Como elaborar um Dicionário Especializado?

A experiência do Grupo TermiSul

Organização

Cleci Regina Bevilacqua
Denise Regina de Sales
Márcia Moura da Silva
Patrícia Chittoni Ramos Reuillard
Sandra Dias Loguercio

editora ZOUK

Como elaborar um Dicionário Especializado?

Porto Alegre • 2023 • 1ª edição

Organização
Cleci Regina Bevilacqua
Denise Regina de Sales
Márcia Moura da Silva
Patrícia Chittoni Ramos Reuillard
Sandra Dias Loguercio

editora
ZOUK

Projeto gráfico e edição: Editora Zouk
Revisão: Cristiane Krause Kilian
Revisão técnica: Silvana de Fátima Bojanoski
Design da capa: Mateus Moura Godinho

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Elaborado por Odilio Hilario Moreira Junior - CRB-8/9949

C735

Como elaborar um dicionário especializado? [recurso eletrônico] / organizado por Cleci Regina Bevilacqua, Denise Regina de Sales, Márcia Moura da Silva, Patrícia Chittoni Ramos Reuillard e Sandra Dias Loguercio - Porto Alegre, RS : Zouk, 2023.

137 p. ; ePUB.

Inclui bibliografia.
ISBN: 978-65-5778-119-7 (Ebook)

1. Dicionário. I. Bevilacqua, Cleci Regina. II. Sales, Denise Regina de. III. Silva, Márcia Moura da. IV. Título.

CDD 403
2023-???? CDU 403

Editora Zouk
Av. Cristóvão Colombo, 1343 sl. 203
90560-004 – Floresta – Porto Alegre – RS – Brasil
f. 51. 3024.7554

www.editorazouk.com.br

Capítulo 3 – Constituição de *corpora*: critérios de coleta, limpeza e organização

Márcia Moura da Silva
Manuela Arcos Machado

Para se construir um *corpus,* é preciso seguir uma série de etapas e fazer uso de programas especiais para processá-lo. Neste capítulo, partindo de uma breve reflexão sobre o uso de *corpus* na pesquisa terminológica e na elaboração de material terminográfico, descrevemos os critérios de coleta, limpeza e organização de um *corpus* e apresentamos dois programas, *AntConc* e *Sketch Engine*, utilizados para processar os textos que compõem a base de dados do projeto do Grupo Termisul na área de *Conservação e Restauração de Bens Culturais Móveis em papel (Projeto Papel)*, realizado entre 2019 e 2021. Em uma primeira etapa (2016 a 2019), foram coletados os termos da área e, em uma segunda etapa (2019 a 2021), as UFEs. A elaboração dessa base contou com textos da área em português, espanhol, francês, inglês, italiano e russo. Alguns dos exemplos que trazemos para apoiar nosso texto foram extraídos dessa base.

Uso de *corpora* na elaboração de material terminográfico

Ao longo deste Manual, falaremos em *corpora* de estudo e de referência. *Corpus* é uma coletânea de textos em formato eletrônico, compilado segundo critérios específicos de acordo com o estudo que se pretende realizar. *Corpus* de estudo é o *corpus* no qual se baseia a pesquisa a ser desenvolvida pelo pesquisador. *Corpus* de referência é o que serve de comparação para o *corpus* de estudo e normalmente deve ter de três a cinco vezes o seu tamanho (Tagnin, 2011). Para as pesquisas terminológicas, em geral, o *corpus* de referência deve estar formado por gêneros textuais que representem a língua comum.

Como visto no Capítulo 2, o pesquisador pode trabalhar com dois tipos de *corpus* – o comparável e o paralelo[1]. Ainda que existam *corpora* gigantescos, como é o caso do *Corpus of Contemporary American English* (Coca), que é o maior *corpus* de livre acesso do mundo (mais de um bilhão de palavras), o mais importante é construir um *corpus* que siga critérios bem definidos e que dê conta dos objetivos da pesquisa. Aluísio e Almeida (2006, p. 158-159) apresentam uma síntese dos principais critérios defendidos por diferentes teóricos da Linguística de *Corpus*: i) autenticidade (textos escritos em linguagem natural e também por falantes nativos); ii) representatividade (textos representativos da língua ou de uma variedade de língua que se queira investigar para que o *corpus* possa representar seu uso efetivo); iii) balanceamento (equilíbrio de gêneros discursivos, tipos de textos[2], títulos ou autores, desde que sejam adequados à investigação pretendida e que a escolha tenha sido feita de maneira criteriosa); iv) amostragem (amostras que incluam toda a variação linguística existente); v) diversidade (de gêneros, tipos de textos e sobretudo de tópicos, visto que a variação desses últimos afeta a frequência de muitas palavras); e vi) tamanho (adequado ao tipo de pesquisa e à metodologia adotada).

Entre as várias áreas do conhecimento em que o uso de *corpora* vem se consolidando, está a Terminologia. Maciel (2006, p. 1) aponta que a disciplina acompanhou outros ramos da Linguística ao adotar o uso dessa metodologia em suas pesquisas. Para a autora, a pesquisa terminológica baseada em *corpus* valoriza o contexto sociolinguístico do termo e consequentemente do texto enquanto "registro do evento comunicativo real". Como observam Bojanoski, Michelon e Bevilacqua (2017), o texto se tornou, a partir dos anos 1990, o objeto central de análise da Terminologia. Pela posição que o texto ocupa hoje nas teorias de Terminologia, as autoras defendem que a elaboração de um *corpus* seguindo critérios bem definidos,

1 No *site* do projeto **Legis** do Grupo Termisul, é possível acessar a base de textos legislativos usados em *corpora* comparáveis, assim como alguns textos alinhados (em língua inglesa e suas traduções para a língua portuguesa, alemã, espanhola e francesa). Disponível em: http://www.ufrgs.br/termisul/legis.php (download > legislação ambiental; download > textos alinhados).

2 Ver Marcuschi (2005, 2008) para uma discussão aprofundada sobre gêneros e tipos textuais.

sobretudo no que diz respeito à representatividade, é fundamental ao trabalho terminológico.

Ainda que a Terminologia venha abraçando essa metodologia, Maciel (2006) adverte que conduzir uma pesquisa baseada em *corpus* não significa necessariamente seguir todos os princípios da Linguística de *Corpus*[3]. Segundo ela, uma das principais vantagens de se construir um *corpus* para um projeto terminológico é a possibilidade de se conduzir uma investigação empírica dos termos ou unidades fraseológicas em uma quantidade considerável de textos especializados. Mas, ainda que seja uma rica fonte de dados e que esses dados evidenciem fatos sobre o padrão de uso que poderiam permanecer imperceptíveis em amostragens menores, o *corpus* não é um "manancial exaustivamente completo" (Maciel, 2006, p. 5).

A autora compartilha com Leech (1991), um dos pioneiros no uso de *corpus* eletrônico do inglês britânico, a noção de que o linguista faz uso da intuição em sua pesquisa, mas vai além e acrescenta à intuição a habilidade de interpretação e o conhecimento do sistema da língua em estudo, seja como falante nativo, falante proficiente ou como linguista. De fato, esses elementos adicionais podem até mesmo indicar ao pesquisador quando é necessário sair dos limites de um *corpus*. No caso do projeto na área de *Conservação e Restauração de Bens Culturais Móveis em papel (Projeto Papel)*, por exemplo, nem sempre foi possível encontrar equivalentes de uma determinada UFE em todas as línguas em seus respectivos *corpora*. Assim, foi necessário interpretar essa ausência de equivalentes para podermos validá-la ou procurar equivalentes fora do *corpus*, em *sites* especializados, como pode ser visto no capítulo 6.

Critérios de coleta, limpeza e organização

Como já mencionado, a construção de um *corpus* demanda critérios claros para sua adequação. A partir do momento que se definem o recorte da pesquisa (temática, tamanho do *corpus*, composição e número mínimo

3 Para saber mais sobre a Linguística de *Corpus* e seus avanços, ver McEnery e Hardie (2012); Tagnin (2011); Berber-Sardinha (2000, 2004); Baker (1995); Biber (1993); e Leech (1991).

de ocorrência do objeto a ser investigado) e o público-alvo do produto terminográfico[4], o primeiro passo será identificar a produção de textos na área do estudo. Nesse sentido, seguindo a proposta de Aluísio e Almeida (2006), sugerimos que se procure equilibrar os gêneros discursivos e tipos de textos. Na construção do *corpus* do nosso projeto na área de *Conservação e Restauração de Bens Culturais Móveis (Projeto Papel)*, por exemplo, incluíram-se textos do gênero acadêmico, como livros, manuais, periódicos, trabalhos de conclusão de curso, dissertações, teses, anais de eventos, relatórios e boletins informativos de associações da área. Vale mencionar que, visto estar a área em situação de estruturação no Brasil, encontrou-se um número reduzido de material em língua portuguesa, tendo-se incluído material de áreas afins, como Biblioteconomia, Arquivologia e Museologia. (Bojanoski; Michelon; Bevilacqua, 2017). A seleção dos textos se deu pela busca das palavras-chave *documento*, *documentação*, *conservação*, *papel*, *patrimônio*, *preservação*, *restauração* e *restauro*, tendo-se tido o cuidado de escolher material que proviesse de fontes confiáveis. Nesse caso, textos acadêmicos, *sites* de universidades, instituições de pesquisa e periódicos reconhecidos na área foram considerados fontes possíveis de coleta desse material.

Uma vez que se tenham selecionado os textos que comporão o *corpus*, é preciso que eles sejam salvos e catalogados. Para que possam ser processados pelas ferramentas de maneira adequada, os textos devem ser salvos em formato <.txt>. Entretanto, diferentes programas pedem diferentes codificações. O *AntConc*[5] por exemplo, requer que os textos sejam salvos em *UTF-8*, mas caso se esteja trabalhando com o programa *ParaConc*, eles devem ser salvos utilizando a codificação *ANSI*[6]. Cada texto é salvo com

4 Ver definição do usuário de material terminográfico no Capítulo 2.

5 Durante a pesquisa, a versão disponível do software era 3.2. Atualmente, a última versão (AntConc 4.2) já oferece a possibilidade de leitura de textos em formato .pdf.

