Metodologias para a extração e identificação de unidades fraseológicas especializadas eventivas em corpora textuais

Methodologies for the extraction and identification of Eventive Specialized Phraseological Units in textual corpora

Metodologías para la extracción e identificación de unidades fraseológicas especializadas eventivas en *corpus* textuales

Manuela ARCOS[1]
Cleci Regina BEVILACQUA[2]

RESUMO: O objetivo deste trabalho é propor diferentes metodologias de extração e identificação de Unidades Fraseológicas Especializadas Eventivas (UFE eventivas) da área de Conservação e Restauração de Bens Móveis em suporte papel a partir da comparação de dois programas computacionais: *AntConc* e *Sketch Engine*. As UFE eventivas se caracterizam pela sua conformação por um Núcleo Eventivo, que pode ocorrer em forma de verbo ou nominalização, e um Núcleo Terminológico, e se caracterizam por denotar ações e processos especializados de um âmbito do conhecimento. O referencial teórico que fundamenta o trabalho segue os princípios da Teoria Comunicativa da Terminologia, da Terminologia Textual e da Fraseologia Especializada. Espera-se, a partir dos resultados, propor diferentes formas de busca para cada um dos programas a partir dos pontos positivos e negativos identificados. Assim, buscamos com este trabalho oferecer uma contribuição para os estudos de Fraseologia Especializada, principalmente no que tange à identificação de UFE eventivas em *corpora* textuais através dos recursos da Linguística de Corpus.

PALAVRAS-CHAVE: Unidades Fraseológicas Especializadas. Terminologia. Linguística de Corpus. Conservação e Restauração de Bens Móveis.

ABSTRACT: The aim of this work is to propose different methodologies for the extraction and identification of Eventive Specialized Phraseological Units (Eventive SPU) in the area of Restoration and Conservation of material assets in paper support to compare two computer programs: *AntConc* and *Sketch Engine*. The Eventive SPU are conformed by an Eventive Nucleus, which may occur as a verb or a nominalization, and a Terminological Nucleus, and are characterized by denoting actions and specialized process of an area of knowledge. The theoretical framework that underlies this work follows the principles of the Communicative Therory of Terminology, the Textual Terminology and the Specialized Phraseology. From the results of this work we expect to propose different types of search for each software, considering their positive and negative aspects. We also expect to offer a

[1] Doutoranda do Programa de Pós-Graduação do Instituto de Letras da Universidade Federal do Rio Grande do Sul – UFRGS. Porto Alegre – RS – Brasil. CEP: 91540-000. E-mail: arcosmanuela@gmail.com
[2] Universidade Federal do Rio Grande do Sul – UFRGS. Instituto de Letras, Departamento de Línguas Modernas. Porto Alegre – RS – Brasil. CEP: 91540-000. E-mail: cleci.bevilacqua@ufrgs.br

contribution for the Specialized Phraseology studies, especially in what concerns the identification of Eventive SPU in textual corpora using the Corpus Linguistics' tools.

KEYWORDS: Specialized Phraseological Units. Terminology. Corpus Linguistics. Restoration and Conservation of material assets.

RESUMEN: Este trabajo propone diferentes metodologías de extracción e identificación de Unidades Fraseológicas Especializadas Eventivas (UFE eventivas) del área de la Conservación y Restauración de Bienes Muebles en soporte papel por la comparación de dos programas computacionales: *AntConc* y *Sketch Engine*. Las UFE eventivas se caracterizan por su conformación por un Núcleo Eventivo, en forma de verbo o nominalización, y un Núcleo Terminológico, y se caracterizan por expresar acciones y procesos especializados de un ámbito del conocimiento. El marco teórico que fundamenta el trabajo sigue los principios de la Teoría Comunicativa de la Terminología, de la Terminología Textual y de la Fraseología Especializada. Con los resultados, esperamos proponer diferentes formas de búsqueda para cada uno de los programas a partir de sus puntos positivos y negativos que identificamos. De esa manera, buscamos con este trabajo ofrecer una contribución para los estudios de la Fraseología Especializada, especialmente en lo que se refiere a la identificación de UFE eventivas en *corpus* textuales a través de las herramientas de la Lingüística de Corpus.

PALABRAS CLAVE: Unidades Fraseológicas Especializadas. Terminología. Linguística de Corpus. Conservación y Restauración de Bienes Muebles.

Introdução

No âmbito dos estudos da Terminologia, a Fraseologia e as Unidades Fraseológicas (UF) têm cobrado cada vez mais atenção de especialistas, tanto no que se refere à elaboração de produtos terminográficos, quanto a sua caracterização e estabelecimento de critérios linguísticos para o seu reconhecimento. Nesse sentido, a Linguística de Corpus e seus recursos têm colaborado na tarefa de processar e extrair grandes quantidades de informações linguísticas de forma semiautomática de *corpora* textuais, incluindo a extração de candidatos a UF.

O objetivo deste artigo é apresentar duas propostas metodológicas de extração e identificação de unidades fraseológicas em *corpora* textuais. Para isso, comparamos duas ferramentas computacionais, *AntConc* e *Sketch Engine* a fim de estabelecer qual *software* oferece melhores recursos para a tarefa. Ressaltamos que este trabalho está inserido em um projeto de pesquisa maior, realizado junto ao grupo Termisul[3] (Projeto Terminológico Cone Sul) cujo propósito é identificar e reunir, em uma base de dados *on-line*, a terminologia da área da Conservação e Restauração de Bens Culturais Móveis em suporte papel. Desse modo, a metodologia aqui proposta será utilizada durante a fase de identificação e extração de UF que comporão, juntamente aos termos simples e sintagmáticos, a base de dados elaborada pelo grupo.

No que se refere ao nosso objeto de estudo, reconhecemos a diversidade conceitual e denominativa que a literatura especializada oferece para as unidades fraseológicas. Desse modo, centraremos nosso estudo nas unidades chamadas Unidades Fraseológicas Especializadas Eventivas (doravante UFE eventivas) conforme a proposta de Bevilacqua (2004), que as concebe como unidades formadas por um núcleo terminológico na forma de um termo, e um núcleo eventivo na forma de um verbo, nominalização ou particípio, por exemplo, *consumir energia*, *consumo de energia* e *energia consumida*. Portanto, sua constituição parte

[3] Projeto Terminológico Cone Sul (TERMISUL) é um grupo de pesquisa teórica e aplicada em Terminologia e Terminografia da Universidade Federal do Rio Grande do Sul (UFRGS). http://www6.ufrgs.br/termisul/

de um núcleo eventivo que se refere a ações e processos especializados de uma determinada área do conhecimento.

O reconhecimento das UFE eventivas se justifica pelo papel importante que desempenham na transmissão de conhecimento dentro da linguagem especializada, uma vez que são responsáveis por comunicar as ações próprias de uma área. Da mesma forma, a distinção dos diferentes tipos de UFE interessa também à elaboração de produtos terminográficos, uma vez que, de acordo com a finalidade ou objetivo que orienta sua elaboração, se faz necessário estabelecer limites entre as diferentes unidades sintagmáticas para decidir quais estarão representadas e de que maneira em um produto terminográfico (BEVILACQUA, 2004, p. 29).

