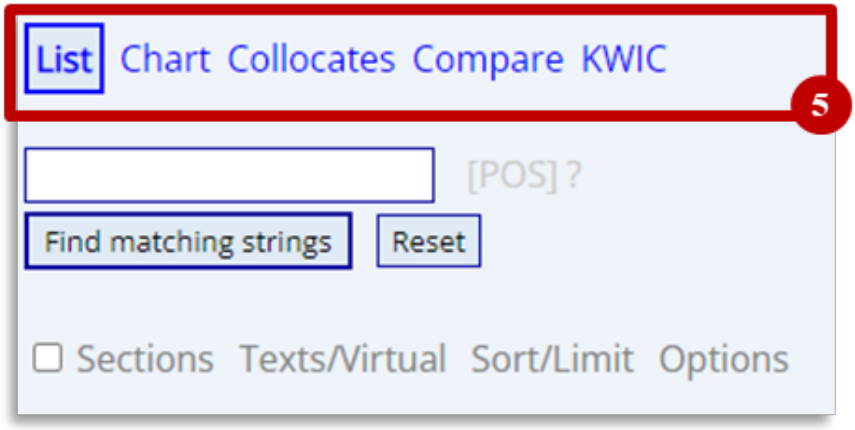


Durante o processo de busca, a depender das ações do usuário, as distintas janelas são automaticamente acionadas pela plataforma. Segundo uma ação em cadeia, a partir de uma busca efetuada na janela pesquisa, a janela frequência é utilizada. Consecutivamente, a partir de uma seleção desempenhada em frequência, a janela de contexto entra em atuação. Dessa forma, a *ativação* de cada uma das áreas da plataforma se apresenta interligada.

Além disso, para o acesso aos recursos do banco de dados, ratificamos a importância do processo de cadastro. É apenas por meio da efetivação do cadastro na plataforma que a realização de buscas pode ser concretizada.

3.4.1.1 A janela de pesquisa e a distribuição dos dados no campo de frequência: o menu e suas funções

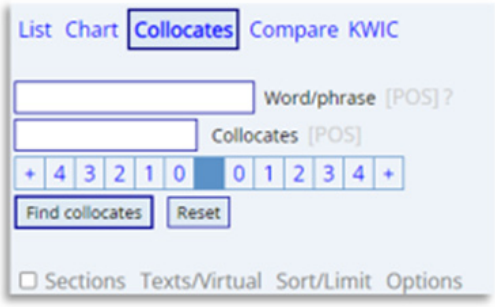
Figura 4: Menu principal da plataforma de busca do *Corpus* do Português.



Na parte superior do menu principal, em 5, temos as distintas opções de busca oferecidas pela plataforma. A opção *List* possibilita a busca por estruturas simples, como o lexema "soco", ou expressões complexas, como "levar um soco", apresentando os resultados da pesquisa efetuada em forma de uma lista. *Chart* assim como em *List*, possibilita uma busca por elementos simples ou expressões; entretanto, o resultado é exposto em forma de gráficos.

A opção *Collocates*, por sua vez, possibilita a busca por combinações, independente do grau de complexidade/tamanho das expressões em jogo. Por meio de seu uso, podemos observar quais elementos ocorrem um com o outro.

Figura 5: Menu de busca da plataforma do *Corpus* do Português NOW segundo a opção *collocates*.



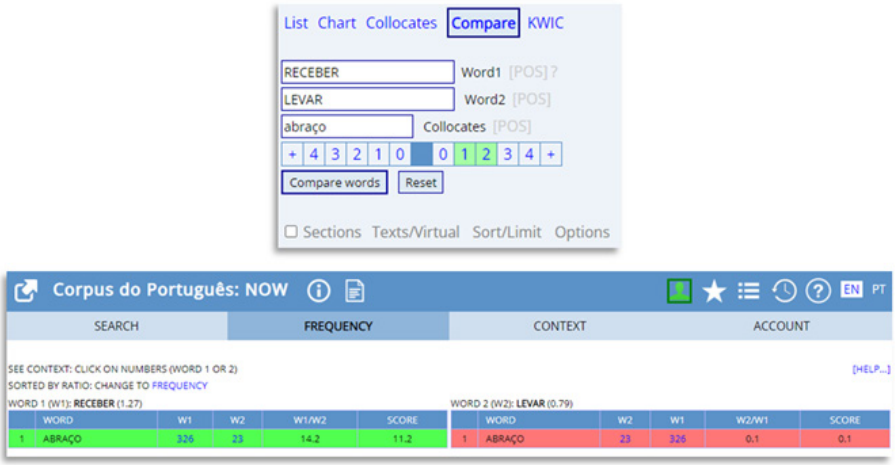
Contamos, nessa opção, com dois espaços de inserção para busca de elementos. O *Word/phrase* é utilizado para determinar qual item será considerado "principal" na combinação e será o referente para o raio/alcance de busca da segunda expressão, apresentada em *collocates*. Abaixo desse, observamos uma escala que determinará esse alcance – a direção e distância do colocado em relação ao referente. O quadrado totalmente preenchido em azul figurará como a expressão exposta em *Word/phrase* e os números, como a posição do item em *collocates* ao se tratar desse referente. A título de exemplificação, observemos a figura a seguir:

Figura 6: Exemplo de busca na opção *collocates* e de seus respectivos resultados.

Ao buscarmos pela combinação "RECEBER" – em *Word/phrase* – mais "abraço" – em *collocates* –, selecionando o número 3 à direita em tal escala, a plataforma me apresentará dados/fragmentos textuais nos quais o lexema "abraçar" aparece até 3 posições após o uso do verbo "receber", como em "receber abraço", "receber um abraço" e "receber um grande abraço".

Em *compare*, pesquisamos por colados associados a dois itens referentes, possibilitando a comparação entre suas configurações estruturais e funcionais. Na Figura 7, temos a busca do colocado "abraço" em relação aos lexemas verbais referentes "RECEBER" e "LEVAR".

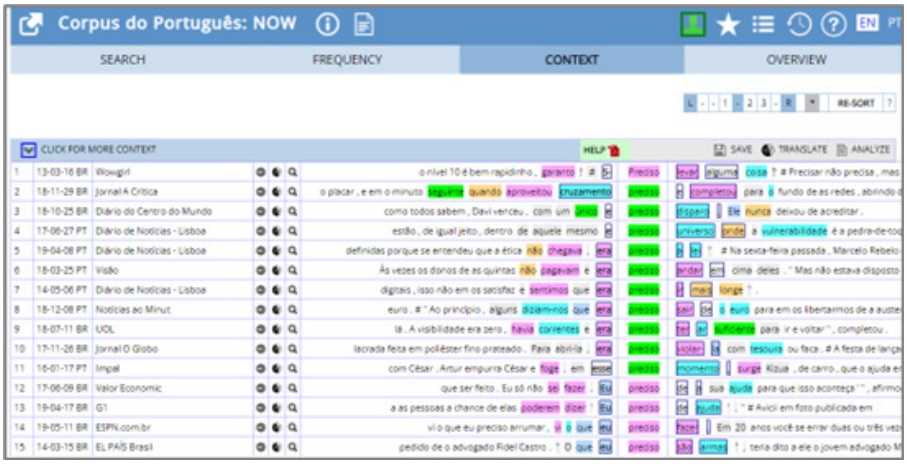
Figura 7: Exemplo de busca na opção *compare* e de seus respectivos resultados.



O vocábulo "abraço" vê-se mais empregado em associação ao lexema verbal RECEBER em detrimento a LEVAR. Tal resultado ratifica indícios observados por Teixeira (2020) ao se tratar de uma possível restrição colocacional do uso do verbo *levar* na configuração de predicadores complexos de passividade. Seu uso vê-se mais associado a elementos nominais de natureza negativa.

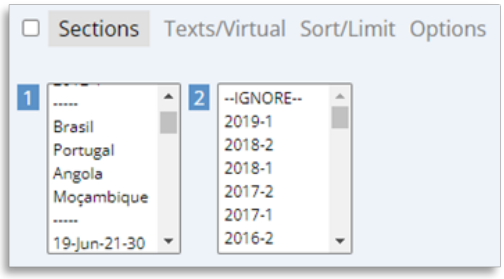
Keyword in Context, denominado *KWIC*, propicia averiguar os padrões nos quais os itens/expressões ocorrem no *corpus*.

Figura 8: Exemplo de resultados para a realização de busca pelo vocábulo "preciso" na opção *KWIC*.