6 Há programas específicos disponíveis gratuitamente na *web* para a conversão de textos em <.pdf> para o formato em <.txt> ou ANSI. Um exemplo é o *AntFileConverter*, do mesmo desenvolvedor do *AntConc*, disponível gratui-

* Sugerimos que esse código apresente as duas primeiras letras do idioma dos textos, duas ou três letras que identifiquem o projeto, seguidas da numeração sequencial. No nosso projeto, por exemplo, o código adotado foi ptPPx, para o qual pt correspondia à língua portuguesa, PP a Projeto Papel e x à numeração do texto. Da mesma forma para os *corpora* de outros idiomas, como espPPx (espanhol), frPPx (francês), ingPPx (inglês) etc.

* Outra alternativa para não recuperar ruídos – isto é, palavras ou estruturas que não são relevantes para a pesquisa, mas que são acusadas pela ferramenta como possíveis palavras-chave, apesar de não as serem – durante a extração de palavras-chave é o uso de *stoplists*. Essas listas, geralmente, são de palavras gramaticais, nomes próprios e de lugares. Como resultado, a ferramenta extrai palavras lexicais. Isto é, aquelas que são representativas da temática do *corpus* de estudo e que poderão ser candidatas a termos. Dependendo do objetivo da pesquisa, outras palavras podem ser acrescentadas às *stoplists*.

um **código** que acaba por se tornar a "identidade" desse texto. Feito isso, todos os textos coletados devem passar por um processo de limpeza em que são retirados todos os **elementos extratextuais***, como tabelas, gráficos e imagens, assim como qualquer informação que não seja significativa para o estudo, como agradecimentos, sumários, referências bibliográficas, notas de rodapé e *links* externos. No caso do *corpus* em língua portuguesa de nosso projeto, 161 textos foram selecionados, totalizando 38.129 *types* (número de palavras diferentes) e cerca de 967.852 *tokens* (número total de palavras).

Em relação à etapa de catalogação[7], ela é necessária para que se tenha um registro do material utilizado para a construção do *corpus*. Fica a critério do pesquisador se o público externo terá acesso aos catálogos ou se esses serão acessados somente pela equipe de pesquisa. O importante é que esse instrumento registre informações como autor, título, fonte, ano de publicação, gênero textual, código de identificação do texto e, se for o caso, o *link* onde o texto esteja disponível e a data de coleta. Essa catalogação pode ser feita em uma tabela em editor de textos (Word, Documentos Google ou outro) ou em uma planilha em <.xlsx> (Excel ou Planilhas Google). Um exemplo de catálogo do Projeto Papel é apresentada no quadro 3.1.

tamente no *site*: https://www.laurenceanthony.net/software/antfileconverter/.

7 Outra alternativa para controlar o registro dos textos que compõem o *corpus* de estudo é a criação de cabeçalhos (*labels*) com informações do texto. Essa informação, no momento da busca, pode ser excluída com os recursos das ferramentas.

Quadro 3.1 – informações para catalogação do corpus textual

Código	Referências	Disponível em	Acesso em
ptPP001	ROSAS, Fernanda Jenner; MENDES, Débora Assis. Resgatando Olinda.	Seminário da ABRACOR, 4, 1988, Gramado. Anais – v. 1. [S.l]: [ABRACOR], 1988. p. 115-124. [comunicação evento]	
ptPP099	CORADI, Joana Paula; EGGERT-STEINDEL, Gisela. Técnicas básicas de conservação e preservação de acervos bibliográficos. Revista ACB, São José, V. 13, n. 2, jul-dez, 2008.	https://revista.acbsc.org.br/racb	05 jul. 2017

Fonte: As autoras.

Esses procedimentos são necessários para garantir a adequação da metodologia empregada. Uma vez concluídos, os textos estão prontos para serem processados pelos programas, que facilitam o trabalho do pesquisador, pois disponibilizam uma série de ferramentas que permitem identificar e extrair o objeto do estudo, produzir listas de palavras, buscar colocados (isto é, palavras que costumam aparecer combinadas ao termo pesquisado), verificar frequência dos termos etc., como veremos a seguir.

Programas de extração

A construção de um *corpus* dá ao pesquisador acesso, em um só instrumento, a um grande número de textos, a partir dos quais poderá verificar padrões de uso em uma determinada língua ou linguagem. Para que isso aconteça, é preciso que os textos que foram criteriosamente selecionados, limpos e catalogados sejam processados em programas especialmente criados para serem usados com *corpora*. Há no mercado uma série de

programas que fazem esse trabalho, como, entre outros, *WordSmith Tools*[8], *AntConc*[9], *Sketch Engine*[10], *Unitext*[11], *TermoStat*[12], *ParaConc*[13], *AntPConc*[14], estes dois últimos para processamento de *corpora* paralelos. Descreveremos algumas das funcionalidades do *AntConc* e do *Sketch Engine*, dois dos programas usados para processar os *corpora* do nosso projeto.

AntConc e *Sketch Engine*

O *AntConc* é um programa de livre acesso, desenvolvido pelo linguista britânico Laurence Anthony, que pode ser baixado e instalado em qualquer computador sem haver necessidade de estar conectado à rede. Por ser um programa gratuito, ele é mais limitado que o *Sketch Engine*, mas tem uma interface amigável e intuitiva[15], permitindo diferentes pesquisas por meio de suas ferramentas e índices estatísticos.

O *Sketch Engine* é um gerenciador de *corpora* e um programa de análise textual criado inicialmente em 2003 por Adam Kilgarriff e Pavel Rychly e desenvolvido pela *Lexical Computing Ltd*[16]. Por ser uma ferramenta *on-line*, ele está em constante atualização, oferecendo ao usuário *corpora* atualizados e ampliados rotineiramente. Para acessá-lo, é necessário criar uma conta através do *site* https://www.sketchengine.eu/. Um dos aspectos negativos dessa ferramenta é que ela não é totalmente gratuita, mas é possível criar uma conta temporária de 30 dias que permite acesso à maioria dos recursos. Nessa conta *trial*, é permitido carregar um *corpus* textual de

8 Disponível em: https://www.lexically.net/wordsmith/

9 Disponível em: https://www.laurenceanthony.net/software/antconc/

10 Disponível em: https://www.sketchengine.eu/

11 Disponível em: https://unitexgramlab.org/pt

12 Disponível em: http://termostat.ling.umontreal.ca/

13 Disponível em: https://paraconc.com/

14 Disponível em: https://www.laurenceanthony.net/software/antpconc/

15 Além das instruções de uso apresentadas na própria ferramenta (clicar em "help"), há vários vídeos na internet em que o próprio Laurence Anthony dá o passo a passo para a execução das principais funções do *AntConc*. Ver em: https://www.youtube.com/watch?v=9TsqFVrUYO0

16 https://www.lexicalcomputing.com/

até 1 milhão de palavras. Para analisar *corpora* maiores, é preciso, obrigatoriamente, comprar uma licença.

Apesar disso, o *Sketch Engine* conta com uma ampla lista de aspectos positivos. Destacamos, entre eles, a possibilidade de **lematização*** automática do *corpus* inserido nele (à diferença do *AntConc*) e também a oferta de *corpora* de referência em diferentes línguas, sem a necessidade de se criar ou de se ter disponível um *corpus* de referência para contrastar com o *corpus* de estudo (como ocorre com o *AntConc*).

* A lematização consiste no processo de juntar todas as formas de uma palavra em sua forma canônica. Um *corpus* lematizado permite que, por apenas uma única busca, recuperem-se todas as formas conjugadas de um verbo, por exemplo. Assim, buscando-se o verbo "registrar", a ferramenta recuperará todas as suas conjugações: registra, registrou, havia registrado, registrara, registraria, registrado, registrando etc. O mesmo aplica-se a substantivos e adjetivos com suas flexões de gênero e número.

Os *corpora* de referência de diferentes línguas ofertados pelo *Sketch Engine* contam com bilhões de palavras e são atualizados regularmente. Também são disponibilizados outros tipos de *corpora*, como alguns paralelos, com textos legislativos traduzidos de diferentes países da Europa, *corpora* especializados de diferentes temas, *corpora* diacrônicos, *corpora* de aprendizes e, inclusive, *corpora* orais.

Vale mencionar que, embora a ferramenta *AntConc* não execute uma lematização automática do *corpus* textual, há disponíveis extensões para o *software* que o fazem. Ainda, existe a possibilidade de lematizar manualmente o *corpus* em .txt. Essa tarefa, no entanto, exige um trabalho longo de codificação. Por fim, outra alternativa no *AntConc* é a possibilidade de realizar buscas com a forma truncada da palavra, utilizando um asterisco. Por exemplo, para identificar todas as ocorrências do verbo *registrar*, pode-se buscar *registr**, o que permitirá chegar às diferentes formas do verbo, como *registra, registrou, registrará* etc.[17]

Além das ferramentas que apresentamos, o *Sketch Engine* conta também com outras de diferentes funções, seja para tarefas de extração terminológica, de análises linguísticas para fins específicos (como ensino de línguas) e também para soluções tradutórias.

17 Para saber mais sobre buscas de estruturas frasais complexas usando formas truncadas no *AntConc*, ver Arcos e Bevilacqua (2018).

Destacamos abaixo a função de algumas das principais ferramentas desses dois programas. Ambos oferecem ferramentas de busca por palavra (*concordance*); lista de palavras (*wordlist*); busca por expressões complexas (n-gramas); lista de palavras-chave (*keyword list*); colocados (*collocate* no *AntConc; Word Sketch* no *Sketch Engine*), porém, o *Sketch Engine* apresenta mais recursos, como é o caso do *Word Sketch difference* e o *parallel concordance*, possibilitando, assim, uma investigação mais minuciosa do *corpus*.

Concordance: permite que o pesquisador procure qualquer palavra no *corpus*. No *AntConc*, quando a palavra é digitada no campo de pesquisa, a ferramenta traz todas as ocorrências dessa palavra, que aparece em cor diferente para identificação imediata, juntamente com o contexto em que se insere (basta clicar em qualquer ocorrência da palavra para que o contexto seja ampliado). A figura 3.1 mostra o resultado da busca pelo termo *acervo* (em azul) com as três palavras à direita em cores diferentes (o número de palavras que acompanham o termo consultado pode ser ajustado, e elas podem também ser destacadas à esquerda do termo). A busca foi feita no *corpus* Papel, ou seja, o *corpus* do projeto.