Para cumprir com os objetivos propostos, em um primeiro momento apresentamos a proposta de definição das UFE eventivas (BEVILACQUA, 2004) dentro do referencial teórico e metodológico da Teoria Comunicativa da Terminologia (TCT) (CABRÉ, 2002), e também trazemos os princípios da Terminologia Textual (BOURIGAULT; SLODZIAN, 2004). Em seguida, descrevemos os passos seguidos para a extração e identificação das UFE eventivas no *corpus* de análise comparando os dois *softwares* de processamento e análise textual: *AntConc* e *Sketch Engine*. Por fim, ilustramos e comentamos os resultados quantitativos e qualitativos a que chegamos a partir da comparação entre cada *software*.

Posicionamento teórico-metodológico

Este trabalho se insere no quadro teórico da Teoria Comunicativa da Terminologia (TCT) (CABRÉ, 2002), que concebe a linguagem especializada não em oposição à língua geral, mas como uma manifestação da mesma em contextos específicos de comunicação. Nessa perspectiva, o termo é definido como uma unidade léxica poliédrica que, utilizada em certas condições discursivas, adquire um valor especializado e, portanto, é ao mesmo tempo igual e diferente das unidades léxicas de uma língua – palavras na lexicologia. A TCT classifica as unidades linguísticas formadas por mais de um lexema como unidades polilexemáticas (UP) (CABRÉ; ESTOPÀ;LORENTE, 1996), que podem se subdividir em:

– unidades terminológicas polilexemáticas (UTP) – são unidades que podem ocupar numa frase a posição sintática de um constituinte mínimo e cujos constituintes estão completamente fixos (ex.: *acervo bibliográfico*; *arquivo público*);

– unidades fraseológicas especializadas (UFE) consideradas como cadeias sintáticas altamente frequentes que não podem ocupar em uma frase uma posição de constituinte mínimo e cujos componentes não estão totalmente fixos, mas podem apresentar variação (ex.: *preservar o acervo bibliográfico, guarda do acervo*).

Neste trabalho focamos nas UFE, mais especificamente nas UFE eventivas, conforme a proposta de Bevilacqua (2004). A autora considera as UFE eventivas como o "ambiente" em que ocorre uma unidade terminológica (UT), ou como a combinação de uma UT e seus coocorrentes. As UFE eventivas se diferenciam das UT, uma vez que representam as ações e processos específicos de um âmbito de especialidade (ex.: *restaurar acervo e restaurar acervo bibliográfico*), enquanto as UT são unidades léxicas simples ou sintagmáticas de caráter denominativo-conceitual e referencial que representam um nó da estrutura conceitual de um âmbito de especialidade (ex.: *acervo, acervo bibliográfico).*

Desse modo, as UFE eventivas são unidades sintagmáticas formadas por um ou mais termos – que constituem seu núcleo terminológico (NT) que pode ser simples ou sintagmático – e por um núcleo eventivo (NE), de caráter terminológico ou não, que pode se manifestar como verbo, nome deverbal ou particípio (BEVILACQUA, 2004, p. 18):

a) $[NE]_V$+ [NT]N[4] : *consumir energía*
b) $[NE]_{Ndev}$ + $[NT]_{SP}$: *consumo de energía*
c) $[NT]_N$ + $[NE]_{PartAdj}$: *energía consumida*[5]

Segundo a autora, o NT é de caráter nominal, conceitual e referencial, e denomina um nó ou núcleo de conhecimento no mapa conceitual da área de especialidade, podendo aparecer como termo simples ou termo sintagmático. O NE, por sua vez, é de categoria verbal ou derivada de verbo (nomes deverbais e particípios), é relacional e denota os processos e ações próprias de determinado âmbito de especialidade. Entre os dois núcleos – terminológico e eventivo – se estabelecem relações semânticas e pragmáticas determinadas pelo seu uso no âmbito temático, que conferem à unidade características como certo grau de estabilidade e fixação, e uma frequência relevante nos textos especializados (BEVILACQUA, 2004, p. 30).

Para a identificação das UFE eventivas em nosso *corpus* de análise, além dos pressupostos da TCT, anteriormente referidos, tomamos a proposta teórica e metodológica da Terminologia Textual (BOURIGAULT; SLODZIAN, 2004), que entende a análise terminológica como uma tarefa descritiva das estruturas lexicais existentes num *corpus* textual, e que deve, portanto, partir de uma análise desse mesmo *corpus*. À vista disso, a Terminologia Textual propõe uma abordagem descritiva da linguagem especializada na sua realização efetiva, isto é, no texto especializado. Nesse sentido, a tarefa de estabelecer uma terminologia de certa área de conhecimento, bem como sua fraseologia, deve partir da análise de *corpora* textuais, uma vez que é nos textos produzidos e utilizados pela comunidade de especialistas que estão expressos e disponíveis os conhecimentos compartilhados por essa comunidade (BOURIGAULT; SLODZIAN, 2004, p. 103).

Antes de apresentar as duas propostas metodológicas de extração e identificação de UFE eventivas a partir dos softwares *AntConc* e *Sketch Engine*, apresentamos informações sobre o projeto em que se insere este estudo, sobre o *corpus* de língua portuguesa de Conservação e Restauração em papel foi compilado e, em seguida, os critérios que seguimos para a identificação das UFE eventivas.

O projeto "A Linguagem do Patrimônio Cultural Brasileiro: Conservação dos Bens Culturais Móveis"

Este trabalho está inserido em um projeto de pesquisa maior, como já referido anteriormente, vinculado ao Grupo TERMISUL, intitulado *A Linguagem do Patrimônio Cultural Brasileiro: Conservação dos Bens Culturais Móveis*, cujo foco é a linguagem da área

[4] O [NE]v é formado por um verbo (indica ação ou processo especializado) que seleciona o $[NT]_N$, cuja função sintática é de objeto direto. Contudo, a estrutura $[NE]_V$ + $[NT]_N$ também se manifesta em UFE eventivas cujo NE é um verbo e o NT tem função sintática de sujeito, como em *la luz incide* (LORENTE, M; BEVILACQUA, R; ESTOPÀ, R, 2002). Neste trabalho, só consideraremos as UFE eventivas cujo NT tem função sintática de objeto direto.

[5] V = verbo; N = nome; Ndev = Nome deverbal; SP = sintagma preposicionado; PartAdj = Particípio/Adjetivo.

de conservação dos bens culturais móveis em suporte papel, isto é, de acervos documentais, bibliográficos e arquivísticos[6]. O objetivo principal do projeto é identificar e organizar a terminologia da área, e seu produto final será uma base de dados terminológica *on-line* em uma perspectiva multilíngue. Para tanto, o projeto tem como língua de partida o português e visa estabelecer as unidades equivalentes em espanhol, alemão, francês, inglês, italiano e russo. Seu público-alvo são os pesquisadores, professores e estudantes da área, bem como profissionais que atuam na produção de textos como tradutores, assessores linguísticos, redatores técnicos, entre outros. Portanto, nosso foco, neste trabalho, é estabelecer uma metodologia de extração das UFE eventivas a partir do *corpus* textual da área de Conservação e Restauração como forma de complementar a primeira etapa do projeto que reúne os termos simples e sintagmáticos da área em língua portuguesa. Num segundo momento, as UFE eventivas extraídas e identificadas deverão ser incluídas um campo específico das fichas terminológicas da base de dados, intitulado "Combinatórias".