A opção *Sections*, por sua vez, viabiliza delimitar em qual(is) seção(ções) – gênero, recorte temporal e/ou dialeto – se deseja efetuar a busca, ensejando a possibilidade de comparar a ocorrência de um objeto em duas seções distintas.

Figura 9: Exemplo de recursos propiciados pela opção *Sections* no menu de busca.



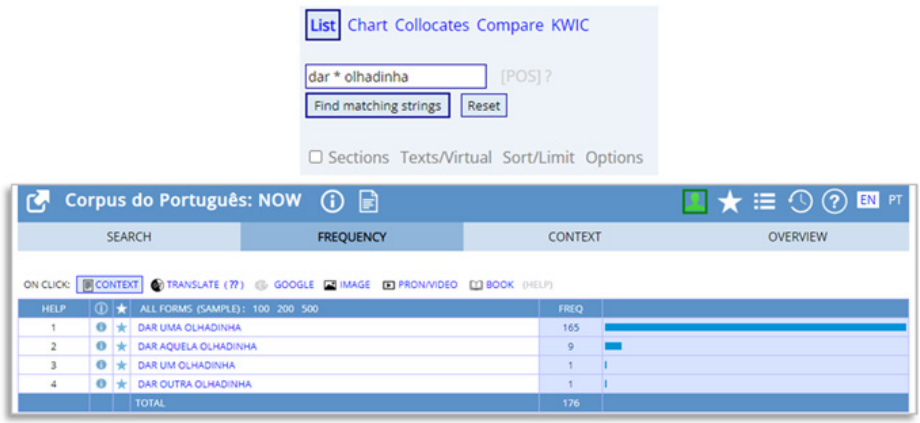
3.4.1.2 Ferramentas de pesquisa

Dentre as ferramentas proporcionadas pela plataforma a fim de otimizar a experiência do usuário/pesquisador, encontra-se a sintaxe específica de busca. Segundo o uso de símbolos especiais, códigos preestabelecidos referentes à distribuição dos elementos em categorias gramaticais (PoS ou _pos), a busca por lemas e listas personalizadas de palavras/expressões, podemos acionar comandos de pesquisa de forma rápida, além de ampliar o escopo de resultados.

3.4.1.2.1 Símbolos especiais

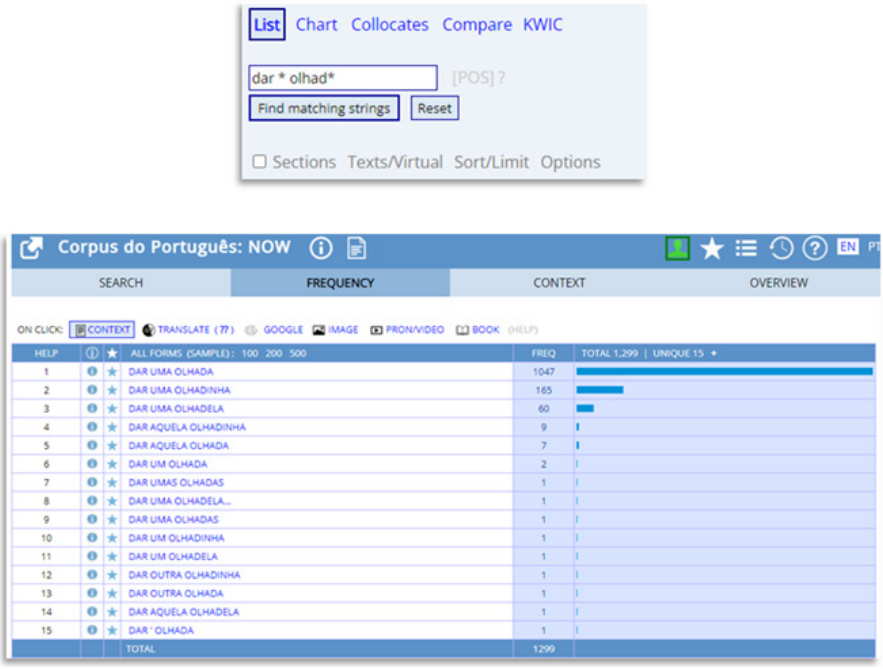
No conjunto de itens especiais, contamos com: o asterisco (*), a barra vertical (|), o símbolo de igual (=) e o símbolo de menos (-). O asterisco (*) se figura como um caractere curinga. Estipula que qualquer espaço/*slot* com a sua presença pode ser preenchido por qualquer elemento. Ao ser empregado de forma independente, antes ou depois de uma palavra, determina que o espaço ocupado pelo asterisco, na expressão, pode estruturalmente representar qualquer lexema.

Figura 10: Exemplo de emprego do símbolo asterisco (*) em um comando de busca.



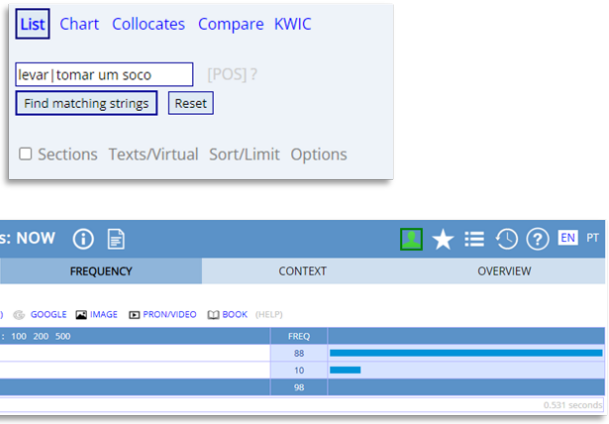
Na Figura 10, ao empregarmos o comando de busca "dar * olhadinha", a plataforma nos propicia ocorrências de uso com expressões como "dar uma olhadinha" e "dar aquela olhadinha", nas quais vemos um elemento entre o verbo "dar" e o item "olhadinha". É importante, nesse caso, realçar a importância do *espaçamento* no processo de pesquisa. O asterisco (*) somente foi lido/compreendido pelo algoritmo de busca como um lexema independente, como uma palavra, devido ao espaço existente entre este e as unidades "dar" e "olhadinha" no comando de pesquisa. Caso esteja acoplado a uma palavra, representará que tal espaço no vocábulo poderá ser preenchido por qualquer caractere ou conjunto de caracteres. Ao delimitarmos, por exemplo, como comando a linha "dar * olhad*", teremos resultados como "dar uma olhada", "dar uma olhadinha", "dar uma olhadela", "dar aquela olhadinha" e "dar aquela olhada" (cf. Figura 11).

Figura 11: Exemplo de emprego do símbolo asterisco (*), de forma independente e associado a uma palavra, em um comando de busca.



A barra vertical (|), por sua vez, indica alternância. Efetiva uma busca por mais de uma opção de elementos. Ao delimitarmos o comando "levar|tomar um soco", temos como resultados as expressões "levar um soco" e "tomar um soco".

Figura 12: Exemplo de emprego do símbolo barra vertical (|) em um comando de busca.

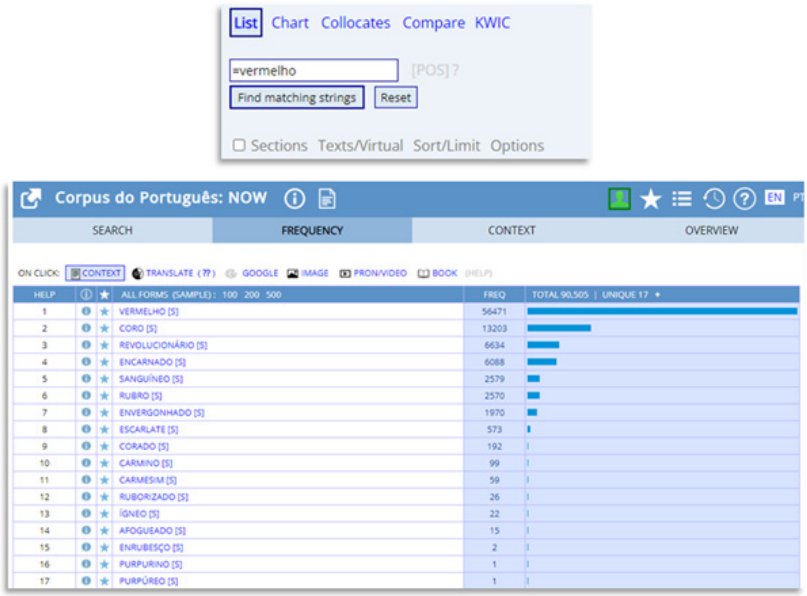


Nesse contexto, chamamos, novamente, atenção para o espaçamento. Não deve haver nenhum espaço entre os elementos de busca alternantes e o símbolo (|). Caso exista, a plataforma indicará que há um erro na sintaxe de busca, pois a expressão não será encontrada dentre o banco de dados (cf. Figura 13).