Figura 3.1 – Concordâncias do termo *acervo* (*Corpus* Papel)

KWIC | Plot | File | Cluster | N-Gram | Collocate | Word | Keyword

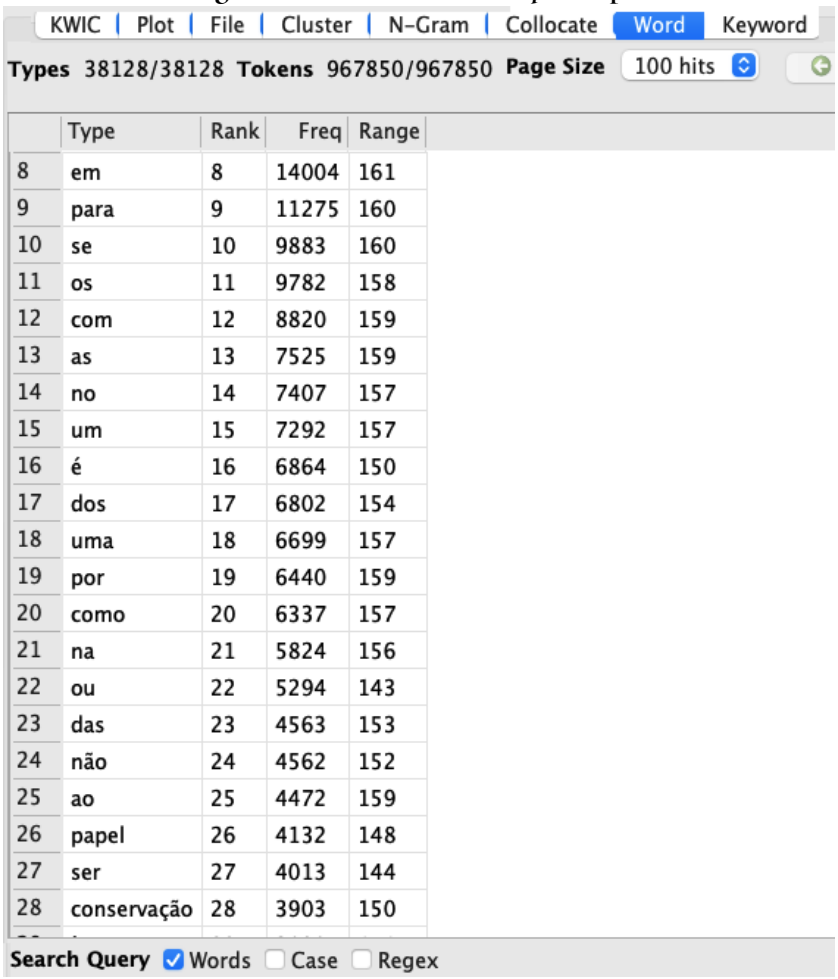
Total Hits: 2422 Page Size 50 hits 701 to 750 of 2422 hits

	File	Left Context	Hit	Right Context
12	ptPP127.txt	nto e melhoria da qualidade nos serviços de preservação do	acervo	documental o Núcleo de Documentação da UFF. In: Anais...
13	ptPP157.txt	incipais fatores ambientais que prejudicam e deterioram um	acervo	documental ou bibliográfico. Assim como os danos relaciona
14	ptPP090.txt	estabelecidas com universidades e instituições de guarda de	acervo	documental para viabilizar esta metodologia. Em 2009 o Dep
15	ptPP095.txt	vulgarmente como brocas ou caruncho. Na deterioração de	acervo	documental por insetos, outras famílias também causam dan
16	ptPP111.txt	arquivo refere-se tanto ao local de guarda de um	acervo	documental quanto ao próprio acervo em si. As duas
17	ptPP097.txt	auro. É parte das atribuições destas instituições preservar o	acervo	documental sob sua guarda, ao mesmo tempo em que
18	ptPP142.txt	003, p. 20-21). Sob esta perspectiva, o termo documento ou	acervo	documental tem em seu significado uma estreita relação com
19	ptPP159.txt	al da UFSM. O desenvolvimento das ações de preservação do	acervo	documental vão ao encontro de um dos objetivos do
20	ptPP153.txt	de transporte, embalagem e seguro. 9.10 Toda unidade do	acervo	a ser emprestada deverá ser conferida na sua saída
21	ptPP153.txt	de transporte, embalagem e seguro. 9.10 Toda unidade do	acervo	a ser emprestada deverá ser conferida na sua saída
22	ptPP111.txt	udiciais a profissionais e pesquisadores que irão lidar com o	acervo	a ser tratado, além de não ter apresentado alterações
23	ptPP142.txt	proposta reduziria, em alguma percentagem, o universo do	acervo	a ser tratado, ao mesmo tempo em que o
24	ptPP112.txt	seja uma inundação ou qualquer outro motivo que leve o	acervo	a ser atingido por água, os primeiros procedimentos devem
25	ptPP148.txt	deve estar adaptado para atender as especificidades de cada	acervo	a ser transportado e a sua localização. Usualmente o

Fonte: AntConc.

Wordlist: a ferramenta gera uma lista de todas as palavras do *corpus*, em ordem alfabética ou por frequência. Conforme a figura 3.2, nota-se, na lista de palavras gerada pelo *AntConc*, que nas primeiras posições aparecem artigos, preposições e conjunções que não são relevantes à pesquisa. O programa permite o uso de *stoplists* para que essas palavras sejam ignoradas na busca e sejam visualizadas apenas as palavras representativas do *corpus* (figura 3.3).

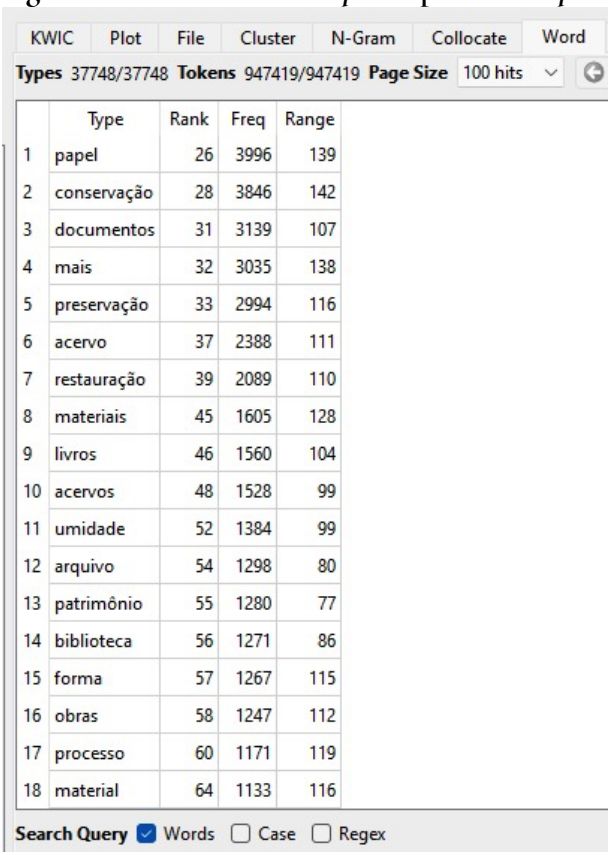
Figura 3.2 – *Wordlist* do *Corpus* Papel



	Type	Rank	Freq	Range
8	em	8	14004	161
9	para	9	11275	160
10	se	10	9883	160
11	os	11	9782	158
12	com	12	8820	159
13	as	13	7525	159
14	no	14	7407	157
15	um	15	7292	157
16	é	16	6864	150
17	dos	17	6802	154
18	uma	18	6699	157
19	por	19	6440	159
20	como	20	6337	157
21	na	21	5824	156
22	ou	22	5294	143
23	das	23	4563	153
24	não	24	4562	152
25	ao	25	4472	159
26	papel	26	4132	148
27	ser	27	4013	144
28	conservação	28	3903	150

Fonte: AntConc.

Figura 3.3 – *Wordlist* do *Corpus* Papel com *stoplist*



	Type	Rank	Freq	Range
1	papel	26	3996	139
2	conservação	28	3846	142
3	documentos	31	3139	107
4	mais	32	3035	138
5	preservação	33	2994	116
6	acervo	37	2388	111
7	restauração	39	2089	110
8	materiais	45	1605	128
9	livros	46	1560	104
10	acervos	48	1528	99
11	umidade	52	1384	99
12	arquivo	54	1298	80
13	patrimônio	55	1280	77
14	biblioteca	56	1271	86
15	forma	57	1267	115
16	obras	58	1247	112
17	processo	60	1171	119
18	material	64	1133	116

Fonte: AntConc.

N-grams: a ferramenta extratora de n-gramas, também chamada de MWE (*multiword expressions*) ou expressões complexas, permite identificar estruturas de diferentes tamanhos (por exemplo, de 2 a 3 palavras, de 2 a 4 palavras, de 2 a 5 palavras, de 3 a 4 palavras, de 3 a 5 palavras etc.) que apresentam uma ocorrência frequente no *corpus* de estudo. O extrator de n-gramas é bastante útil para identificar expressões complexas típicas do *corpus* de estudo e pode ser empregado como extrator de termos sintagmáticos e unidades candidatas a fraseologias especializadas. Contudo, cabe ressaltar que ele não filtrará somente expressões complexas iniciadas e terminadas por palavras lexicais (por exemplo, *estado de conservação*), mas também estruturas recorrentes iniciadas e terminadas por palavras

gramaticais (por exemplo, *de Conservação e*). A figura 3.4 ilustra n-gramas do *Corpus* Papel do tamanho de 3 a 4 palavras no *Sketch Engine*.