Quanto ao *corpus* de estudo, sua compilação foi feita por bolsistas de iniciação científica do grupo Termisul, que seguiram os seguintes critérios para a seleção dos textos: a) conter as palavras-chave "documento, documentação, conservação, papel, patrimônio, preservação, restauração e restauro"; b) pertencer aos gêneros acadêmicos – livros, manuais, revistas científicas, trabalhos de conclusão de curso, dissertações, teses e boletins informativos de associações da área; c) estar incluídos em fontes confiáveis – *sites* de universidades, instituições de pesquisa, dentre outros, cuja língua original fosse o português. Todos os textos coletados passaram por um processo de limpeza (retirada de todas as informações não significativas para a pesquisa), por um processo de conversão para formato <*.txt*>, e, por fim, por catalogação. Todos esses procedimentos fazem parte de uma metodologia para a construção de *corpus*, de modo a garantir a qualidade na posterior extração das unidades terminológicas e, no nosso caso, as UFE eventivas. Atualmente, o *corpus* de língua portuguesa contém 161 textos com, aproximadamente, 38.129*types* e 967.852 *tokens*[7].

Critérios e metodologia de análise

Antes de apresentar a metodologia seguida, esclarecemos os critérios estabelecidos para o trabalho de identificação das UFE eventivas. Para tanto, nos embasamos em seus critérios de constituição: critério linguístico (sintáticos e semânticos), critério pragmático-discursivos e o critério quantitativo (BEVILACQUA, 2004).

Com relação ao critério linguístico (sintático e semântico), consideramos as duas estruturas morfossintáticas superficiais V+N; N+SP (prep+(art+)N) derivadas das estruturas $[NE]_V$+ $[NT]_N$ e$[NE]_{Ndev}$ + $[NT]_{SP}$, respectivamente, pois serão as estruturas incluídas na base de dados terminológica que está sendo elaborada pelo grupo Termisul. Assim, não consideraremos as UFE eventivas formadas por particípios. Nesse sentido, as UFE eventivas devem incluir, no mínimo, uma unidade terminológica (NT simples ou sintagmático) referente à área da conservação e restauração, e um elemento eventivo (NE), na forma de verbo ou

[6] Não abrangem mapas, gráficos nem material fotográfico.
[7] O conceito de *types* se refere ao número de palavras diferentes que ocorrem em um *corpus* e *tokens* ao número total de palavras do *corpus*.

nominalização que represente as ações e os processos caracterizadores da área, conforme ilustra o quadro 1.

Quadro 1 – Estruturas das UFE eventivas a serem identificadas

$[NE]_{V}$ + $[NT]_{N}$	V+N	*abrigar acervo*
$[NE]_{Ndev}$ + $[NT]_{SP}$	N+SP(prep+(art+)N)	*higienização de/do acervo*

Fonte: As autoras.

Quanto ao critério pragmático-discursivo, salientamos duas características do texto especializado: a fixação temática e a fixação do texto pelas propriedades pragmático-discursivas. Dessa forma, as UFE eventivas adquirem valor especializado não somente pela temática (Conservação e Restauração), mas também pela função comunicativa que desempenham no contexto de uso.

Por fim, o critério quantitativo refere-se à alta frequência de aparição no texto especializado, um traço típico das UFE eventivas. Decidimos estabelecer como critério um corte de frequência mínima de 10 ocorrências de cada UFE eventiva no *corpus*, uma vez que é o mesmo corte de frequência adotado para identificação dos termos simples e sintagmáticos a serem incluídos como entradas da base de dados terminológica. Contudo, vale ressaltar que esse recorte quantitativo pode compreender:

> – unidades que apresentam variação morfossintática (*conservação do acervo; conservação de acervo; conservação deste(s) acervo(s); conservação desse(s) acervo(s); conservação do(s) seu(s) acervo(s)*);
> – unidades que possuem um mesmo NT e um mesmo NE, mas que apresentam outros unidades léxicas inseridas em seu interior (*conservação do acervo*/ *conservação* ***adequada*** *do acervo*).

Desse modo, considerando as variações morfossintáticas das UFE eventivas, podemos chegar ao seu número total de ocorrências, sem correr o risco de eliminar estruturas que podem ter frequência inferior a 10 ocorrências pela sua variação morfossintática.

Para estabelecer uma metodologia de extração das UFE eventivas de nosso *corpus* textual, usamos os programas *AntConc* e *Sketch Engine* e suas ferramentas. Uma vez que nosso objetivo é estabelecer um método de extração para tais combinatórias, determinando qual ferramenta oferece melhores resultados, partimos da análise somente do NT *acervo*, uma vez que é um dos termos mais frequentes (3959 ocorrências) e representativos do nosso *corpus* de análise. Portanto, nossa metodologia consiste em identificar as UFE eventivas formadas a partir desse termo.

A seguir, ilustramos como foi feita a identificação e extração das unidades fraseológicas a partir das ferramentas acima referidas, para demonstrar os pontos positivos e as limitações de cada *software* na execução da tarefa proposta.

As ferramentas *Sketch Engine* e *AntConc*

Nesta sessão, apresentaremos as duas ferramentas a partir das quais proporemos metodologias para a extração e identificação de UFE eventivas: *Sketch Engine*[8] e *AntConc*[9].

SketchEngine

O *Sketch Engine* (SE) é uma ferramenta *on-line* que permite criar e fazer pesquisas em *corpora* textuais. Trata-se de um *software* pago, que oferece 30 dias de uso gratuito para testagem, e que exige conexão à internet para realizar as consultas. A ferramenta permite que se carregue um *corpus* textual de até 1 milhão de palavras (na versão gratuita), que pode ser lematizado automaticamente. Além do recurso de lematização automática, o SE também conta com um *corpus* de referência já inserido no sistema.

Para identificar e extrair UFE eventivas usamos o filtro *Word Sketch*, que consiste numa ferramenta que oferece, automaticamente e a partir de uma única busca, diferentes colocados do termo pesquisado e seus índices de frequência, separados por classes gramaticais e funções sintáticas. Portanto, o usuário escolhe as classes gramaticais e as funções sintáticas dos colocados do termo pesquisado. Por exemplo, para o termo *acervo* o usuário poderá recuperar estruturas em que o termo tem função sintática de objeto direto de um verbo. Assim, a ferramenta recupera estruturas como ***abrigar*** *acervo*, em que *abrigar* é o verbo selecionado por *acervo*. Nesse caso, o *Word Sketch* filtra, automaticamente, estruturas que podem conformar UFE do tipo $[NE]_V + [NT]_N$. Desse modo, de acordo com a busca escolhida pelo usuário, a ferramenta pode direcionar a determinadas estruturas de unidades fraseológicas (ver Fig. 1).

Adotamos os seguintes critérios para identificação das UFE eventivas:

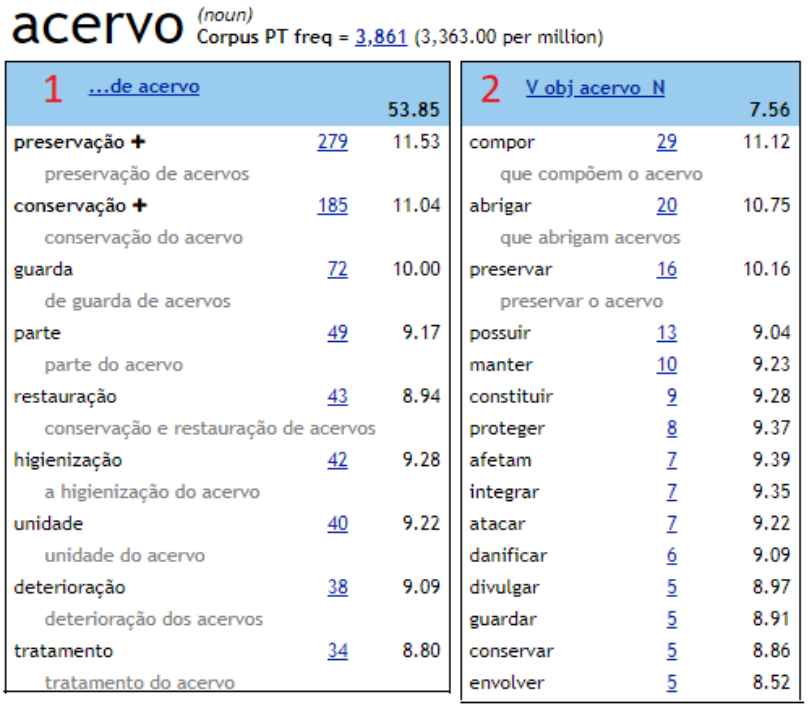
a) Critérios sintáticos:
   i. Partir do NT (base) para identificar os NE (colocados);
   ii. Identificar o colocado do sintagma preposicionado "de + NT", que será um **nome deverbal**, portanto, busca-se identificar combinatórias do tipo *conservação de acervo*;
   iii. Identificar a estrutura "V + NT", onde o colocado será um **verbo** cujo objeto direto é o NT pesquisado, que permite identificar combinatórias como *conservar o acervo*.

b) Critério quantitativo: o corte de frequência mínima para cada candidato a UFE eventiva deve ser igual ou superior a 10 ocorrências.

Exemplificamos abaixo o uso do filtro *Word Sketch* com o NT *acervo*:

[8] https://www.sketchengine.co.uk/
[9] http://www.laurenceanthony.net/software.html

Figura 1 – *Word Sketch.*



acervo *(noun)* Corpus PT freq = 3,861 (3,363.00 per million)

1 …de acervo		53.85
preservação +	279	11.53
preservação de acervos		
conservação +	185	11.04
conservação do acervo		
guarda	72	10.00
de guarda de acervos		
parte	49	9.17
parte do acervo		
restauração	43	8.94
conservação e restauração de acervos		
higienização	42	9.28
a higienização do acervo		
unidade	40	9.22
unidade do acervo		
deterioração	38	9.09
deterioração dos acervos		
tratamento	34	8.80
tratamento do acervo		

2 V obj acervo N		7.56
compor	29	11.12
que compõem o acervo		
abrigar	20	10.75
que abrigam acervos		
preservar	16	10.16
preservar o acervo		
possuir	13	9.04
manter	10	9.23
constituir	9	9.28
proteger	8	9.37
afetam	7	9.39
integrar	7	9.35
atacar	7	9.22
danificar	6	9.09
divulgar	5	8.97
guardar	5	8.91
conservar	5	8.86
envolver	5	8.52

Fonte: Adaptado de https://www.sketchengine.co.uk/

O método de busca 1 aponta para a estrutura $[NE]_{Ndev}$ + $[NT]_{SP}$, na qual o NE (colocado identificado pela ferramenta) é uma ação ou processo expressos por um nome deverbal (*conservação*), e o NT é um sintagma preposicionado (*do acervo*), como em *conservação do acervo*:

1) Sintagma preposicional **"de... acervo"** – oferece os colocados (geralmente substantivos e nomes deverbais) que formam sintagmas preposicionados com o termo pesquisado: *preservação do acervo, conservação do acervo*.

Já o método 2 indica unidades de estrutura: $[NE]_V$+ $[NT]_N$ (*abrigar acervo*):

2) "**V obj_acervo_N**" – oferece os verbos (V) cujo colocado é o termo pesquisado (N) com função sintática de objeto direto, formando combinatórias do tipo *compor acervo, abrigar acervo*.

Contudo, os resultados oferecidos pelo *Word Sketch* recuperam muito ruído por se tratar de uma ferramenta automática que oferece os colocados seguindo o critério de classe gramatical. Portanto, faz-se necessária uma segunda etapa de análise manual. Assim, após tomarmos os dados obtidos automaticamente pela ferramenta *Word Sketch*, confirmamos quais unidades conformam ações ou processos especializados da área. Para isso, verificamos as concordâncias (contextos) de cada um dos resultados. Essa etapa permitiu descartar combinações como *compor acervo* – cujo NE *compor* não se refere a uma ação ou processo especializado, mas a uma característica do acervo de ser composto por determinados tipos de

documentos –, e selecionar unidades como *abrigar acervo*, *preservar acervo*, etc., que são verbos eventivos próprios da área.

O mesmo trabalho manual é necessário para as estruturas em que o NE é uma nominalização. Dado que o filtro recupera os substantivos que aparecem junto do sintagma preposicionado *de acervo*, precisamos descartar unidades que são nomes, como ***unidade*** *do acervo*, e selecionar os NE que são nomes deverbais (nominalizações) como ***guarda*** *de acervo*.

Em suma, o *software* SE oferece, a partir de apenas uma busca, todas as estruturas candidatas a UFE eventivas. Essa totalidade dos dados se deve ao fato de que a ferramenta lematiza automaticamente os *corpora* nele inseridos reconhecendo todas as variações morfológicas da palavra pesquisada. Dessa forma, quando pesquisamos um NT como *acervo*, o SE gera resultados para as formas *acervo* e *acervos*, sem que sejam necessárias duas buscas diferentes. O mesmo ocorrerá com as unidades que acompanham o NT, ou seja, a ferramenta oferece todas as variações, como *preservação* ***de/do(s)/deste(s)/do(s)*** *acervo(s)* a partir de uma busca somente.

Outra característica que garante a totalidade dos dados é o *span* que o SE aplica em suas análises automáticas. Conforme Sinclair (1990), o *span* consiste na distância que há entre a base da colocação (neste trabalho, o NT) e seus colocados (NE). Nesse sentido, a ferramenta reconhece colocações mesmo quando os dois elementos apresentam entre eles outras unidades inseridas. Por exemplo, a UFE eventiva *preservação de acervos* pode ocorrer na estrutura *preservação de* ***seus*** *acervos*, ou *abrigar acervo* que ocorre no *corpus* também como *abrigar* ***seu rico*** *acervo*. Desse modo, embora as unidades da UFE eventiva não estejam imediatamente juntas, a ferramenta as reconhece e as contabiliza nos resultados finais da busca.

Cabe ressaltar que, muito embora a ferramenta *Word Sketch* ofereça todas as unidades candidatas a UFE eventivas a partir de uma única pesquisa, os resultados trazem muito ruído. Por exemplo, quando buscamos por estruturas nas quais o NT *acervo* forma um sintagma preposicionado (*de acervo*), a ferramenta não recupera somente os NE em forma de nominalizações, mas recupera também substantivos, como em *parte do acervo*, *unidade do acervo*, etc., o que reforça a necessidade da análise manual.

AntConc

O *AntConc* é um *software* gratuito, de livre acesso e que não exige conexão à rede para ser utilizado, posto que pode ser baixado e instalado em qualquer computador e mesmo em um *pen drive*. O programa permite diversos tipos de pesquisas por meio de seus filtros e índices estatísticos. Contudo, em contraponto ao SE, o *AntConc*, não oferece um *corpus* de referência nem tampouco lematiza automaticamente o *corpus* de estudo inserido nele[10].