Figura 13: Exemplo de erro indicado pela plataforma.

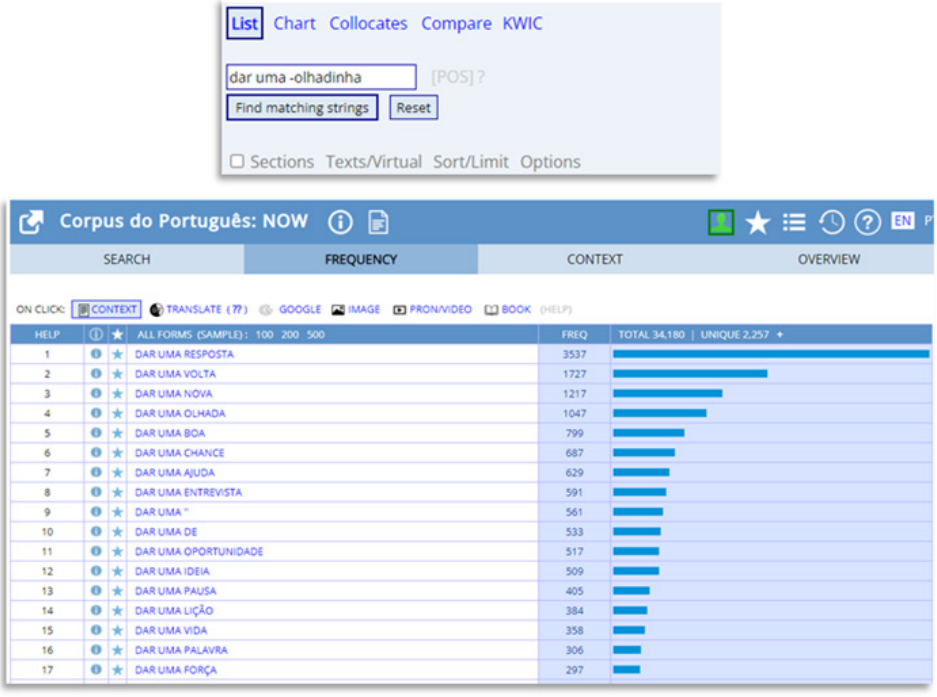
O símbolo de igual (=), ao ser aplicado antes de uma palavra, possibilita a busca por seus sinônimos. Logo, por meio de seu emprego, teremos acesso tanto aos usos de sinônimos dessa palavra no interior do *corpus* quanto de seus usos em si. Ao pesquisarmos por "=vermelho", obtemos como resultados expressões como "coro", "rubro" e "escarlate".

Figura 14: Exemplo de emprego do símbolo de igual (=) em um comando de busca.



Já o símbolo de menos (-) indica uma restrição de busca. Na pesquisa, serão inclusos todos os elementos possíveis, menos aquele precedido pelo sinal de menos. Na busca "dar uma –olhadinha", mesmo que a expressão "dar uma olhadinha" seja produtiva no *corpus*, a plataforma apresentará somente o uso de expressões compostas por "dar uma" mais elementos que não sejam "olhadinha".

Figura 15: Exemplo de emprego do símbolo de menos (-) em um comando de busca.



3.4.1.2.2 Codificação de *Part of Speech* (POS e/ou _pos)

Uma das características marcantes da organização textual do *Corpus* do Português é o fato de se figurar como um banco de dados de *corpora* anotados. Cada uma de suas unidades vocabulares recebe *tags*/etiquetas referentes às suas características estruturais e funcionais (segundo um ponto de vista (morfo-)sintático). Como exemplo, vemos a segmentação dos lexemas em conjuntos de acordo a sua classe – em verbos, substantivos pronomes, dentre outras –, a sua função sintática na sentença – sujeito ou objeto – e a atributos morfológicos relativos ao seu número, gênero, e no caso dos itens verbais, pessoa – singular ou plural, feminino ou masculino e primeira, segunda ou terceira pessoa verbais. Assim, é possível fazer uso de códigos/*tags* que possibilitam a restrição/especificação dos itens que irão compor nossa busca.

Tais códigos são apresentados na plataforma por meio da função *Parts of Speech* (POS ou _pos). Nessa, classificações de partes do discurso/texto são utilizadas como recursos de busca. A seguir, podemos visualizar exemplos de códigos empregados na plataforma segunda essa função:

Tabela 1: Exemplos de códigos empregados no *Corpus* do Português via função *Parts of Speech* (POS e _pos)

Códigos PoS (versões longas e abreviadas)	Códigos _pos	Tipo de itens pesquisados	Exemplos de resultados
NOMES			
NOUN / **N**	_n	Substantivos comuns em geral	bola, mesa, soco, fome
NMS	_nms	Substantivos comuns masculinos no singular	caderno, carrinho, menino, homem
NMP	_nmp	Substantivos comuns masculinos no plural	cadernos, carrinhos, meninos, homens
NFS	_nfs	Substantivos comuns femininos no singular	borracha, menina, bicicleta, mesa
NFP	_nfp	Substantivos comuns femininos no plural	borrachas, meninas, bicicletas, mesas
O / **NAME**	_o	Substantivos próprios	Vitória, Copa, Maria, Roberto
DETERMINANTES			
DET / **D**	_d	Determinantes em geral	seu, este
DD	_dd	Determinantes demonstrativos	isto, aquilo, este
DP	_dp	Determinantes possessivos	meu, nosso, seu
ARTIGOS			

ART	_l	Artigos em geral	a, as, um, uns
L			
LD	_ld	Artigos definidos	a, as, o, os
LI	_li	Artigos indefinidos	um, uns, uma, umas
PRONOMES			
P	_p	Pronomes em geral	que, se, ele, ela
PD	_pd	Pronomes demonstrativos	isto, isso, aquilo
PI	_pi	Pronomes indefinidos	nada, algo, alguém
PO	_po	Pronomes pessoais com função de objeto	os, me, lhe
PS	_ps	Pronomes pessoais com função de sujeito	eu, ele, ela
PR	_pr	Pronomes relativos	que, quem, cujo
PREPOSIÇÕES			
PREP	_e	Preposições em geral	para, com, de
E			
CONJUNÇÕES			
CONJ	_c	Conjunções em geral	porque, pois, embora
C			
ADJETIVOS			
ADJ	_j	Adjetivos em geral	maior, grande, melhor
J			
	_jms	Adjetivos masculinos no singular	novo, necessário, bom
	_jmp	Adjetivos masculinos no plural	novos, necessários, bons
	_jfs	Adjetivos femininos no singular	nova, necessária, boa
	_jfp	Adjetivos femininos no plural	novas, necessárias, boas
	_jcs	Adjetivos uniformes no singular	possível, importante, nacional
	_jcp	Adjetivos uniformes no plural	possíveis, importantes, nacionais
ADVÉRBIOS			
ADV	_r	Advérbios em geral	não, também, ainda
R			
INTERJEIÇÕES			
I	_i	Interjeições em geral	caramba, ah, ei
NUMERAIS			