Figura 3.4 – n-gramas do *Corpus* Papel

3–4-grams, word (items: 14,821 , total frequency: 156,296)

	Word	Frequency
1	conservação e restauração	389
2	Rio de Janeiro	327
3	de conservação e	299
4	temperatura e umidade	286
5	a necessidade de	240
6	estado de conservação	238
7	a fim de	237
8	de temperatura e	225
9	o uso de	224
10	a partir de	216
11	e restauração de	216
12	de acordo com	213
13	por meio de	207
14	obras de arte	203
15	conservação e restauração de	197
16	o processo de	191
17	de temperatura e umidade	188
18	Conservação e Restauração	176
19	a partir da	176
20	do século XIX	174
21	e umidade relativa	174
22	de conservação e restauração	171
23	umidade relativa do	169
24	uma vez que	155
25	de Conservação e	155
26	temperatura e umidade relativa	147
27	para a preservação	145
28	relativa do ar	144
29	preservação e conservação	143
30	do patrimônio cultural	143
31	de preservação e	142
32	e Restauração de	141
33	de bens culturais	139
34	do Rio de	138

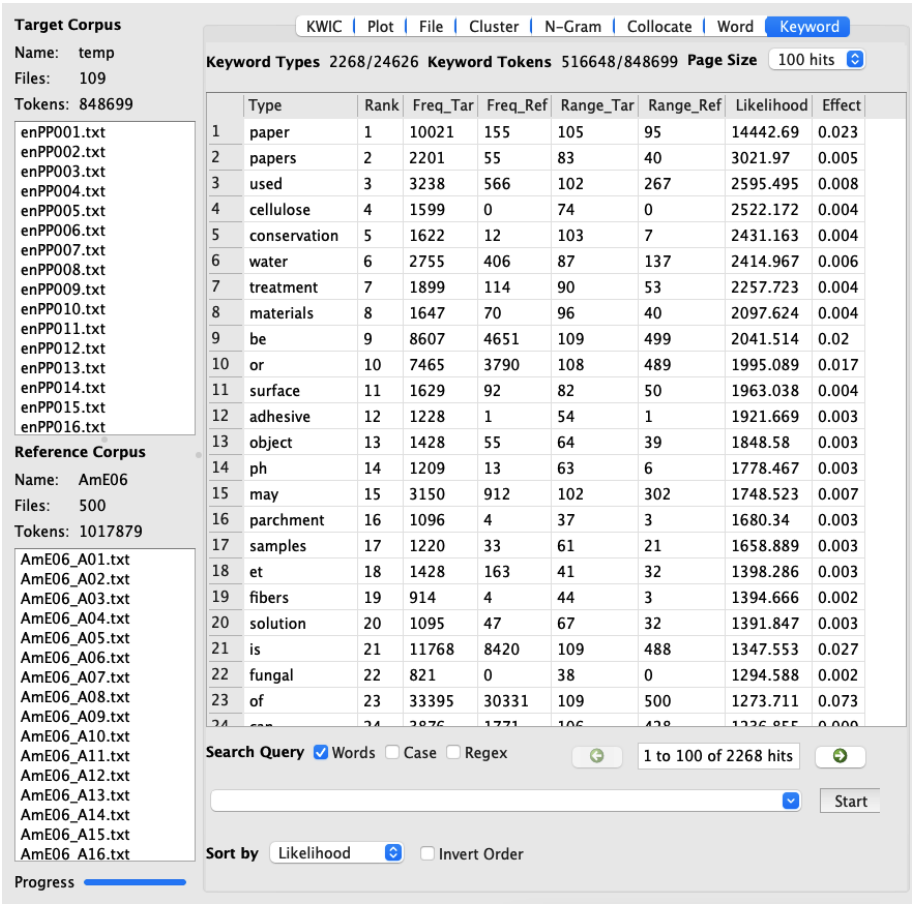
Fonte: Sketch Engine.

***Keyword list*:** essa lista de palavras-chave é resultado da comparação das palavras do *corpus* de estudo com as de um *corpus* de referência, como é o caso do *BNC* (*British National Corpus*[18]), cuja importação pode ser feita para o *AntConc*. Tal comparação permite que o pesquisador identifique palavras que são estatisticamente mais frequentes no *corpus* de estudo. Quanto

18 *Corpus* de livre acesso constituído de textos de inglês britânico falado e escrito com mais de cem milhões de palavras.

maior for a frequência estatística de uma palavra nesse *corpus*, maior será sua especificidade. Em termos simples, o cálculo feito pelo extrator consiste na comparação da proporção de ocorrências de uma palavra no *corpus* de estudo frente à proporção de ocorrências dessa mesma palavra no *corpus* de referência. No trabalho terminológico, essa ferramenta é usada para chegar a candidatos a termos. A figura 3.5 mostra uma lista de palavras-chave do *corpus* em língua inglesa do projeto Papel, em que a palavra *paper* aparece como sendo a mais frequente (10.021 ocorrências), sendo que no *corpus* de referência essa mesma palavra aparece com uma frequência de 155 ocorrências, o que indica sua alta especificidade no *corpus* de estudo.

Figura 3.5 – *Keywords* do *Corpus* Papel

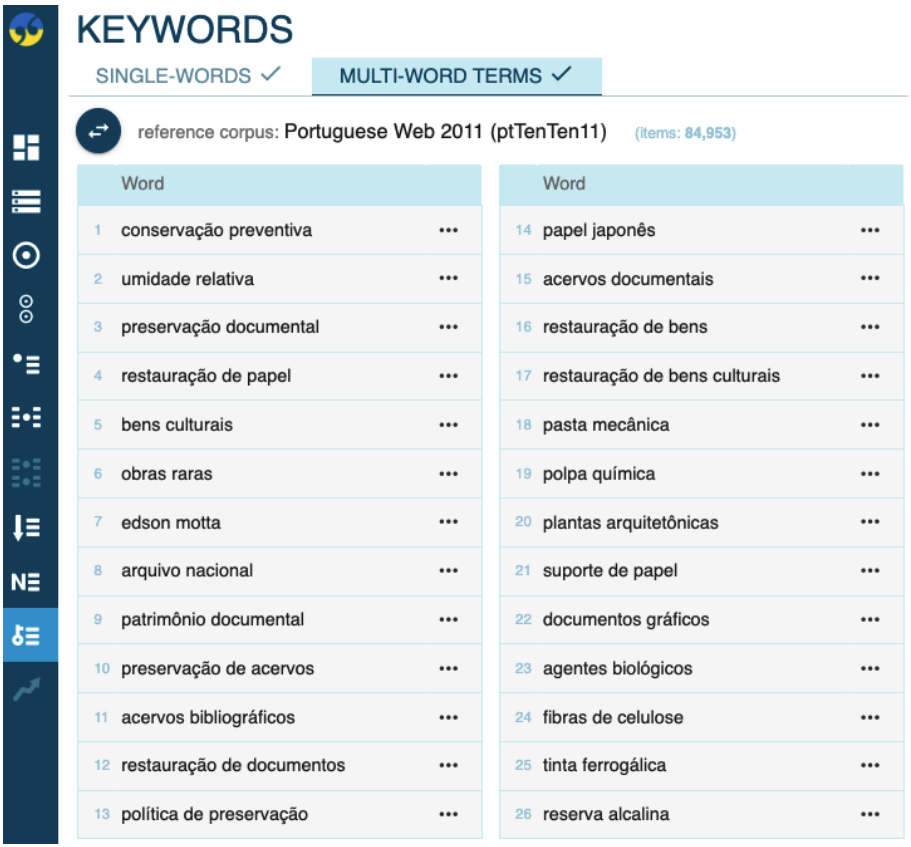


	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Likelihood	Effect
1	paper	1	10021	155	105	95	14442.69	0.023
2	papers	2	2201	55	83	40	3021.97	0.005
3	used	3	3238	566	102	267	2595.495	0.008
4	cellulose	4	1599	0	74	0	2522.172	0.004
5	conservation	5	1622	12	103	7	2431.163	0.004
6	water	6	2755	406	87	137	2414.967	0.006
7	treatment	7	1899	114	90	53	2257.723	0.004
8	materials	8	1647	70	96	40	2097.624	0.004
9	be	9	8607	4651	109	499	2041.514	0.02
10	or	10	7465	3790	108	489	1995.089	0.017
11	surface	11	1629	92	82	50	1963.038	0.004
12	adhesive	12	1228	1	54	1	1921.669	0.003
13	object	13	1428	55	64	39	1848.58	0.003
14	ph	14	1209	13	63	6	1778.467	0.003
15	may	15	3150	912	102	302	1748.523	0.007
16	parchment	16	1096	4	37	3	1680.34	0.003
17	samples	17	1220	33	61	21	1658.889	0.003
18	et	18	1428	163	41	32	1398.286	0.003
19	fibers	19	914	4	44	3	1394.666	0.002
20	solution	20	1095	47	67	32	1391.847	0.003
21	is	21	11768	8420	109	488	1347.553	0.027
22	fungal	22	821	0	38	0	1294.588	0.002
23	of	23	33395	30331	109	500	1273.711	0.073

Fonte: AntConc.

O *Sketch Engine*, como mencionamos anteriormente, já possui *corpora* de referência acoplados a ele[19], inclusive do português brasileiro; assim, não há necessidade de se buscar um *corpus* de referência externo para gerar uma lista de palavras-chave. Outra vantagem dessa ferramenta é que ela identifica não somente candidatos a termos simples, mas também sintagmáticos (ver figura 3.6).

Figura 3.6 – *Keyword*: extração de candidatos a termos sintagmáticos

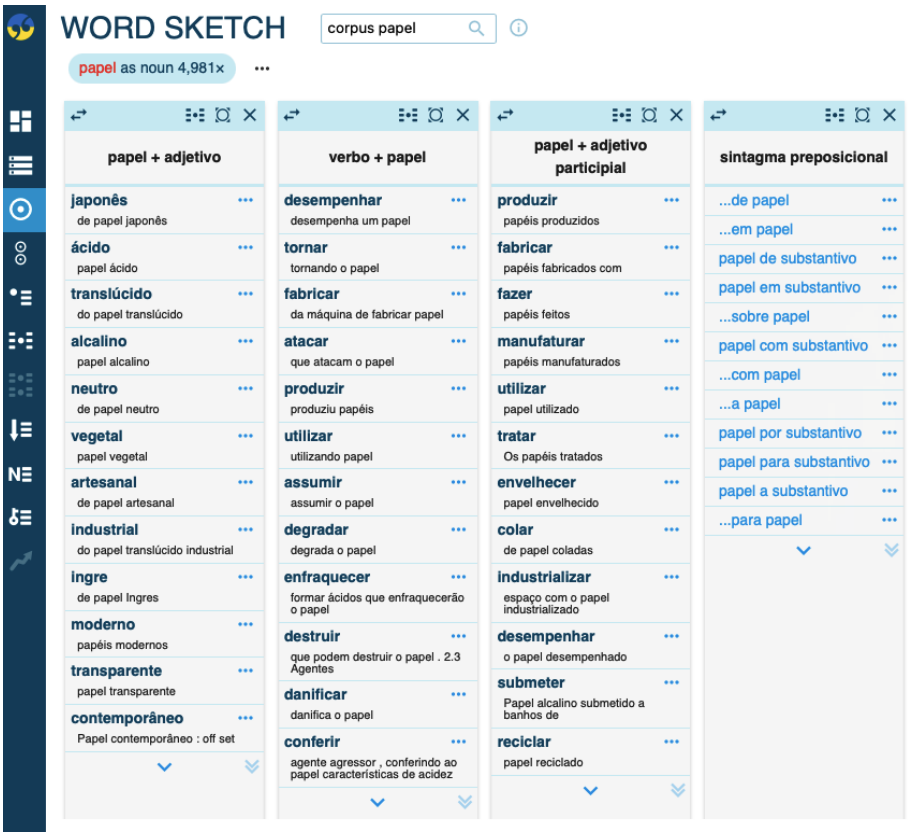


Fonte: Sketch Engine.