Para identificar e extrair as UFE eventivas dos tipos $[NE]_{Ndev}$ + $[NT]_{SP}$ e $[NE]_{V}$+ $[NT]_{N}$ a partir do *AntConc*, utilizamos a ferramenta *Clusters/N-gramas*[11] para chegar às unidades candidatas seguindo os seguintes critérios:

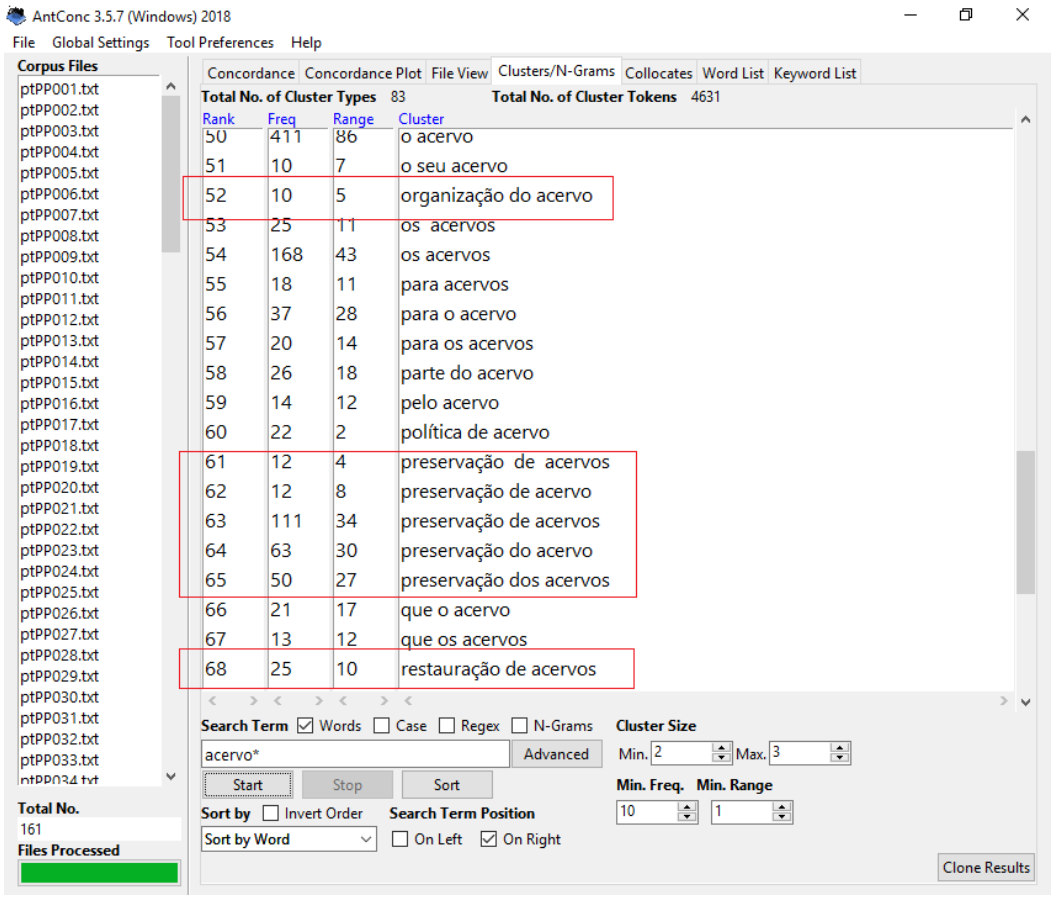
[10] Tanto a lematização quanto o *corpus* de referencia podem ser adicionados ao *software* como extensões. Contudo, o usuário deve criar as lematizações e o *corpus* de referência manualmente.

[11] Esta ferramenta lista as ocorrências de grupos de palavras combinadas (chamados na Linguística de Corpus de *clusters* ou *n-gramas*) repetidos ao longo de um texto ou *corpus* com indicação do seu número de frequência.

a) Partir do NT que poderá conformar a UFE eventiva;
b) Buscá-lo utilizando a forma truncada, por exemplo, *acervo**[12]
c) Recuperar *n-gramas* com um tamanho mínimo de 2 unidades e máximo de 3[13];
d) Buscar colocados em que o NT esteja à direita;
e) Estabelecer como corte de frequência mínima de cada *n-grama*10 ocorrências;

A partir desses critérios de busca, limitamos que o tamanho das unidades que queremos identificar esteja de acordo com as estruturas das UFE eventivas buscadas, e garantimos, através do filtro de frequência, que sejam unidades representativas da linguagem da área. A Fig. 2 ilustra a identificação de UFE eventivas partindo do NT 'acervo' com os critérios de busca aplicados.

Figura 2 – *AntConc*



Fonte: *AntConc 3.5.4*

[12] A busca da forma truncada no *AntConc* se dá pelo uso da *wildcard* "*", que significa zero ou mais caracteres. Com isso, garantimos que os resultados recuperados a partir da forma truncada serão todas as suas variações morfológicas (no caso de *acerv** serão *acervo*/*acervos*).

[13] Esse tamanho de cluster se deve ao tamanho das estruturas que estamos buscando identificar. As UFE eventivas do tipo $[NE]_V$ + $[NT]_N$, como *abrigar acervo*, apresenta duas unidades, já as UFE eventivas do tipo $[NE]_{Ndev}$ + $[NT]_{SP}$, como preservação de acervo, apresenta três unidades.

A partir da Fig. 2, observamos que, ao aplicar os critérios de busca, o NT *acervo* gera 83clusters, isto é, unidades que seguem o padrão dos critérios estabelecidos para os colocados do termo *acervo*.

Contudo, o *AntConc* reconhece todas as combinações possíveis entre palavras, independente de serem palavras gramaticais ou lexicais, isto é, o *software* considera como um *cluster* unidades que podem ser iniciadas por uma palavra gramatical. Por isso, nessa primeira etapa, o *AntConc* gera muitos resultados como *do acervo*, *de acervo*, *para o acervo* que não são relevantes para nossa pesquisa. Por outro lado, uma vez que *AntConc* não lematiza o *corpus* nele inserido, as variações morfossintáticas de uma mesma UFE serão apresentadas em resultados diferentes, como mostra a figura para *preservação* ***de*** *acervo*, *preservação* ***de*** *acervo****s***, *preservação* ***do*** *acervo*, *preservação* ***dos*** *acervo****s***.

Por isso, numa segunda etapa, os *clusters* formados por palavras gramaticais como *para o acervo* deverão ser eliminadas manualmente para que sejam filtradas somente as estruturas que constituam unidades fraseológicas como *organização do acervo*, *preservação do acervo* e *restauração do acervo*, por exemplo. Concomitantemente, para obter a frequência real das unidades, é necessário somar cada uma das variações morfossintáticas, como no caso da UFE eventiva *preservação de acervo* que terá uma frequência final de 280 ocorrências. Ainda assim, na ferramenta *clusters/n-gramas*, não é possível recuperar todas as estruturas morfossintáticas de determinada UFE em razão de não atenderem o critério de frequência igual ou maior a dez. É o caso, por exemplo, de *preservação desse(s) acervo(s)* que tem duas ocorrências no *corpus*.

Portanto, observamos que a limitação do *AntConc* no que se refere à falta da lematização gera dois problemas:

1) Exige do usuário o trabalho manual de somar as frequências de cada uma das variações morfossintáticas para obter a frequência real de cada unidade;
2) Gera silêncio na recuperação de determinadas unidades, isto é, certas unidades que não serão recuperadas porque suas variações morfossintáticas não estão dentro dos critérios estabelecidos para a busca, uma vez que a ferramenta não lematiza automaticamente o *corpus*.