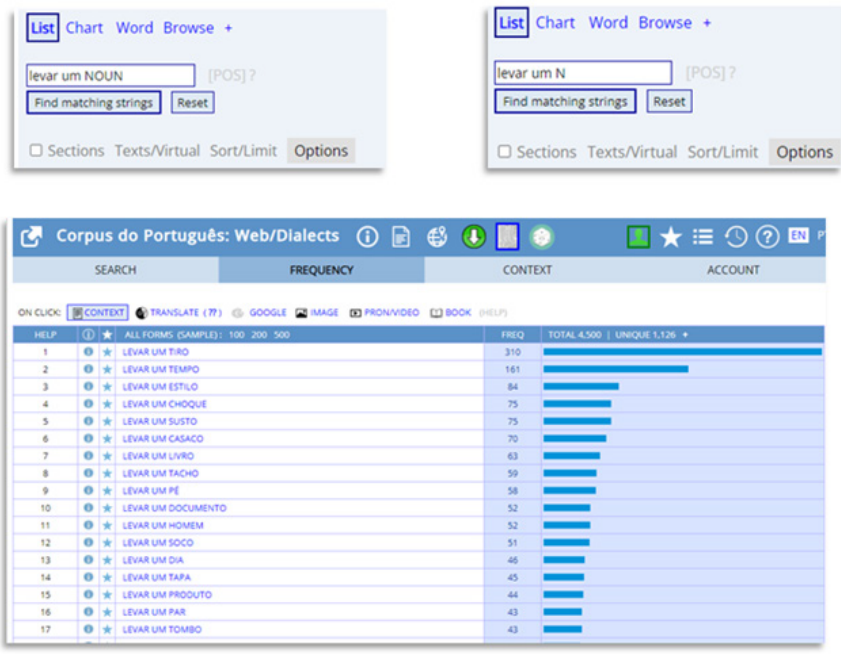
NUM	_m	Numerais em geral	dois, 6, primeiro
M			
MO	_mo	Números ordinais	último, segundo, terceiro
MC	_mc	Números cardinais	3, quatro, 20
VERBOS			
VERB	_v	Verbos em geral	receber, ter, levar
V			
VIP	_vip	Verbo no presente do indicativo	é, está, vai
VIF	_vif	Verbo no futuro do indicativo	levará, estarei, comeremos
VIS	_vis	Verbo no pretérito do indicativo	foi, disse, afirmou
VII	_vii	Verbo no pretérito imperfeito do indicativo	estava, eram, queria
VSP	_vsp	Verbo no presente do subjuntivo	seja, possa, tenham
VSJ	_vsj	Verbo no pretérito imperfeito do subjuntivo	fossem, tivesse, ficássemos
VPP	_vpp	Verbo no particípio presente	sendo, incluindo, fazendo
VPS	_vps	Verbo no particípio passado	feito, realizado, publicado
VC	_vc	Verbo condicional	teria, poderia, estaria
VR	_vr	Verbo no infinitivo	ser, ter, estar
CLASSIFICAÇÃO DE PESSOA E NÚMERO VERBAL			
X[2] -1S	_X-1s	Verbo na primeira pessoa do singular	como, tivesse, andei
X -2S	_X-2s	Verbo na segunda pessoa do singular	tens, pertences, estás
X -3S	_X-3s	Verbo na terceira pessoa do singular	tem, come, anda
X -1P	_X-1p	Verno na primeira pessoa do plural	comemos, tivéssemos
X -2P	_X-2p	Verbo na segunda pessoa do plural	ficais, moveis
X -3P	_X-3p	Verbo na terceira pessoa do plural	têm, comem, andam

Fonte: Autoral.

2 Nesse caso, o elemento "X" não se figura como parte estrutural do código. Apenas representa a possibilidade de preenchimento do *slot* a ele relativo com qualquer um dos códigos previamente expostos referentes ao tempo e modo verbais (vip, vsp, vpp, dentre outros).

Podemos fazer uso de dois tipos de codificação de *Part of Speech*: *POS* ou *_pos*. As *tags POS* são empregadas em letras maiúsculas e fazem referência a tipos de lexemas. Utilizamos códigos *POS* no lugar de vocábulos a fim de efetuarmos uma busca ampla pautada na classificação sintática do tipo de item cuja realização desejamos averiguar no banco de dados.

Figura 16: Exemplo de emprego dos códigos *POS* "NOUN" e "N" em comandos de busca e seus respectivos resultados.

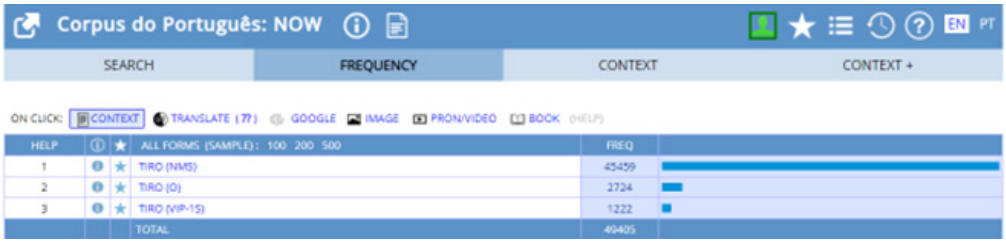


Na Figura 16, temos os comandos "levar um NOUN" e "levar um N", nos quais empregamos, no *slot* relativo ao item não verbal a compor o padrão composto pelo verbo "levar", os códigos "NOUN" e "N". Assim, indicamos ao algoritmo de busca da plataforma uma especificação em relação à configuração do tipo de expressões cuja realização desejamos observar: dados constituídos pelo verbo "levar", o artigo "um" e qualquer item caracterizado como substantivo, nessa determinada ordem. Como resultados, encontramos "levar um tiro", "levar um tempo", "levar um estilo", dentre outros, em que o terceiro elemento é um substantivo, independente da natureza da expressão em si.

Os códigos *_pos*, em contrapartida, não possibilitam, primordialmente, a substituição de um item de busca por uma codificação pautada nos atributos que desejamos delimitar, mas sim, restringir o campo de atuação de um item predeterminado. Considerando a multifuncionalidade que elementos linguísticos e/ou padrões estruturais podem desempenhar a depender de seu contexto de uso, podemos, segundo o uso de códigos *_pos*, especificar os atributos dos elementos que fazem parte do nosso interesse de pesquisa.

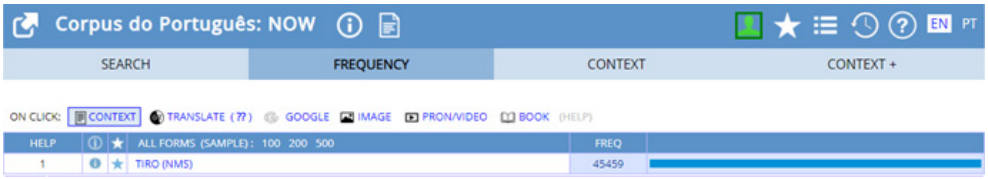
Como exemplo de padrão multifacetado, é possível observar o padrão vocabular "tiro", que pode se referir, dentro dos *corpora*, conforme a Figura 17, a um elemento nominal no masculino singular (*levar um tiro*), a um elemento verbal na primeira pessoa no singular do presente do indicativo (*Eu tiro o lixo todo os dias*) ou a parte de um nome próprio (*O alistamento é feito em a Junta de o Serviço Militar, em o Tiro de Guerra*). Assim, caso desejemos efetuar uma busca que enseje a análise do uso do lexema nominal "tiro", é oportuno realizarmos especificações no nosso comando de busca. Se indicarmos como comando de pesquisa simplesmente o padrão "tiro", encontraremos tanto usos referentes ao substantivo "tiro" quanto ao verbo "tirar" na primeira pessoa no singular do presente do indicativo.

Figura 17: Exemplo de resultados para o comando de busca "tiro".



Logo, é possível restringirmos nosso escopo de busca segundo o comando "tiro_nms", por meio do qual nos serão apresentados resultados referentes somente ao uso do substantivo no masculino singular "tiro".

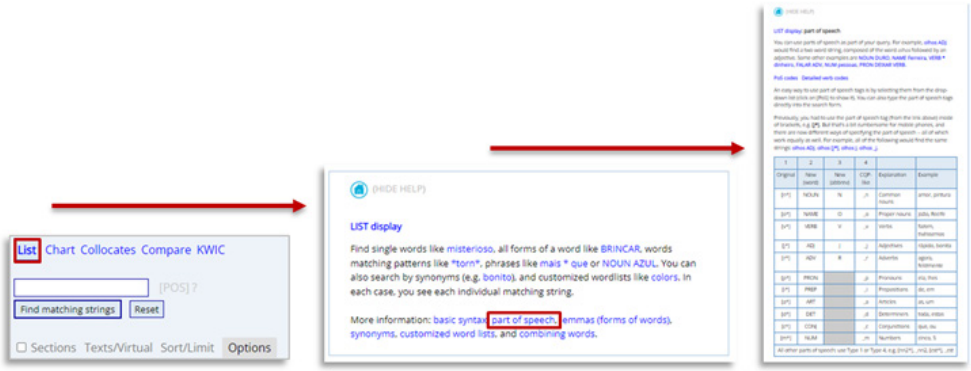
Figura 18: Exemplo de resultados para o comando de busca "tiro_nms".