19 O *corpus* de referência do português oferecido pelo *Sketch Engine*, o *Portuguese Web 2018* (ptTenTen18), está formado por aproximadamente 4 bilhões de palavras e representa o português brasileiro.

Word Sketch: a ferramenta *Word Sketch* oferece o padrão colocacional de uma palavra pesquisada, reunindo, em uma única página, o sumário do comportamento colocacional do item buscado, organizado por suas relações sintáticas. Por exemplo, se no *Corpus* Papel do Termisul for buscado o termo *acervo*, o *Word Sketch* indicará, por estrutura sintática, as palavras que costumam aparecer juntas com esse termo, considerando não somente a frequência, mas também outros índices matemáticos, como o *Mutual Information* (MI) (Church; Hanks, 1990), responsáveis por identificar estruturas complexas cujas palavras apresentam uma atração semântica entre si. Assim, o *Word Sketch* permitirá identificar para o termo *papel* colocações especializadas formadas por diferentes estruturas morfossintáticas como, por exemplo, nome + adjetivo: *papel japonês, papel translúcido, papel vegetal;* verbo + nome: *degradar papel, enfraquecer papel, danificar papel*; nome + particípio: *papel envelhecido, papel reciclado*; nome deverbal + preposição + nome: *restauração de papel, deterioração de papel*; e outras estruturas possíveis (ver figura 3.7).

Figura 3.7 – *Word Sketch* do termo *papel*



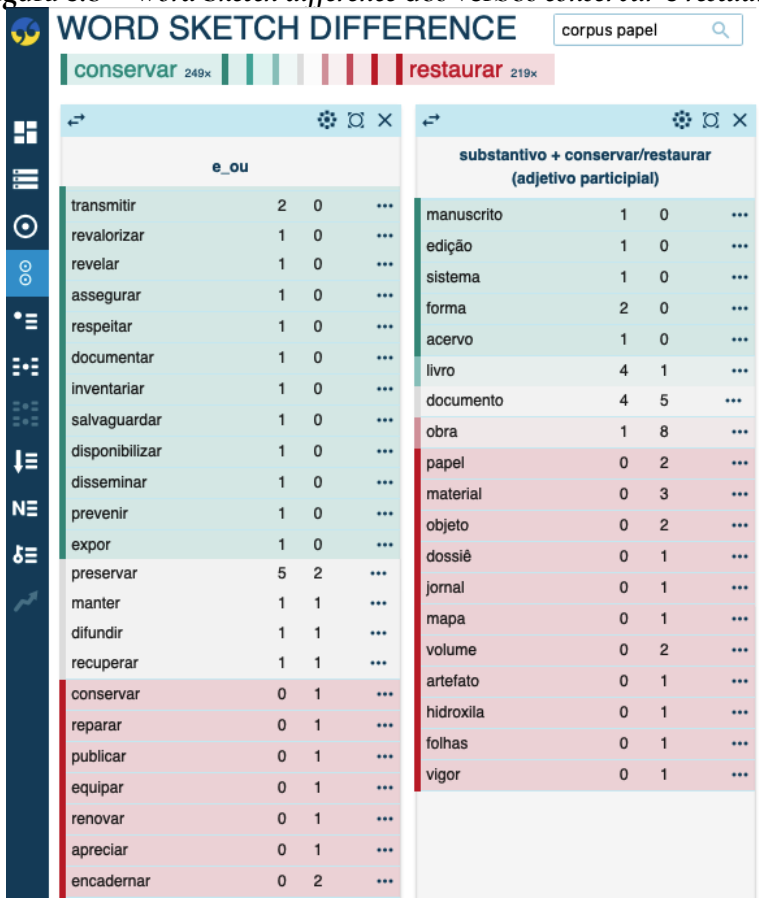
Fonte: Sketch Engine.

Word Sketch difference: segue o mesmo funcionamento do *Word Sketch*, porém é usado para comparar e contrastar o padrão colocacional de duas palavras diferentes. Ou seja, duas palavras são pesquisadas no *corpus* e a ferramenta indicará com quais outros itens lexicais cada uma costuma aparecer. Cada palavra pesquisada tem uma cor – vermelho ou verde – que será aplicada aos colocados, identificando o grau de atração que há entre eles e o item de pesquisa. As palavras indicadas na parte central de fundo cinza são aquelas que costumam ser usadas com ambos os itens pesquisados.

Esse é um recurso bastante útil para a tradução de língua geral, por exemplo, para decidir entre palavras que apresentam valores sinônimos,

mas que não são empregadas nas mesmas combinações de palavras. No caso da terminologia, essa ferramenta pode ser bastante proveitosa para a conceitualização (através da criação de árvores de domínio ou mapas conceituais) de unidades de valor especializado. A figura 3.8 ilustra a comparação entre os verbos *conservar* e *restaurar*, termos que definem a área de estudo do *Corpus* Papel. Observa-se, a partir do padrão colocacional desses verbos, que *conservar* é um termo associado, por um lado, a uma noção de cuidado: *guardar, salvaguardar, prevenir, valorizar*; e, por outro, a uma noção de registro: *transmitir, disseminar, documentar, expor*. Já o verbo *restaurar* associa-se às ações práticas da área: *reparar, encadernar, renovar.*

Figura 3.8 – *Word Sketch difference* dos verbos *conservar* e *restaurar*

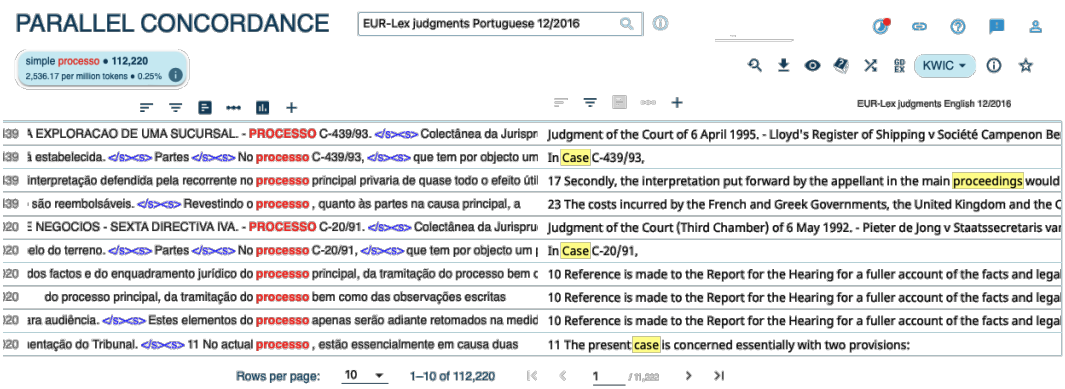


Fonte: Sketch Engine.

Também é possível compreender os valores desses verbos e as diferenças conceituais entre eles pela estrutura de colocação participial (segunda coluna da figura 3.8). Enquanto *conservar* é uma prática realizada com *acervos* e *sistemas*, a prática de *restaurar* é realizada com os bens em suporte papel em si, como *folhas, jornais, mapas, artefatos* etc. Com essa análise, pode-se chegar a conclusões de que *conservar* é uma tarefa mais voltada para o âmbito institucional e de gestão, enquanto o ato de *restaurar* envolve procedimentos práticos realizados nos bens materiais.

Parallel concordance: para usar essa ferramenta, é necessário ter *corpora* paralelos disponíveis. É o caso do *corpus* EUR-LEX JUDGMENTS, disponibilizado pelo próprio programa, que oferece textos jurídicos traduzidos. *Corpora* paralelos são bastante úteis para a tradução especializada. A figura 3.9 demonstra o exemplo da busca por um equivalente, em inglês, do termo jurídico *processo*. As concordâncias paralelas oferecem diferentes contextos do termo *case* sendo empregado em sua maioria, além da opção *proceedings*, cabendo ao tradutor analisar e optar pela melhor solução tradutória.

Figura 3.9 – *Parallel concordance* do termo *processo*



Fonte: Sketch Engine.

Fechamos o capítulo com a proposta de atividades que permitirão que você pratique todas as etapas da construção de um *corpus* aqui descritas

e seu processamento nas duas ferramentas apresentadas. Esperamos que elas sirvam como um ponto de partida para futuras pesquisas com o uso de *corpus*.

ATIVIDADES: *compilação de* corpus *de estudo*

Agora você pode praticar um pouco o que abordamos até agora, construindo um pequeno *corpus* e testando as ferramentas de extração. Sugerimos que construa um *corpus* em língua portuguesa e depois experimente construir outro com uma língua estrangeira para observar diferenças e semelhanças entre elas.

1. Escolha uma área de seu interesse e pense em um tópico dentro dessa área para fazer o seu recorte. O nosso projeto, por exemplo, é da área da *Conservação e Restauração de bens móveis em suporte papel (Projeto Papel)*, sendo nosso objetivo identificar os termos e as UFEs dessa área.

2. Agora que já tem a área e o recorte definidos, busque na internet 20 textos de diferentes gêneros sobre o tópico escolhido. Você pode usar a pesquisa avançada do Google, que permite identificar textos de países determinados, escritos em determinado idioma, e que também oferece o filtro de busca por palavras-chave específicas.