O segundo problema é a principal limitação do *AntConc* para a tarefa de identificação e extração de UFE eventivas, pois a partir dele outros problemas surgem:

2.1) As frequências das unidades nunca são as frequências reais, posto que unidades que por ventura possuam frequência menor a 10 (critério estabelecidos) e que sejam uma variação morfossintática da UFE não serão recuperadas, por exemplo: *higienização do****s*** *acervo****s*** que possui 4 ocorrências e, portanto, não é recuperada pelos critérios de busca.

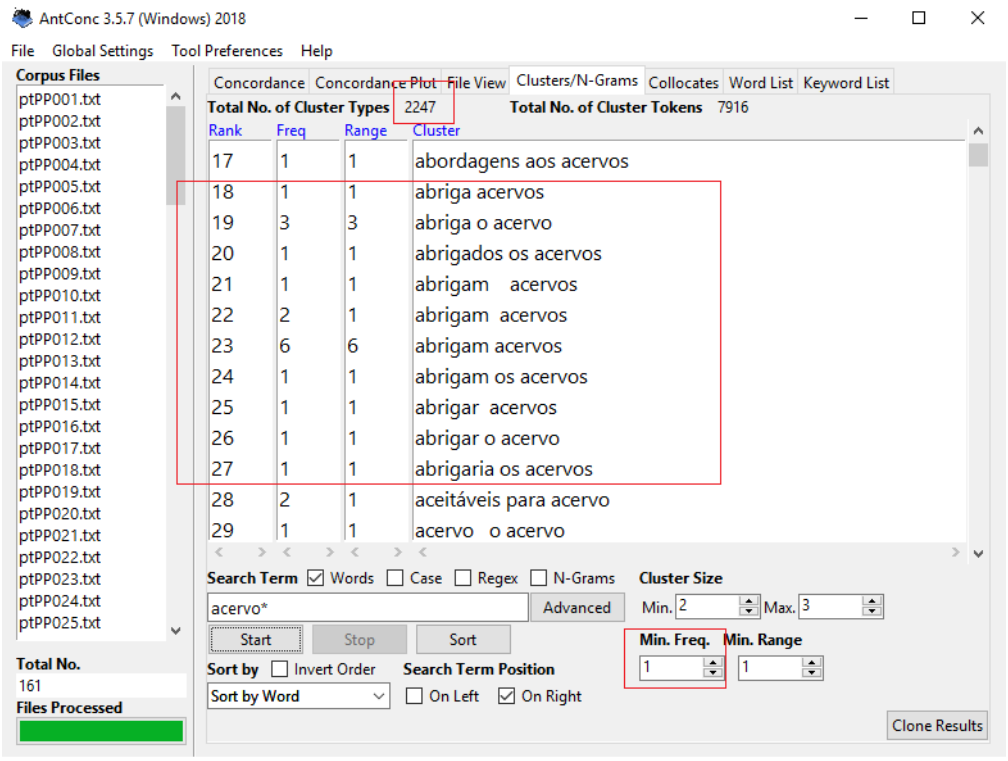
2.2) Em função do comentado em 2.1, algumas unidades sequer serão recuperadas, como no caso de *transporte de acervo*. Essa UFE eventiva apresenta como variações *transporte de acervos* (5 ocorrências) e *transporte do acervo* (5 ocorrências). Juntas, as variações morfossintáticas computam 10 ocorrências e, portanto, deveriam ser incluídas na base de dados. Entretanto, em função da limitação da ferramenta, essa unidade não é recuperada.

Uma possível solução para esse problema seria diminuir o critério de frequência, de modo que a ferramenta recuperasse estruturas como *higienização dos acervos*, que apresenta 4 ocorrências, ou *transporte de/do acervo*, que apresenta 5 ocorrências cada. No entanto, ao diminuir o critério de frequência para 1, isto é, se todas as variações morfossintáticas forem

recuperadas, o número de *clusters* formados a partir do termo *acervo* aumentará de 83 para 2247, tornando a análise manual impraticável.

2.3) Os problemas 2.1 e 2.2 estão diretamente ligados à quase impossibilidade de recuperar UFE eventivas formadas por $[NE]_V$+ $[NT]_N$ seguindo os critérios pré-estabelecidos. Isso ocorre uma vez que os verbos (NE) apresentarão variação morfológica (diferentes tempos e modos verbais), nem sempre computando cada um mais de 10 ocorrências. Por exemplo, na busca do NT *acervo*, seguindo os critérios, somente a forma *comp**õem** o acervo* foi recuperada, pois apresenta ocorrência igual a 10 (embora não seja uma UFE eventiva, como já mencionado, por não se tratar de uma ação ou processo especializado da área). Novamente, a solução para poder identificar UFE eventivas formadas por $[NE]_V$+ $[NT]_N$ seria diminuir a frequência para 1. Assim, podemos identificar todas as variações morfológicas dos NE que apresentam o NT *acervo* como objeto direto. A Fig.3 ilustra o caso.

Figura 3 – *AntConc*



Fonte: *AntConc 3.5.4*

Diminuindo o critério de frequência para 1, a ferramenta recupera as formas verbais que conformam a UFE eventiva *abrigar acervo*, com uma frequência de aparição no *corpus* igual a 10. Contudo, deverão ser analisados 2247 *clusters* para chegar a esses resultados, e, para cada unidade, as frequências deverão ser somadas manualmente.

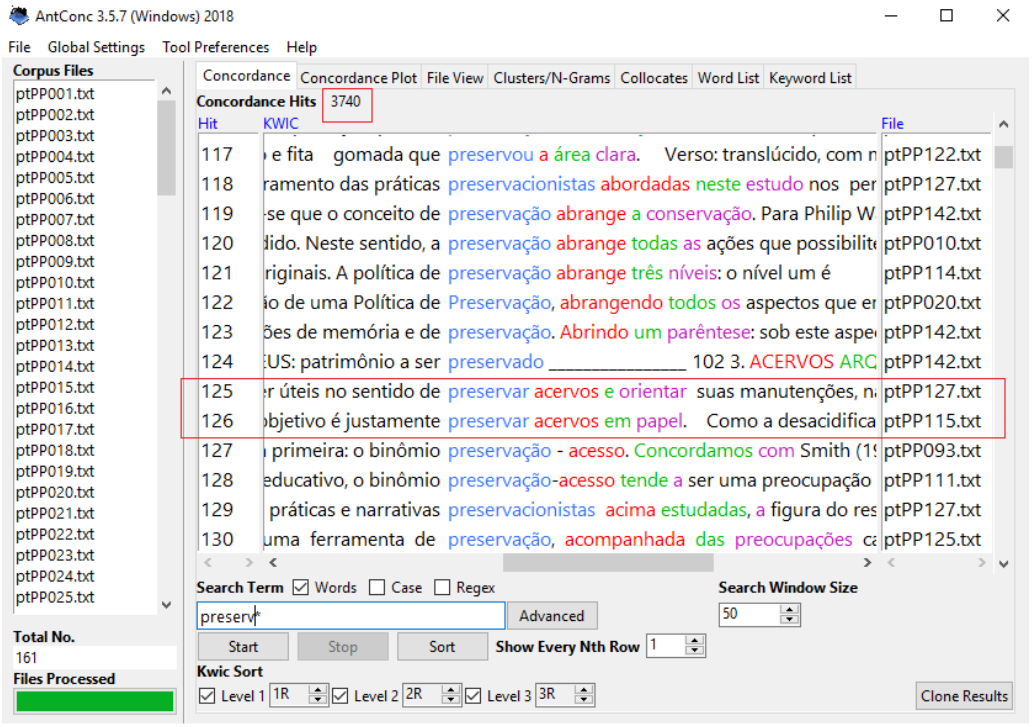
Outra possibilidade para a identificação de UFE eventivas de padrão morfossintático $[NE]_V$+ $[NT]_N$ seria buscar as unidades a partir das nominalizações. Assim, após identificar unidades como *preservação de acervo*, *transporte de acervo*, *guarda de acervo*, etc., que seguem o critério de frequência mínima 10, procura-se a UFE eventiva a partir do NE. Desse modo, procuramos na ferramenta *concordanciador* as formas truncadas dos verbos: *preserv**, *transport** e *guard**. No entanto, essa busca também é limitante, no sentido em que:

– somente recuperaremos UFE eventivas que apresentam uma correspondência morfossintática, como *preservação de acervo→preservar acervo*, e,

– a análise da frequência será manual, sendo necessário somar individualmente as estruturas para conferir se somam 10 ocorrências.