Nos casos citados, temos conhecimento prévio dos marcadores utilizados para classificar os elementos linguísticos em pauta. Entretanto, nem sempre isso é possível. Logo, é essencial considerarmos meios para identificar a categorização dos itens que nos interessam no interior dos *corpora*. Para isso, contamos com dois caminhos metodológicos:

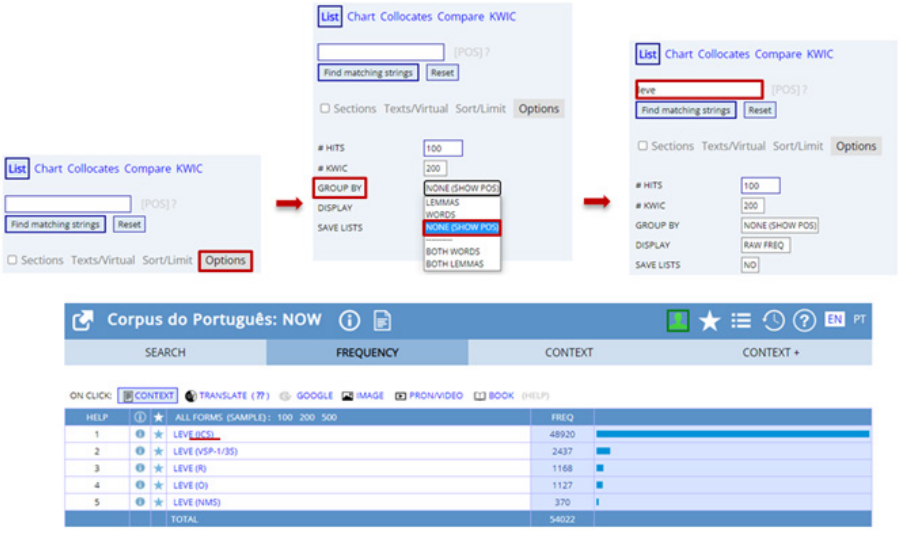
(1) Ir no menu de busca, selecionar a opção "*list*", que abrirá a caixa de ajuda referente a essa função na parte direita da tela e, em seguida, clicar na opção "*part of speech*", a qual promoverá acesso a uma tabela com os códigos *POS* e *_pos* básicos utilizados na plataforma, assim como uma breve explicação de seus usos em conjunto com exemplos diversificados.

Figura 19: Procedimento metodológico para acessar a caixa de ajuda referente aos códigos *POS* e *_pos.*



(2) No menu de busca, clicar em "*Options*", referente à forma na qual os resultados são expostos na janela de frequência dos resultados, e, em "*GROUP BY*", selecionar a opção "*NONE (SHOW POS)*". Em sequência, é necessário apenas digitar o vocábulo que deseja saber a codificação/*tag* e visualizar os resultados na janela de frequência, que serão organizados com base na classificação do vocábulo no banco de dados. Aparecerá o vocábulo e, ao final, entre parênteses, sua classificação/codificação.

Figura 20: Procedimento metodológico para acessar os resultados de buscas segundo seus códigos *POS*.

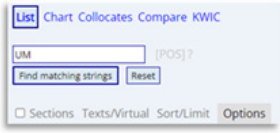


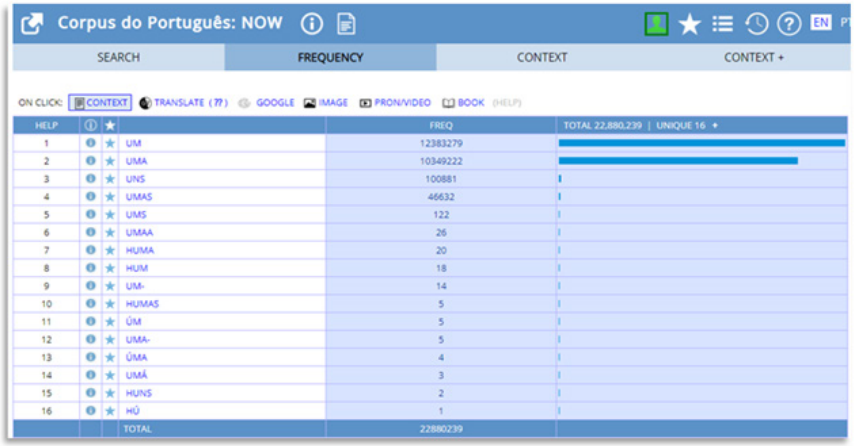
3.4.1.2.3 Uso de lemas

Outra ferramenta é a utilização de uma unidade lexical em letras maiúsculas a fim de efetuar buscas por formas flexionadas que se encontrem associadas a essa unidade base. Todas as unidades linguísticas do banco de dados encontram-se vinculadas a um item lexical que funciona como seu representante, de forma similar a uma entrada de dicionário. Por exemplo, a estrutura verbal no infinitivo "receber", em um dicionário, atua como representante de todos os itens que se encaixam nesse paradigma verbal, como "recebi", "recebeu", "recebido", dentre outros. Na plataforma, o mesmo ocorre. Temos como efetuar buscas a partir de uma unidade representante, e isso não se restringe a pesquisas centradas em composições verbais.

Ao realizarmos uma pesquisa segundo o comando "UM", no qual fazemos uso do artigo indefinido masculino em letras maiúsculas, conforme a Figura 21, o determinante será compreendido como um lema. Dessa forma, na janela de resultados, a plataforma nos apresentará dados de usos relativos a todas as unidades que são associadas a esse item representante, como "um", "uma", "uns", "umas".

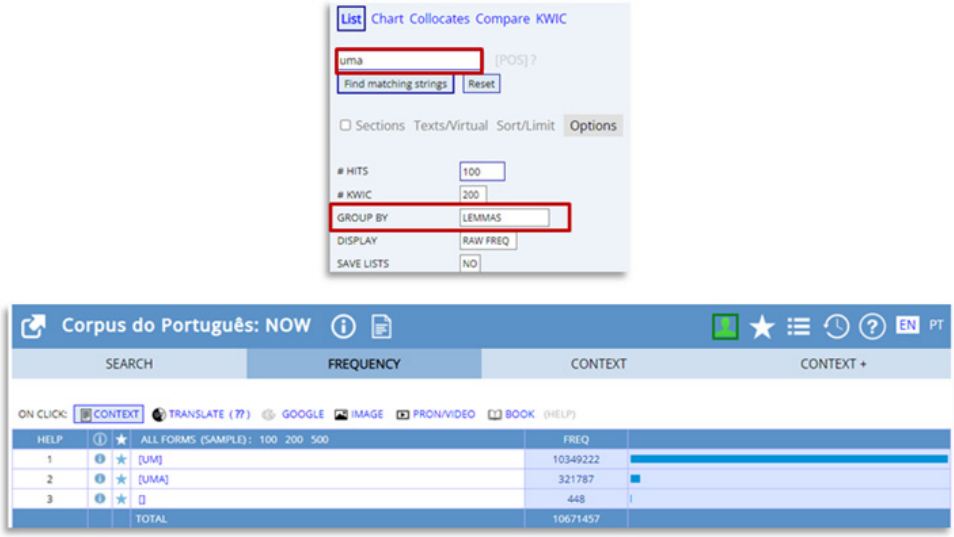
Figura 21: Exemplo de busca pelo lema "UM" e seus respectivos resultados.





Para acessar informações referentes aos lemas das unidades linguísticas, contamos com a função "*options*" disposta no menu de busca principal. É necessário apenas clicar em "*options*" e, em "*GROUP BY*", selecionar a opção "*LEMMAS*", que, ao efetuar uma busca por uma expressão, os lemas a que se vê associado serão apresentados entre colchetes como resultados na janela de frequência.

Figura 22: Procedimento metodológico para acessar o(s) código(s) *lemma* associados ao artigo "uma".