3. A seguir, os textos precisam passar por uma limpeza, serem salvos em <.txt> (codificação UTF-8) e catalogados. Veja acima, na seção "Critérios de coleta, limpeza e organização", o que falamos sobre os elementos que precisam ser retirados dos textos antes de serem salvos. Não se esqueça de criar seu próprio sistema de código para ordenar os textos e incluí-los no seu catálogo.

4. Após compilar o *corpus,* gere sua lista de palavras usando o *AntConc*. A primeira palavra lexical dessa lista deve ser representativa da temática do seu *corpus*, isto é, provavelmente corresponderá a uma das palavras-chave

que você usou para a busca e seleção de textos. Se isso não acontecer, reveja os critérios de seleção e limpeza dos textos escolhidos para compor seu *corpus*.

5. Se quiser identificar termos e fraseologias especializadas da área de seu interesse sobre a qual você compilou seu *corpus*, você pode fazer uma conta gratuita no programa *Sketch Engine* e compilar seu *corpus* de estudo ali. Em seguida, gere a *keyword list* e identifique os principais termos que aparecem nessa lista (para saber mais sobre identificação de candidatos a termo e sobre o valor terminológico da unidade lexical, leia o capítulo 4).

6. Depois de identificar alguns termos, use a ferramenta *Word Sketch* para observar se esses termos formam fraseologias especializadas. Para identificá-las, pesquise o termo no *Word Sketch* (sugerimos que, se seu *corpus* tiver menos de 1 milhão de palavras, escolha um corte de frequência que fique entre 2 e 5 ocorrências). Em seguida, observe e selecione as estruturas verbais e nominalizadas com as quais esse termo aparece registrado, isto é, [verbo + nome(termo)] e [nome deverbal + de + nome(termo)]. Essas serão as estruturas candidatas a unidades fraseológicas especializadas da linguagem que seu *corpus* de estudo representa. Para saber mais sobre identificação e extração de fraseologias especializadas, leia, também, o capítulo 4.

Você também pode identificar os candidatos a termos do seu *corpus* com a ferramenta *Keyword list*, do *AntConc*; contudo, para isso, você precisará ter um *corpus* de referência para contrastar com o seu de estudo. Da mesma forma, uma vez identificados alguns termos, você também pode buscar pelas fraseologias usando a ferramenta *Clusters/N-grams* do *AntConc*.

PARA SABER MAIS

ARCOS, Manuela; BEVILACQUA, Cleci R. UFE eventivas na área da conservação e restauração de bens culturais móveis em suporte papel: identificação e análise. *Debate Terminológico*, [Porto Alegre], n. 18, p. 4-18, 2020. Disponível em: https://seer.ufrgs.br/riterm/article/view/98700. Acesso em: 17 set. 2021.

ARCOS, Manuela; BEVILACQUA, Cleci R. Metodologias para a extração e identificação de unidades fraseológicas especializadas eventivas em *corpora* textuais. *Guavira Letras*, v. 27, p. 75-95, 2018. Disponível em: http://websensors.net.br/seer/index.php/guavira/issue/view/34/showToc. Acesso em: 7 jun. 2022.

BAKER, Mona. Corpora in Translation Studies. An overview and suggestions for future research, *Target*, v. 7, n. 2, p. 223-243, 1995.

BAKER, Mona. Corpus Linguistics and Translation Studies: implications and applications. *In*: BAKER, Mona; FRANCIS, Gill; TOGNINI-BONELLI, Elena (org.). *Text and Technology:* in honour of John Sinclair. Amsterdam: John Benjamins, 1993. p. 233-250.

BERBER SARDINHA, Antônio Paulo. *Linguística de Corpus*. Barueri: Manola, 2004.

BERBER SARDINHA, Antônio Paulo. Histórico e problemática. *D.E.L.T.A.*, v. 16, n. 2, p. 323-367, 2000.

BIBER, Douglas. Representativeness in corpus design. *Literary and Linguistic Computing*, v. 5, n. 4, p. 243-257, 1993.

MARCUSCHI, Luiz Antônio. *Produção textual, análise de gêneros e compreensão*. São Paulo: Parábola Editorial, 2008.

MARCUSCHI, Luiz Antônio. Gêneros textuais: definição e funcionalidade. *In*: DIONÍSIO, Ângela P.; MACHADO, Anna Raquel; BEZERRA, Maria Auxiliadora (org.). *Gêneros textuais e ensino*. 4. ed. Rio de Janeiro: Lucerna, 2005.

MCENERY, Tony; HARDIE, Andrew. *Corpus Linguistics: method, theory and practice*. Edinburgh: Cambridge University Press, 2012.

TAGNIN, Stella E. O. Glossário de linguística de *corpus*. *In: Corpora no ensino de línguas estrangeiras.* São Paulo: HUB Editorial, 2011. p. 357-361.

Referências

ALUÍSIO, Sandra M.; ALMEIDA, Glades M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, v. 4, n. 3, p. 156-178, 2006.

ALVES, Ieda M. A unidade lexical neológica: do histórico-social ao morfológico. *In*: ISQUERDO, Aparecida N.; KRIEGER, Maria da Graça. *As Ciências do Léxico*. Lexicologia, Lexicografia, Terminologia. Campo Grande: Editora UFMS, 2004. p. 77-100. v. II.

ARCOS, Manuela; BEVILACQUA, Cleci R. UFE eventivas na área da conservação e restauração de bens culturais móveis em suporte papel: identificação e análise. *Debate Terminológico*, [Porto Alegre], n. 18, p. 4-18, 2020. Disponível em: https://seer.ufrgs.br/riterm/article/view/98700. Acesso em: 17 set. 2021.

ARCOS, Manuela; BEVILACQUA, Cleci R. Metodologias para a extração e identificação de unidades fraseológicas especializadas eventivas em *corpora* textuais. *Guavira Letras*, v. 27, p. 75-95, 2018. Disponível em: http://websensors.net.br/seer/index.php/guavira/article/view/714 . Acesso em: 17 set. 2021.

BAKER, Mona. Corpora in Translation Studies. An overview and suggestions for future research. *Target*, v. 7, n. 2, p. 223-243, 1995.

BAKER, Mona. Corpus Linguistics and Translation Studies: implications and applications. *In*: BAKER, Mona; FRANCIS, Gill; TOGNINI-BONELLI, Elena (org.). *Text and Technology:* in honour of John Sinclair. Amsterdam: John Benjamins, 1993. p. 233-250.

BERBER SARDINHA, Tony. Histórico e problemática. *D.E.L.T.A.*, v. 16, n. 2, p. 323-367, 2000.

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

BEVILACQUA, Cleci R. Fraseologia Especializada: panorama das pesquisas realizadas no Brasil. *In*: SILVA, Suzete (org.). *Fraseologia & Cia*: entabulando diálogos reflexivos. Campinas, SP: Pontes, 2020. p. 41-66. v. 2.

BEVILACQUA, C. R. Investigación Sistemática en Terminología. *In*: ÁLVAREZ CATALÁ, Sara; BARITÉ, Mario (org.). *Teoría y praxis en Terminología*. Montevidéu:

Ediciones Universitarias, Unidad de Comunicación de la Universidad de la República, 2017. p. 69-90. v. 1.

BEVILACQUA, Cleci R. *Unidades Fraseológicas Especializadas Eventivas: descripción y reglas de formación en el ámbito de la energía solar.* Tese (Doutorado). Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 2004.

BEVILACQUA, Cleci R. Terminologia mono/bi/multilíngue: algumas propostas e reflexões referentes às unidades fraseológicas especializadas. *TradTerm*, n. 8, p. 135-147, 2002.

BEVILACQUA, Cleci R. *A fraseología jurídico-ambiental*. 1996. Dissertação (Mestrado em Estudos da Linguagem). Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Letras, Porto Alegre, 1996.

BEVILACQUA, Cleci R.; FINATTO, Maria José B.; REUILLARD, Patrícia C. R. Glossário de gestão ambiental: estabelecimento de equivalentes em alemão, espanhol e francês. *Tradução & comunicação:* Revista Brasileira de Tradutores, São Paulo, n. 19, p. 61-72, 2009.

BEVILACQUA, Cleci R.; MACIEL, Anna Maria B. A variação terminológica em uma base de dados de combinatórias léxicas especializadas: descrição e tratamento. *In*: ISQUERDO, Aparecida N.; DAL CORNO, Giselle O. M. (org.). *As Ciências do Léxico*, Campo Grande: Ed. UFSM, 2018. p. 273-290. v. VIII.

BEVILACQUA, Cleci R *et al.* Combinatórias léxicas especializadas da linguagem legislativa: uma abordagem orientada pelo *corpus*. *In*: MURAKAWA, Clotilde; NADIN, Odair Luiz (ed.). *Terminologia: uma ciência interdisciplinar*. São Paulo: Cultura Acadêmica, 2013. p. 227-243.

BEVILACQUA, Cleci R. *et al.* Acervo Termisul: implantação das bases textuais. *In*: CONGRESSO INTERNACIONAL DA ASSOCIAÇÃO BRASILEIRA DE LINGUÍSTICA (ABRALIN), 7, 2009, João Pessoa. *Anais...* João Pessoa: Ideia, 2009. v. 1. p. 815-824. 2009.

BIBER, Douglas. Representativeness in corpus design. *Literary and Linguistic Computing*, v. 5, n. 4, p. 243-257, 1993.

BOJANOSKI, Silvana F. *Terminologia em Conservação de bens culturais em papel: produção de um glossário para profissionais em formação.* Tese (Doutorado) –Programa de Pós-Graduação em Memória Social e Patrimônio Cultural, UFPEL, 2018.

BOJANOSKI, Silvana F.; MICHELON, Francisca; BEVILACQUA, Cleci Regina. Criação do *corpus* para um estudo terminológico da área da conservação e restauração de bens culturais. *Debate Terminológico,* n. 17, p. 33-45, 2017.

BOURIGAULT, Didier; SLODZIAN, Monique. Pour une terminologie textuelle. *Terminologies Nouvelles*, n. 19, déc. 1998-juin. 1999.

CABRÉ, María Teresa. *Terminología: representación y comunicación.* Una teoría de base comunicativa y otros artículos. Barcelona: Universitat Pompeu Fabra; Instituto Universitário de Linguística Aplicada, 1999.

CABRÉ, María Teresa. *La terminologia: teoría, metodologia, aplicaciones.* Barcelona: Editorial Empúries, 1993.