A Fig. 4 ilustra a busca a partir do NE *preservar:*

Figura 4 – *AntConc*

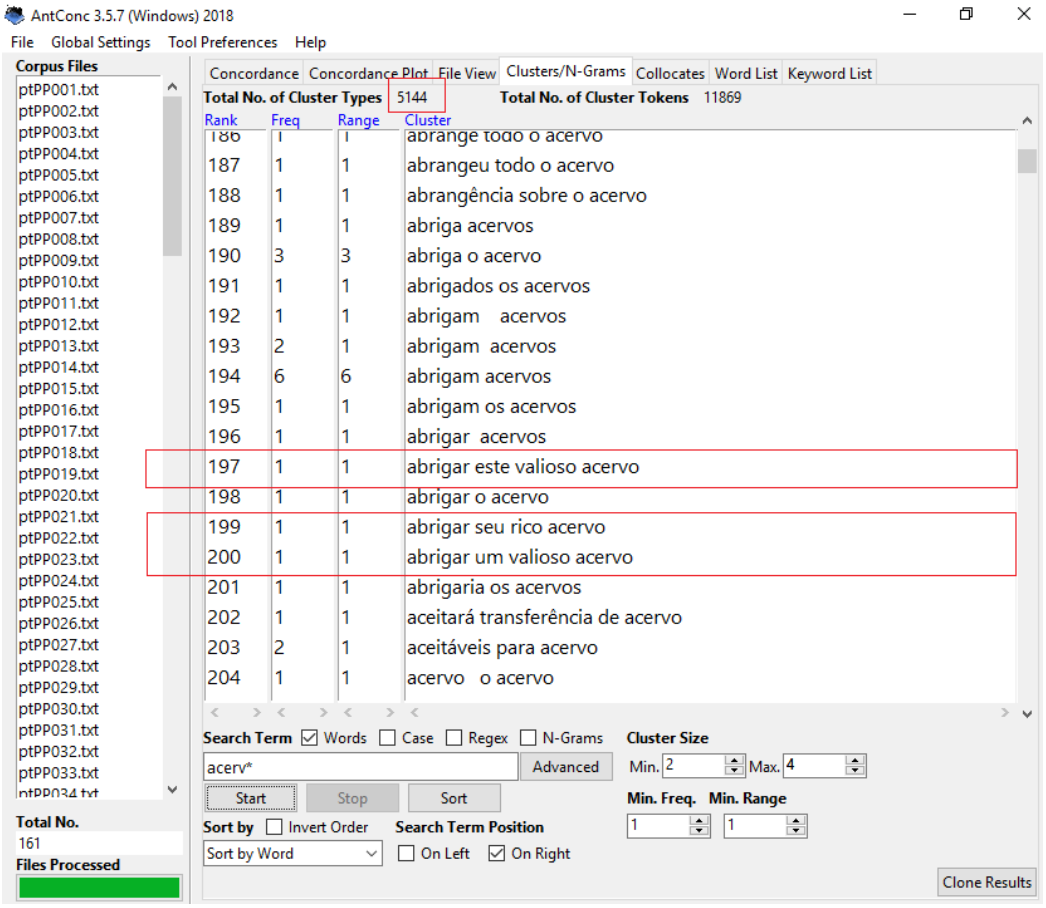


Fonte: *AntConc 3.5.4*

Embora seja possível identificar UFE eventivas de estrutura $[NE]_V$+ $[NT]_N$ (*preservar acervos*), novamente constatamos que a análise manual se torna impraticável, pois são 3740 concordâncias para analisar individualmente e chegar ao índice de frequência das unidades.

2.3) Finalmente, outro problema do *AntConc* é o *span*. Como vimos no SE, algumas unidades nem sempre apresentam seus itens imediatamente um ao lado do outro, gerando uma distância de outras palavras entre a base e o colocado (*span*). A única possibilidade de solucionar o problema do *span* no *AntConc* é aumentar o tamanho do *cluster*. Por exemplo, para recuperar unidades como *abrigar seu rico acervo* devemos buscar clusters de mínimo 2 unidades e máximo 4, com uma frequência mínima (1). A Fig. 5 ilustra essa possibilidade.

Figura 5 – *AntConc*



Fonte: *AntConc 3.5.4*

Com isso, identificamos unidades que apresentam outros itens intercalados, como *abrigar **este valioso** acervo*, *abrigar **seu rico** acervo*, *abrigar **um valioso** acervo*, que fazem parte da UFE eventiva *abrigar acervo*. Porém, a tarefa de identificação das unidades e a soma de suas frequências deverão ser feitas manualmente nos 5144 *clusters* que essa busca gera.

Desse modo, após analisarmos todas as possibilidades que o *AntConc* oferece, julgamos que os métodos mais razoáveis de usá-lo para identificar UFE eventivas seja:

1) Para identificar UFE eventivas de padrão morfossintático **[NE]$_{Ndev}$ + [NT]$_{SP}$**:
   a. Realizar a busca com a forma truncado do NT (como *acervo**);
   b. Recuperar *n-gramas* com um tamanho mínimo de 2 unidades e máximo de 3 em função das estruturas buscadas;
   c. Estabelecer como corte de frequência mínima de 10 ocorrências para cada unidade;
   d. Buscar os NE em forma de nominalização em que o NT esteja à direita (por exemplo, ***preservação** do acervo*);

2) Para reconhecer UFE eventivas de padrão morfossintático **[NE]$_{V}$+ [NT]$_{N}$**:
   a. Identificar as formas verbais das nominalizações extraídas na etapa anterior (por exemplo, *preservação→preservar*);

b. Na ferramenta *concordanciador*, buscar a forma truncada do verbo (por exemplo, *preserv**), identificar manualmente as unidades e somar suas frequências para garantir que tenham o mínimo de 10 ocorrências.

É importante esclarecer que o *AntConc* oferece outras possibilidades de busca. Por exemplo, para identificar unidades de tipo $[NE]_V$+ $[NT]_N$ poderíamos gerar uma busca como:

– *conserv* * * *acerv**. Essa busca gera 180 *clusters* (um número plausível de análise), a partir dos quais podemos identificar UFE eventivas como *conserva o acervo*. Contudo, o número de frequência ainda assim não seria o número real, pois outros itens podem estar intercalados na UFE eventiva, exigindo uma busca compreendendo um *span* maior.