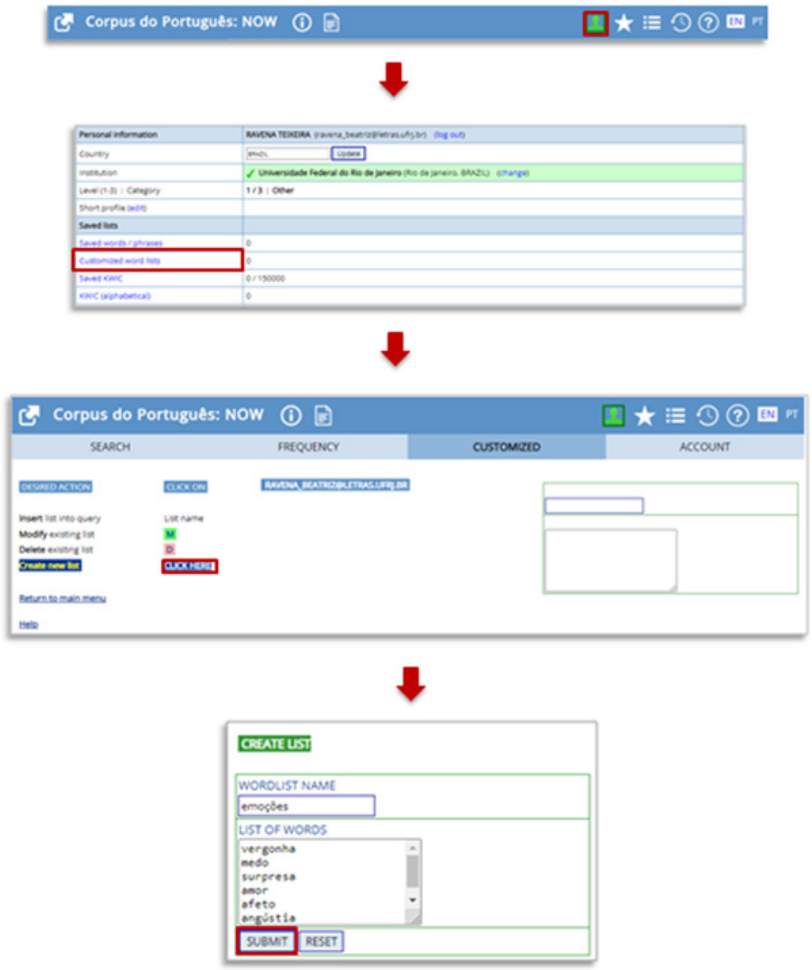
No exemplo exposto na Figura 22, podemos perceber que o artigo indefinido feminino "uma" encontra-se relacionado a mais um lema: "UM", "UMA" e unidades vazias (casos em que a lematização não foi especificada pelo sistema). Por conseguinte, chamamos atenção para o fato de que uma unidade pode estar associada a mais de um representante. Logo, no uso da categorização de lema no processo de pesquisa, é essencial, a fim de considerar por total os recursos dos *corpora*, verificar, segundo os passos previamente expostos, se os vocábulos a serem focalizados na busca se veem associados a mais de um item base/representante.

3.4.1.2.4 Listas de busca personalizadas

A plataforma também possibilita a criação de listas de busca personalizadas, ou criadas do zero pelo próprio usuário ou salvas segundo resultados de buscas, além daquelas já disponibilizadas como exemplos. Cada uma das listas desenvolvidas será salva no banco de dados do perfil do usuário e poderão ser acessadas independente do momento ou local de uso da plataforma.

Para criar uma lista personalizada, primeiramente é necessário acessar a página relativa ao perfil do pesquisador na plataforma – possível por meio do botão "*login*", em verde, na parte superior da tela –, clicar em "*Customized word lists*", em seguida, clicar em "*CLICK HERE*", inserir, nos quadros apresentados à direita da tela, o nome da lista de busca e os lexemas/vocábulos que deseja incluir na lista e salvar as informações por meio do botão "*SUBMIT*", conforme o esquema a seguir:

Figura 23: Procedimento metodológico para criar uma lista de busca personalizada.



Na lista, cada um dos vocábulos a ser considerado deve ser disposto sozinho, um abaixo do outro, em linhas diferentes, de forma que o sistema possa compreendê-los como unidades distintas. Além disso, é importante ratificar que apenas itens simples podem ser adicionados na construção de uma lista (*bola*, *carrinho*, *bicicleta*, entre outras). Expressões complexas, formadas por mais de um elemento lexical, estes divididos por espaçamento, não são aceitas pelo sistema.

3.4.1.2.5 Combinações entre ferramentas de busca

Considerando os recursos associados à sintaxe de busca propiciada pela interface do banco de dados, é essencial tratar da combinação entre estes. O usuário pode, no momento de construção do comando de pesquisa, lançar mão de mais de uma ferramenta a fim de alcançar os seus objetivos e otimizar o processo de trabalho. A seguir, temos exemplos de comandos nos quais uma associação entre ferramentas é observada:

Quadro 2: Exemplos de comandos de busca com a presença de mais de uma ferramenta de pesquisa

Comando	Explicação	Exemplos de resultados
LEVAR uma *ada	Vê-se o uso do lema "LEVAR", possibilitando a busca por todas as formas flexionadas desse paradigma verbal e o emprego do caractere curinga (*), propiciando uma busca por palavras terminadas por "ada".	levar uma facada, levou uma facada, levei uma bolada, levando uma pancada
levar|tomar uma *ada	Utiliza-se a barra vertical, para indicar a possibilidade de busca por um elemento verbal ou outro, e o caractere (*), delimitando uma busca por palavras terminadas em (ada), já que o mesmo viabiliza o preenchimento de sua posição por qualquer conjunto de letras.	levar uma pancada, levar uma bolada, levar uma década, tomar uma gelada, tomar uma goleada, tomar uma pancada
VIP-3P UM TIRO	Faz-se uso do código POS "VIP-3P", o qual delimita que sua posição pode ser preenchida por qualquer unidade verbal na terceira pessoa no plural do presente do indicativo, e dos lemas "UM" e "TIRO", indicando que suas posições podem ser ocupadas por qualquer elemento associado aos seus respectivos paradigmas.	dão um tiro, levam um tiro, disparam um tiro, recebem um tiro, mandam um tiro, dão uns tiros
VERMELHO de @emoções	Utiliza-se o lema "VERMELHO", assinalando que sua posição pode ser preenchida por qualquer unidade associada a esse representante e da lista de palavras "@emoções", propiciando uma busca na qual sua posição pode ser ocupada por qualquer item pertencente a essa lista.	vermelho de raiva, vermelho de vergonha, vermelha de vergonha, vermelhas de vergonha
TER_vis-3s UM NOUN	Empregam-se: o lema "TER", para indicar o paradigma verbal em foco, delimitando a flexão verbal de interesse a partir do código _pos "_vis-3s" para indicar que se deseja observar dados em que o verbo ter encontre-se na terceira pessoa do singular do pretérito do indicativo; o lema "UM", assinalando que sua posição pode ser preenchida por qualquer elemento associado a essa unidade representante e o código POS "NOUN", o qual sinaliza que essa posição deve ser preenchida por qualquer item etiquetado como substantivo.	teve um papel, teve um aumento, teve uma queda, teve um problema, teve um gol

Fonte: Autoral.

Observamos, logo, que combinações são possíveis, mas também devem ser coerentes. Se procuramos, a título de exemplificação, por expressões compostas pelo verbo "levar" e elementos nominais terminados em "-ada", caso consideremos a presença de um artigo indefinido na configuração da expressão, este deve concordar com o gênero do item nominal, sendo empregado no feminino. Dessa forma, ao compor um comando de busca segundo as ferramentas de pesquisa disponibilizadas, devemos avaliar quais são as melhores opções para atingir nossa meta de pesquisa e refletir sobre os benefícios e possíveis adversidades que podem se apresentar devido seu uso.

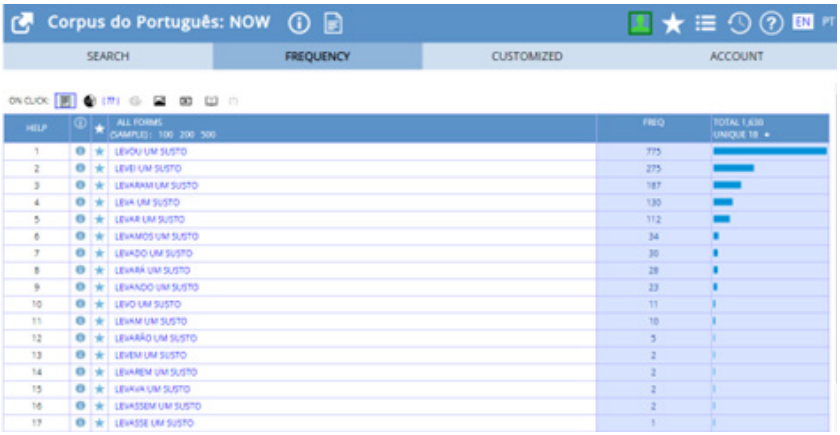
3.4.1.3 As janelas de resultados e acesso aos dados e suas fontes

No que concerne à exposição dos resultados de buscas e ao acesso aos dados, é importante ressaltar a organização das janelas de frequência, contexto e contexto expandido. Na janela de frequência, temos a apresentação inicial dos resultados de uma forma compacta, sua distribuição a depender da opção de busca selecionada – *list*, *chart*, *collocates*, *compare* ou *KWIC*.