CHICHORRO, Caroline L. C. M. *Terminologia do Licenciamento Ambiental em português e inglês*. Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, 2016.

CHURCH, Kenneth W.; HANKS, Patrick. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, n. 16, p. 22-29, 1990.

DIKI-KIDIRI, Marcel. Eléments de terminologie culturelle. *Cahiers du Rifal*, v. 26, 2007.

FABER, Pamela; MÁRQUEZ, Carlos; VEGA, Miguel. Framing Terminology: A Process-Oriented Approach. M*eta: journal des traducteurs* / Meta: Translators› Journal, v. 50, n. 4, 2005. Disponível em: https://www.erudit.org/en/journals/meta/2005-v50-n4-meta1024/019916ar.pdf. Acesso em: 15 mar. 2022.

FAULSTICH, E. A socioterminologia na comunicação científica e técnica. *Ciência e Cultura*, v. 58, n. 2, 2006. Disponível em: http://cienciaecultura.bvs.br/pdf/cic/v58n2/a12v58n2. Acesso em: 8 abr. 2022.

FINATTO, Maria José B. A definição de termos técnico-científicos no âmbito dos estudos de terminologia. *Rev. Est. Ling.*, Belo Horizonte, v. 11, n. 1, p. 197-222, jan./jun. 2003. Disponível em: http://periodicos.letras.ufmg.br/index.php/relin/article/view/2351 Acesso em: 18 abr. 2022.

FINATTO, Maria José B. Termos, textos e textos com termos: novos enfoques dos estudos terminológicos de perspectiva linguística. *In*: ISQUERDO, Aparecida

N.; KRIEGER, Maria da G. (org.). *As Ciências do Léxico.* Campo Grande, MS: Editora UFMS, 2004. v. II.

FINATTO, Maria José. B. *Definição terminológica*: fundamentos teórico-metodológicos para sua descrição e explicação. 2001. Tese (Doutorado em Estudos da Linguagem) – Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Letras, Porto Alegre, 2001.

FISH, Stanley E. *Is There a Text in This Class?*: The Authority of Interpretive Communities. Cambridge, Mass: Harvard University Press, 1980.

FROMM, Guilherme *et al*. Wordsmith Tools e Sketch Engine: um estudo analítico-comparativo para pesquisas científicas com uso de corpora. *Revista de Estudos Linguísticos*, Belo Horizonte, v. 28, n. 3, p. 1.101-1.248, 2020.

GAUDIN, François. *Pour une socioterminologie*. Des problèmees sémantiques aux pratiques institutionnelles. Rouen: Publications de l'Université de Rouen, 1993. Disponível em: http://websensors.net.br/seer/index.php/guavira/issue/view/34/showToc. Acesso em: 7 jun. 2022.

KILIAN, Cristiane K. *A retomada de unidades de significação especializada em textos em língua alemã e portuguesa sobre gestão de resíduos: uma contribuição para a tradução técnico-científica.* Tese (Doutorado) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Instituto de Letras, Porto Alegre, 2007.

KILIAN, Cristiane K.; LOGUERCIO, Sandra D. Fraseologias de gênero em resumos científicos de Linguística, Engenharia de Materiais e Ciências Econômicas. *Tradterm*, n. 26, p. 241-267, 2015.

KRIEGER, Maria da Graça. Terminografia: entre teoria e aplicações. *In*: ISQUERDO, Aparecida N.; DAL CORNO, Giselle O. M. *As Ciências do Léxico*. Lexicologia, Lexicografia, Terminologia, Campo Grande, MS: Editora UFMS, 2018. p. 329-346. v. VIII.

KRIEGER, Maria da Graça. Porque Lexicografia e Terminologia: relações textuais. *In*: ENCONTRO DO CELSUL (Círculo de Estudos Linguísticos do Sul), 8., 2008. *Anais...* Pelotas: Educat, 2008.

KRIEGER, Maria da Graça. *Terminologias em construção: procedimentos metodológicos. In*: CONGRESSO INTERNACIONAL DA ABECAN (Associação Brasileira

de Estudos Canadenses), 8., 2005. *Anais*... Gramado, 2005. Disponível em: http://www.ufrgs.br/termisul/files/file112160.pdf. Acesso em: 10 nov. 2021.

KRIEGER, Maria da Graça. Do reconhecimento de terminologias: entre o linguístico e o textual. *In*: ISQUERDO, Aparecida N.; KRIEGER, Maria da Graça. *As Ciências do Léxico*. Lexicologia, Lexicografia, Terminologia, Campo Grande, MS: Editora UFMS, 2004, p. 327-339. v. II.

KRIEGER, Maria da Graça. Sobre Terminologia e seus objetos. *In*: LIMA, Marília; RAMOS, Patrícia C. (org.). *Terminologia e ensino de segunda língua*: Canadá e Brasil. Porto, Alegre: NEC, IL, UFRGS/Abecan, 2001. p. 45-53.

KRIEGER, Maria da Graça. Terminologia revisitada. *DELTA*, v. 16, n. 2, p. 209-228, 2000.

KRIEGER, Maria da Graça. Terminologia em contextos integradores: funcionalidade e fundamentos. *Organon*, v. 12, n. 26, p. 19-31, 1998.

KRIEGER, Maria da Graça; FINATTO, Maria José B. *Introdução à Terminologia*: Teoria & Prática. São Paulo: Contexto, 2004.

KRIEGER, Maria da Graça; MACIEL, Anna Maria Becker; FINATTO, Maria José Bocorny. Terminografia das leis do meio ambiente: princípios teórico-metodológicos. *In*: KRIEGER, Maria da Graça; MACIEL, Anna Maria Becker (org.). *Temas de terminologia*. Porto Alegre/São Paulo: Ed. Universidade/UFRGS/Humanitas/USP, 2001. p. 317-335.

KRIEGER, Maria da Graça *et al*. *Dicionário de Direito Ambiental*: terminologia das leis do meio ambiente. 2. ed. rev. e atualizada. Rio de Janeiro: Lexicon, 2008.

KRIEGER, Maria da Graça *et al*. *Glossário de gestão ambiental*. Barueri, SP: Disal, 2006.

KRIEGER, Maria da Graça *et al*. *Glossário Multilíngue de Direito Ambiental Internacional*. Rio de Janeiro: Ed. Forense, 2004.

KRIEGER, Maria da Graça *et al*. *Dicionário de direito ambiental*: terminologia das leis do meio ambiente. Porto Alegre: Editora da Universidade/UFRGS, 1998.

LAZZARIN, Renan. *AGROTÓXICO E PFLANZENSCHUTZMITTEL: estudo exploratório da variação terminológica e proposição de equivalentes tradutórios no par de línguas português-alemão*. Trabalho de conclusão de curso (Bacharel em Letras – Tradutor Português e Alemão) – Universidade Federal do Rio Grande do Sul.

Instituto de Letras, Porto Alegre, 2007. Disponível em: https://lume.ufrgs.br/handle/10183/178858. Acesso em: 22 ago. 2023.

LEECH, Geoffrey Corpora. *In*: MALMKJAER, Kirsten (ed.). *The Linguistics Encyclopedia*. London: Routledge, 1991. p. 73-80.

LOGUERCIO, Sandra D. A linguagem comum do artigo científico em português brasileiro: um estudo baseado em corpus. *ANTARES*, v. 12, n. 25, p. 140-164, jan./abr. 2020.

LOGUERCIO, Sandra D. Entre buscar contribuir e la contribution: a modalização em resumos científicos em português/francês. *Linguagem & Ensino*, v. 22, n. 3, p. 881-995, jul./set. 2019.

LOGUERCIO, Sandra D.; KILIAN, Cristiane K. Fraseologias de gênero de resumos de artigos científicos (português, alemão e francês). In: Claudia Zavaglia; Angélica Karim Garcia Simão. (Org.). Reflexões, tendências e novos rumos dos estudos fraseoparemiológicos. 1ed.São José do Rio Preto (SP): UNESP/IBILCE, 2017, v., p. 94-108.

MACIEL, Anna Maria B. Reflexão sobre a pesquisa terminológica em *corpus*. *In*: ENCONTRO NACIONAL DA ANPOLL, 21, São Paulo. *Domínios do Saber*: História, Instituições, Práticas, 2006. Disponível em: https://silo.tips/queue/xxi--encontro-nacional-da-anpoll-associaao-nacional-de-pos-graduaao-e-pesquisa--em?&queue_id=-1&v=1654593928&u=MmEwMTo0YjAwOjg0NGQ6YWIwMDo5YzM3OmVlZjplNzMxOmE3ZmM=. Acesso em: 7 jun. 2022.

MACIEL, Anna Maria B. *Para o reconhecimento da especificidade do termo jurídico*. Tese (Doutorado em Estudos da Linguagem) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Letras, Porto Alegre, 2001.

MACIEL, Anna M.; BEVILACQUA, Cleci R. A fraseologia da legislação do Direito Ambiental em línguas e sistemas jurídicos distintos. *In*: ZAVAGLIA, Claudia; SIMÃO, Angélica (org.). *Reflexões, tendências e novos rumos dos Estudos Fraseoparemiológicos.* São José do Rio Preto: Unesp, 2017. p. 46-56.

MACIEL, Anna Maria B.; REUILLARD, Patrícia C. R. Abordagem da variação terminológica em uma base de dados de combinatórias léxicas. *Tradterm*, São Paulo, v. 26, p. 223-240, 2015. Disponível em: https://doi.org/10.11606/issn.2317-9511.v26i0p223-240. Acesso em: 12 out. 2021.

MARCUSCHI, Luiz Antônio. *Produção textual, análise de gêneros e compreensão*. São Paulo: Parábola Editorial, 2008.

MARCUSCHI, Luiz Antônio. Gêneros textuais: definição e funcionalidade. *In*: DIONÍSIO, Ângela P.; MACHADO, Anna Raquel; BEZERRA, Maria Auxiliadora (org.). *Gêneros textuais e ensino*. 4. ed. Rio de Janeiro: Lucerna, 2005.

MCENERY, Tony; HARDIE, Andrew. *Corpus Linguistics: method, theory and practice.* Edinburgh: Cambridge University Press, 2012.

NORD, Christiane. *Traducir, una actividad con propósito.* Introducción a los enfoques funcionalistas. Berlim: Frank & Timme GmbH, 2018.

NORD, Christiane. Lealdade em vez de fidelidade: proposta de uma tipologia funcional da tradução. *Cadernos de Tradução*, Porto Alegre, Número Especial, p. 9-24, 2016.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). *Convenção das Nações Unidas sobre Direito do Mar*. 1990. Disponível em: http://www.unbciencia.unb.br/images/Noticias/2019/12-Dez/Convencao_das_Nacoes_Unidas_sobre_Direito_do_Mar_Montego_Bay.pdf. Acesso em: 8 jul. 2022.

REUILLARD, Patrícia C. R. Neologismos lacanianos e equivalência tradutória. Tese (Doutorado em Estudos da Linguagem) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Letras, Porto Alegre, 2007. Disponível em: https://lume.ufrgs.br/handle/10183/12506. Acesso em: 23 ago. 2023.

SWALES, John M. *Genre Analysis*: English in Academic and Research Settings. Cambridge: CUP, 2002[1990].

TAGNIN, Stella E. O. Glossário de linguística de *corpus*. *In: Corpora no ensino de línguas estrangeiras.* São Paulo: HUB Editorial, 2011. p. 357-361.

TEMMERMAN, R. *Towards new ways of Terminology description.* Amsterdam: John Benjamins, 2000.

TUTIN, Agnès. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue Française de Linguistique Appliquée*, n. 2, v. XII, p. 5-14, 2007.

Chave de respostas das atividades propostas

Capítulo 1 – Quando a teoria e a prática se encontram
O capítulo 1 não possui atividades por ser um capítulo teórico e que embasa os demais capítulos do livro.

Capítulo 2 – As decisões prévias
As respostas para as atividades propostas no capítulo 2 dependem das obras selecionadas para a realização das atividades, razão pela qual não apresentamos um gabarito.

Capítulo 3 – Constituição de *corpora*: critérios de coleta, limpeza e organização
As respostas para as atividades propostas no capítulo 3 dependem da área a ser selecionada para a construção de *corpus*, razão pela qual não apresentamos um gabarito.

Capítulo 4 – Seleção de unidades terminológicas: estratégias de extração e princípios de identificação

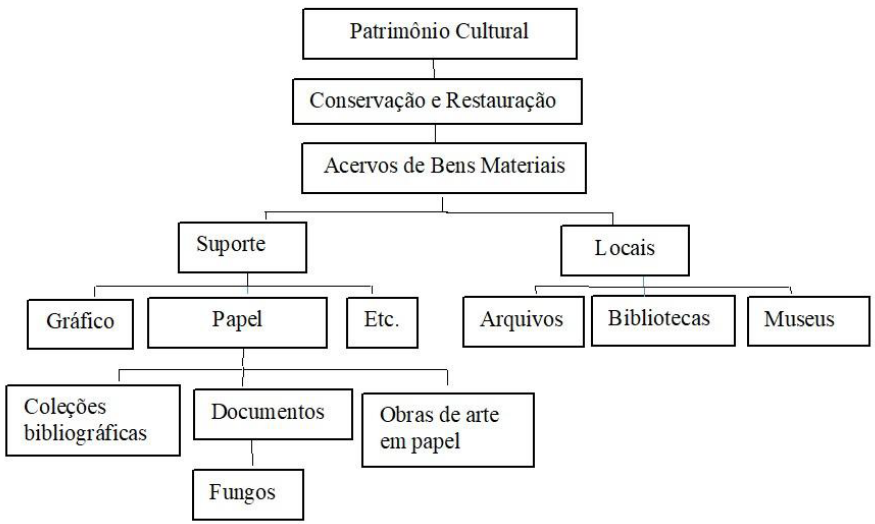
Exercício 1: O termo definido no trecho do *Corpus* Papel é *arquivo*. Nesse fragmento, o termo apresenta uma frequência de cinco ocorrências. Além disso, o termo *arquivo* aparece acompanhado pelo verbo *definir* em três contextos definitórios, sendo eles: 1) "[...] o arquivo é definido como: um conjunto de documentos produzidos e recebidos por órgãos públicos (...)", 2) "[...] o arquivo não se define pela forma dos documentos ou por sua origem, mas pela razão para que foram criados e por sua forma de acumulação orgânica" e 3)" [...] os elementos que definem os arquivos podem ser resumidos em três fatores que são abstratos [...]".

Exercício 2: As UFEs formadas a partir do termo arquivo são do tipo colocação (nesse caso, UFE eventivas), pois estão formadas por [verbo + termo]

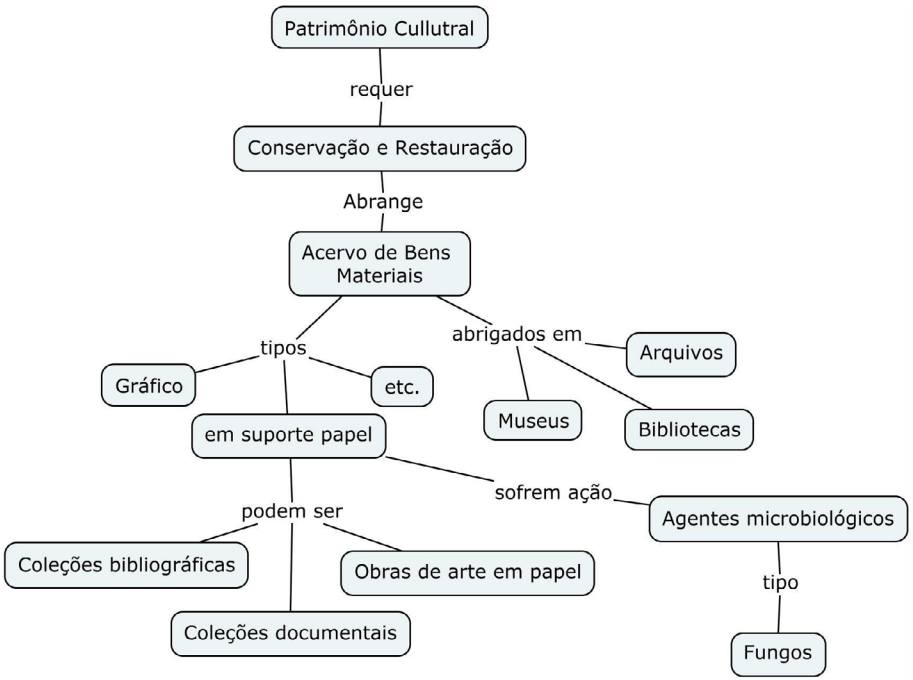
ou [nominalização + de + termo]. São elas: *abrigar arquivo*, *organização de arquivo*, *conservação de arquivo*, *catalogação de arquivo* e *microfilmagem de arquivo*.

Exercício 3: A área de conhecimento pode ser identificada, mais amplamente, como sendo do **Patrimônio Cultural** (cf. linhas 1, 2 e 5), e mais especificamente, como a de **Conservação e Restauração**, vista na referência ao *corpus* de onde foi extraído o texto. Já o assunto abordado é **fungos em acervo de papel** (introduzido nas linhas 14 a 16 e especificado nas linhas 20, 23, 29 e 30). Isso é feito em um **artigo científico**, gênero identificado pelo registro escrito, pela estrutura textual-discursiva (texto segmentado em parágrafos que trazem contextualização da área e do tema, justificativa da pesquisa, indicação do objeto de estudo e dos objetivos etc.) e por unidades lexicais e fraseológicas que remetem mais especificamente ao relato científico.

Exercício 4:
Sugestão de árvore de domínio



Sugestão de mapa conceitual



Exercício 5: O léxico relativo ao gênero artigo científico (também chamado de léxico metacientífico) torna-se saliente no excerto a partir da linha 14, com *No presente trabalho optou-se por*, em que **trabalho** faz referência ao próprio artigo e a fórmula introduz o tema geral do estudo. Também podem ser identificadas as seguintes unidades: **orientar esta pesquisa**, **esta pesquisa pretende**, **por meio de uma investigação** (l. 26), **estudo de caso** (l. 28), [estudar] **métodos de tratamento para** (l. 29), **a pesquisa busca** (l. 30). Também podemos pensar em palavras como: **trabalho, pesquisa, investigação, estudo de caso, estudar, método(s)**.

Exercício 6: c / d / e / a / b

Capítulo 5 – A ficha terminológica

Exercício 1:

TERMO: água

Língua: português
Contexto: No tanque superior se dá o processo da reenfibragem, que é a passagem de uma solução de água + polpa de papel através de uma tela semipermeável onde está o documento a ser restaurado. Como resultado esperado temos o depósito da polpa nas áreas do documento onde houve perdas de material. No tanque inferior armazena-se a água após o processo de reenfibragem que, por ser deionizada e trifiltrada, é de custo elevado, portanto não deve ser desperdiçada. (ptPP023)
Ver também:
água quente
água deionizada
água destilada
água desmineralizada
Equivalente(s) em Inglês:
water 2
Equivalente(s) em Espanhol:
agua 2
Equivalente(s) em Francês:
eau 2
Equivalente(s) em Italiano:
acqua 2
Equivalente(s) em Russo:
вода 2 [voda]

Exercício 2: Como explicado no capítulo, a ficha vai variar de acordo com os diversos fatores envolvidos. Lembre-se de que ela costuma ter Entrada; Categoria gramatical, Gênero e Número; Fonte da entrada; Definição; Fonte da definição; Contexto; Fonte do contexto; Remissivas; Equivalentes; e Notas.

Exercício 3: ver respostas do exercício 1.

Capítulo 6 – Busca e identificação de equivalentes em línguas estrangeiras

Exercício 1:

Língua	Termo	Equivalente
Espanhol	cartão alcalino	cartón libre de ácido
Francês	envelhecimento do papel	vieillissement du papier
Inglês	atmosfera anóxia	anoxic atmosphere
Italiano	banho aquoso	lavaggio acquoso
Russo	solubilidade de tintas	водное растворение чернил [vodnoe rastvorienie tchernil]

Para identificar os equivalentes das atividades 2 e 3, você pode consultar as bases do grupo Termisul disponíveis em www.ufrgs.br ou outras fontes confiáveis de consulta, como *sites* de universidades, de outros grupos de pesquisa e o portal de periódicos da Capes, por exemplo.