3.4.1.3.1 Distribuição dos resultados na janela de frequência

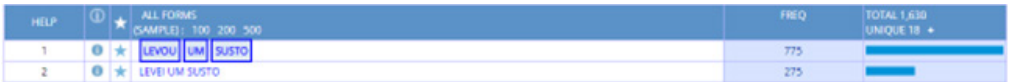
Segundo as opções *list* e *collocates*, como previamente indicado, os resultados são apresentados na janela de frequência na forma de uma lista disposta por meio de uma matriz, na qual cada expressão exposta como resultado funciona como um botão que possibilita o acesso aos dados referentes ao seu uso.

Figura 24: Exemplo de resultado de busca segundo função *list* na janela de frequência.



Na primeira coluna, temos o número da expressão de acordo com o seu posicionamento dentre os resultados. Na segunda, vemos o botão que possibilita angariar mais informações sobre o(s) elemento(s) que compõe(m) o resultado.

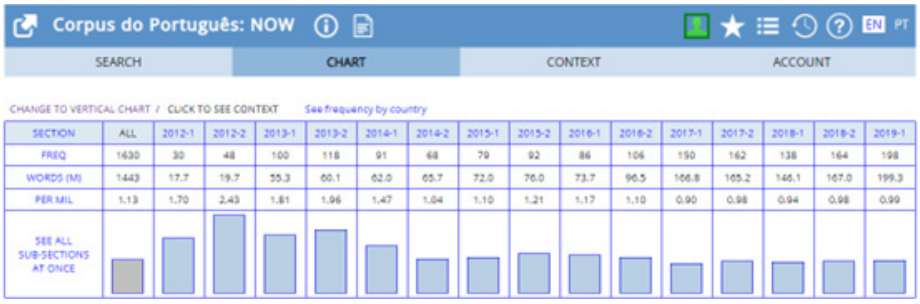
Figura 25: Exemplo de uso de botão referente a informações dos lexemas que compõem os resultados na janela de frequência.



Já na terceira coluna, temos a opção de salvar[3] a expressão disposta como resultado, a fim de visualizá-la ou utilizá-la em um momento posterior. Na quarta e quinta coluna, temos respectivamente, a apresentação dos resultados e sua frequência bruta de ocorrências no *corpus*.

A opção *chart*, por sua vez, apresenta os resultados em forma de gráfico (vertical ou horizontal). As expressões resultantes são expressas segundo segmentações pautadas nas diferentes seções disponíveis nos *corpora*, por data de ocorrência ou por país, no caso da aba *NOW*.

Figura 26: Exemplo de resultados segundo a função *chart* na janela de frequência.



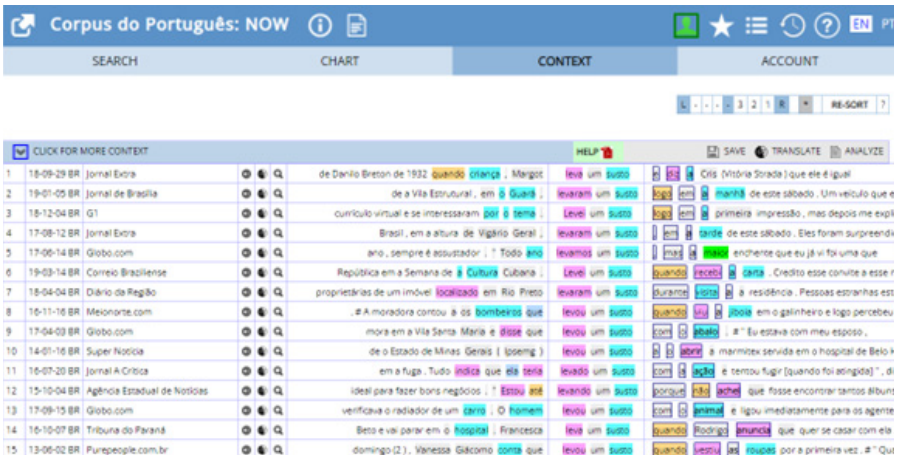
Em *compare*, comparamos a frequência de uso de um item em relação à sua associação com outros dois elementos. Logo, os resultados são expostos em duas matrizes distintas, uma levando em consideração a frequência de ocorrência do item colocado em correlação à expressão 1, e outra considerando sua correlação com a expressão 2, conforme a Figura 27.

Figura 27: Exemplo de resultados segundo a função *compare* na janela de frequência.

A busca segundo a opção *KWIC* não propicia o acesso à janela de frequência, mas sim à visualização do arranjo de elementos no entorno da expressão pesquisada na janela de contexto. Assim, cada elemento cotextual recebe uma demarcação de cor distinta, a fim de possibilitar uma melhor visualização de padrões por parte do usuário.

3 As palavras/expressões salvas são acessáveis segundo o perfil do usuário, em "Saved words / phrases".

Figura 28: Exemplo de resultados segundo a função *KWIC*.



3.4.1.3.2 Distribuição dos resultados na janela de contexto

Na janela de contexto, temos acesso aos dados de uso dos resultados observáveis na aba de frequência. Independente da função empregada no processo de busca – *list*, *chart*, *collocates*, *compare* ou *KWIC* –, a janela de contexto apresenta-se uniforme para todos os resultados. É composta por uma tabela na qual temos acesso às seguintes informações:

1ª coluna: número da expressão de acordo com o seu posicionamento dentre os resultados;

2ª coluna: data de publicação do texto fonte do dado em jogo e o país de referência do mesmo;

3ª coluna: botão para ouvir a leitura automática do dado no tradutor do Google;

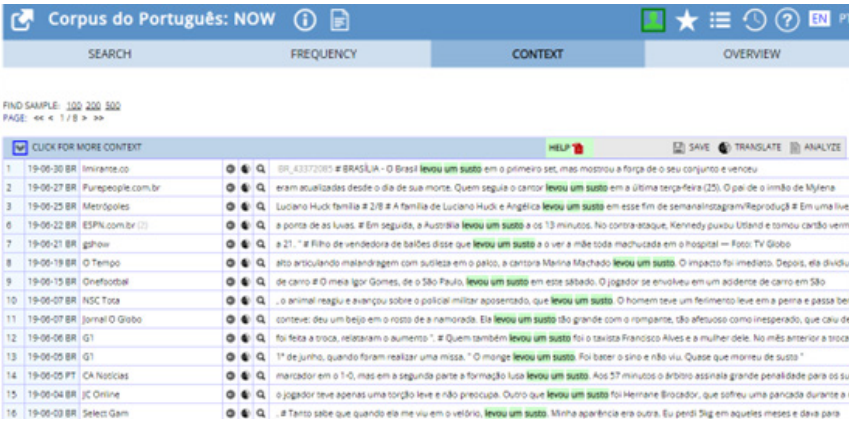
4ª coluna: botão para acessar a tradução do dado no tradutor do Google;

5ª coluna: botão que possibilita a transformação dos vocábulos do dado em *links* para pesquisas no interior do *corpus*; e

6ª coluna: linha de segmento textual no qual o dado do resultado encontra-se em destaque.

A seguir, podemos visualizar a distribuição dos resultados na janela *context*.

Figura 29: Exemplo de resultados na janela *context*.

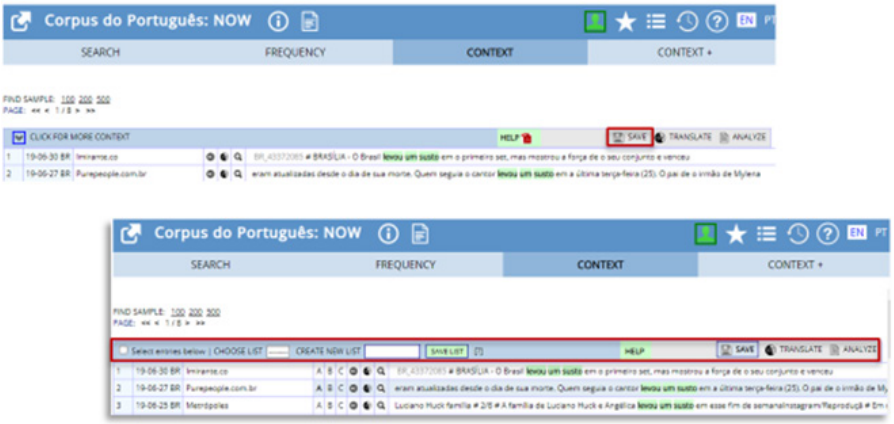


Ao clicarmos na segunda coluna, somos redirecionados para uma janela de contexto expandido, *Context +*, na qual observamos a presença de um trecho expandido do dado selecionado, assim como com informações relativas ao *link* de acesso do texto fonte do dado, e a data de publicação e título do mesmo.

Figura 30: Exemplo de resultados na janela *Context +*.

Já ao clicarmos na terceira coluna da janela de contexto, somos redirecionados para a página eletrônica referente à fonte do dado em questão. Além disso, a janela *context* oferece como recurso a possibilidade de salvar dados selecionados em uma lista (já existente ou nova) que ficará disponível no perfil do usuário para acesso em momento posterior.

Figura 31: Exemplo de uso da função *SAVE* na janela *Context*.



Assim, por meio de pesquisas na plataforma, podemos, a partir dessa ferramenta, criar *sub-corpora* pautados nas necessidades e objetivos dos usuários.

3.4.2 O PROCESSO DE COLETA NO *CORPUS* DO PORTUGUÊS

A fim de ilustrar o processo de coleta relativo a bancos de dados digitais, em especial na plataforma do *Corpus* do Português, selecionamos tratar da angariação de dados relativos a construções de predicação de passividade compostas por verbo (semi-)suporte no Português Brasileiro (PB), a partir de Teixeira (2020). Por meio de descrição metodológica apresentada pela autora, lidaremos com questões concernentes aos cuidados e passos a serem levados em consideração na coleta de dados em plataformas digitais, assim como apresentar reflexões acerca dos aspectos positivos e negativos que permeiam o processo.

Para ratificar a presença de perífrases compostas por verbo (semi-) suporte na rede de predicadores complexos de passividade do PB, e mapear os atributos (formais e funcionais) daqueles, o estudo desenvolvido por Teixeira (2020) abarca o uso de padrões constituídos pelos verbos *levar*, *tomar*, *sofrer*, *receber* e *ganhar* –, como: *levar um soco*, *tomar uma pancada*, *receber um fora*, *sofrer um gol*, *ganhar um tiro*.

Ao elaborar uma análise de cunho sincrônico, a autora recorre a aba NOW, a qual conta em sua base de dados, artigos de revistas, jornais e blogs associados ao Google News cuja publicação se deu a partir do ano de 2012. Em um primeiro, em uma busca exploratória, com o intuito de apreender distinções de colocações relativas ao uso dos elementos verbais em foco, deu-se o uso da opção *compare* do menu de busca. Foram observados os itens aos quais se compatibilizam os verbos (semi-)suporte a serem considerados na análise a fim de visualizar seus sentidos em uso concreto e estabele-

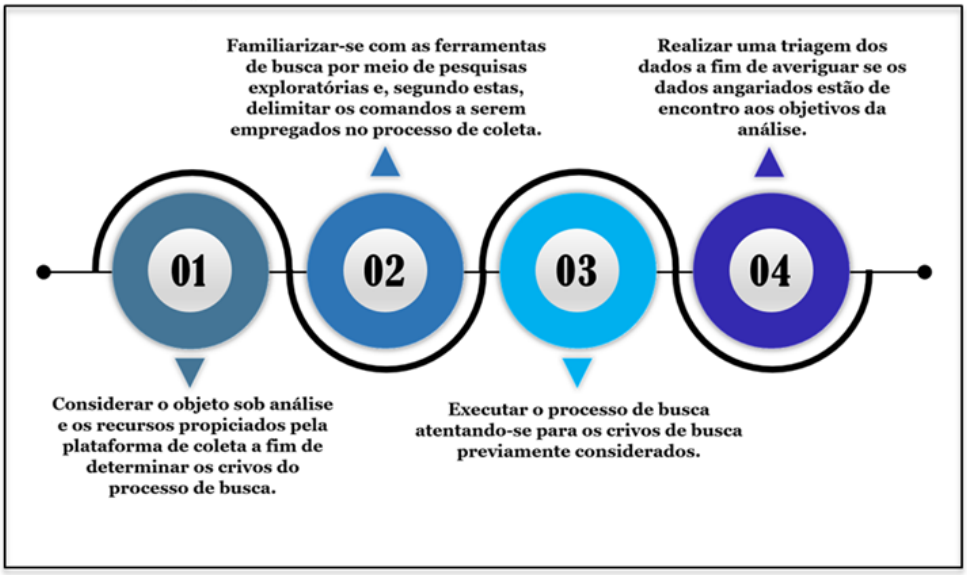
cer uma melhor compreensão dos tipos de dados aos quais se depararia no processo efetivo de coleta.

Após considerar as características configuracionais dos predicadores complexos, deu-se o seguinte padrão de busca: (i) no campo *Word/phrase* acrescentou os verbos em caixa alta para, assim, dar conta de todas as suas possibilidades de expressão; (ii) em *collocate* – uma das ferramentas que a interface de coleta do banco de dados disponibiliza – empregou os elementos nominais e/ou sintagmáticos que se associam ao elemento verbal; e (iii) definiu-se, como padrão de rastreio para o elemento nominal e/ou sintagmático, sua ocorrência até a quarta posição, levando em consideração o número de palavras que antecedem ou precedem o verbo.

Com o intuito de contemplar as possíveis realizações, fez-se uso, como chave de busca, de escolhas mais específicas e mais vastas do padrão construcional em análise, por exemplo, [LEVAR uma/um + NOUN], que viabiliza observar qual(is) elemento(s) nominal(is) ocorrem em adjacência ao verbo (*LEVAR uma + pancada; LEVAR uma + surra*). Tal procedimento foi repetido com as outras formas verbais em jogo (*tomar*, *sofrer*, *receber*, e *ganhar*). Nessa investigação, vinculada à variedade do Português do Brasil, como parâmetro, leva em consideração somente os textos publicados em sites brasileiros, com isso, em sessões, delineamos o campo de rastreio para Brasil. Finalizado o processo de busca no *corpus*, realizou-se uma triagem dos dados, a fim de verificar se os fragmentos de uso angariados eram relevantes para a investigação.

Dessa forma, para o processo de coleta em banco de dados digital, destacamos quatro passos elementares, conforme o esquema a seguir:

Esquema 1: Passos relativos ao processo de coleta em bancos de dados digitais.



Fonte: Autoral.

Com o intuito de desenvolver um trabalho de coleta coerente e otimizá-lo, o pesquisador deve, em primeiro plano, considerar seu foco de estudo, assim como os recursos que a plataforma/interface de busca pode lhe oferecer. Entretanto, uma compatibilização entre teoria e prática deve ser realizada de modo a evitar, e considerar, possíveis problemas durante o processo de coleta. Logo, é essencial executar pesquisas de caráter exploratório na plataforma para familiarizar-se com as suas ferramentas e suas implicações que concerne ao acesso aos dados. Em seguida, dá-se, em efetivo, o processo de coleta no qual o estudo se pautará e, consecutivamente, a triagem dos dados angariados em tal processo, pois, nem sempre os dados aos que temos acesso são relevantes/adequados para nossos objetivos particulares de análise.

3.5 CONSIDERAÇÕES FINAIS

Buscamos, neste capítulo, introduzir, ao (futuro)pesquisador, o universo de coleta online, em especial a busca de dados por meio do gerenciador de *corpora Corpus* do Português. Com isso, ressaltamos a relevância do uso de *corpora* virtuais para o campo de análise linguística, uma vez que, contribuem para a otimização do processo de coleta de dados. Ressaltamos que, ao recorrer a *corpora* online, o pesquisador encontra: fácil acessibilidade, rápido rastreamento de dados de diversas línguas e variedades, mais agilidade e menos esforço para a obtenção de dados linguísticos, fácil seleção das fontes e/ou gêneros textuais e acesso a informações sobre os dados. Em contrapartida, depara-se também com: páginas indisponíveis e páginas bloqueadas devido à política de privacidade do conteúdo. Por fim, destacamos a importância de delimitar as diretrizes de coleta segundo os recursos dos *corpora* disponíveis, bem como adaptá-las aos objetivos da análise linguística.

REFERÊNCIAS

DAVIES, M. O corpus do português. *Corpus do Português*. Disponível em: https://www.corpusdoportugues.org/x.asp, 2016. Acesso em: 15 mar. 2022.

SARDINHA, T. B. Linguística de corpus: histórico e problemática. *In.: Delta:* documentação de estudos em linguística teórica e aplicada, São Paulo, v. 16, n. 2, 2000.

TEIXEIRA, R. B. de S. *Estruturas com verbo (semi)suporte*: a variação sob um prisma construcionista. Dissertação de Mestrado. Universidade Federal do Rio de Janeiro, Faculdade de Letras, 2020.