Assim, seria necessário pesquisar:

– *conserv* * * * *acerv**, que gera 68 *clusters* e permite identificar UFE eventivas em concordâncias como ***conservar** o referido **acervo***. Porém, algumas UFE eventivas também estão registradas em *span* ainda maiores, como em ***conservar** e preservar um **acervo** não precisa ser uma atividade onerosa*. Para identificá-la, seria necessário buscar *conserv* * * * * *acerv**, que gera 122 *clusters*. Com isso, para obter a totalidade de frequência e confirmar se certas unidades conformam UFE eventivas do tipo $[NE]_V$+ $[NT]_N$ seriam necessários diversos tipos de busca.

Por isso, com o *AntConc* somente identificaremos UFE eventivas de padrão morfossintático $[NE]_V$ + $[NT]_N$ que apresentem correspondência com unidades de padrão $[NE]_{Ndev}$ + $[NT]_{SP}$, pois sua identificação partirá da nominalização. Infelizmente, a tarefa de identificar UFE eventivas de tipo $[NE]_V$ + $[NT]_N$ sem correspondência de verbo e nominalização se torna praticamente impossível pelo alto número de clusters para analisar (entre 4 e 6 mil, aproximadamente), como já explicado.

Na seção seguinte, apresentamos os resultados quantitativos e qualitativos a que chegamos a partir das metodologias adotadas para a extração de UFE eventivas nas ferramentas *AntConc* e *Sketch Engine*.

Resultados dos métodos de identificação e extração de UFE eventivas

Depois das etapas de identificação de UFE eventivas formadas a partir do termo *acervo* nas duas ferramentas, extraímos todas as unidades identificadas para duas planilhas de Excel para cada *software* (uma para os resultados do SE e outra para os resultados do *AntConc*). Organizamos os dados por estrutura morfossintática da UFE eventiva e por índice de frequência de cada unidade.

SketchEngine

A partir da ferramenta *Word Sketch* do SE, seguindo os critérios de busca estabelecidos, identificamos um total de 20 UFE eventivas, com uma frequência total de 896 ocorrências, de padrão morfossintático $[NE]_{Ndev}$ + $[NT]_{SP}$ formadas pelo NT *acervo*. A Tab. 1 ilustra os dados de frequência de cada unidade.

Tabela 1 – Resultado numérico das UFE eventivas[NE]$_{Ndev}$ + [NT]$_{SP}$ identificadas pelo SE

Sketch Engine	
[NE]$_{Ndev}$ + [NT]$_{SP}$	Frequência
1. *preservação de acervo*	279
2. *conservação de acervo*	185
3. *guarda de acervo*	72
4. *restauração de acervo*	43
5. *higienização de acervo*	42
6. *deterioração de acervo*	38
7. *tratamento de acervo*	34
8. *degradação de acervo*	25
9. *salvaguarda de acervo*	21
10. *uso de acervo*	20
11. *organização de acervo*	19
12. *manutenção de acervo*	19
13. *aquisição de acervo*	14
14. *diagnóstico de acervo*	14
15. *transporte de acervo*	13
16. *gestão de acervo*	13
17. *conservação-restauração de acervo*	12
18. *proteção de acervo*	12
19. *administração de acervo*	11
20. *perda de acervo*	10

Quanto às UFE eventivas de padrão morfossintático [NE]$_{V}$+ [NT]$_{N}$, identificamos 3 tipos de unidades, conforme ilustra a Tab. 2.

Tabela 2 – Resultado numérico das UFE eventivas[NE]V + [NT]N identificadas pelo SE

Sketch Engine	
[NE]V+ [NT]N	Frequência
1. *abrigar acervo*	20
2. *preservar acervo*	16
3. *manter acervo*	10

AntConc

A partir da pesquisa da forma truncada do termo *acerv**na ferramenta *AntConc* e seguindo os critérios de busca estabelecidos já citados, identificamos um total de 9 UFE

eventivas, com uma frequência total de 573 ocorrências, formadas por [NE]Ndev + [NT]SP. A Tab. 3 ilustra os dados.

Tabela 3 – Resultado numérico das UFE eventivas $[NE]_{Ndev}$ + $[NT]_{SP}$ identificadas pelo *AntConc*

AntConc	
$[NE]_{Ndev}$ + $[NT]_{SP}$	Frequência
1. *preservação de/do/dos acervo(s)*	248
2. *conservação de/do/dos acervo(s)*	148
3. *guarda de/do acervo(s)*	54
4. *restauração de/do acervo(s)*	36
5. *higienização do acervo*	27
6. *deterioração de/do/dos acervo(s)*	23
7. *uso de acervo*	14
8. *diagnóstico do acervo*	13
9. *organização do acervo*	10

O resultado numérico a que chegamos para cada unidade foi feito manualmente, somando as frequências de cada variação morfológica das UFE eventivas. Vale ressaltar que as unidades às quais chegamos são as que se encaixam nos critérios de busca estabelecidos (frequência mínima igual a 10 ocorrências e tamanho do *cluster* de 2 a 3 unidades). Com isso, algumas unidades com *span* maior não são identificadas, gerando silêncio na recuperação dos dados, como, por exemplo, em ***conservação do*** *referido* ***acervo***.

Quanto à identificação de UFE eventivas de padrão morfossintático **$[NE]_V$+ $[NT]_N$**, não foi possível fazer a análise de identificação pelo alto número de *clusters* que essa busca gera (como já referido e ilustrado anteriormente). Uma vez que o propósito deste trabalho é estabelecer uma metodologia para a identificação de UFE eventivas que conformarão uma base de dados que terá aproximadamente 400 termos, analisar de 100 a 5000 *clusters* para **cada** NT se torna inviável. Justificamos essa impossibilidade ilustrando, na Tab. 4, o número de *clusters* gerados a partir da busca de cada NE.

Tabela 4 – Número de clusters para análise na identificação de UFE eventivas $[NE]_V$ + $[NT]_N$

AntConc	
[NE]V+ [NT]N	*Clusters* para análise
*preserv**+ [NT]N	3740
*conserv**+ [NT]N	4836
*guard**+ [NT]N	850
*restaur**+ [NT]N	3284
*higieniz**+ [NT]N	515
*deterior**+ [NT]N	768
*us**+ [NT]N	2314
*diagnostic**+ [NT]N	36
*organiz**+ [NT]N	612

Cabe ressaltar que, caso analisássemos todos os *clusters* gerados a partir do NE, ainda assim não identificaríamos todas as unidades de padrão $[NE]_V$+ $[NT]_N$ formadas a partir do termo *acervo*, uma vez que estamos partindo dos NE identificados nas nominalizações das estruturas $[NE]_{Ndev}$ + $[NT]_{SP}$. Desse modo, um NE como *abrigar* em *abrigar acervo* (fig. 5) não seriam recuperados por não apresentarem forma nominalizada.

Observamos, assim, uma discrepância quantitativa relevante entre as diferentes UFE eventivas de estruturas morfossintáticas $[NE]_{Ndev}$ + $[NT]_{SP}$ e $[NE]_V$+ $[NT]_N$ identificadas pelo SE e pelo *AntConc*, bem como uma grande diferença nos índices de frequência de cada unidade. A Tab. 5 compara os dados finais.

Tabela 5 – Resultado quantitativo final das UFE eventivas identificadas pelo SE e *AntConc*.

Sketch Engine			*AntConc*		
Padrão morfossintático	UFE eventivas identificas	Frequência total	Padrão morfossintático	UFE eventivas identificas	Frequência total
$[NE]_V$ + $[NT]_N$	3	46	$[NE]_V$ + $[NT]_N$	0[14]	-
$[NE]_{Ndev}$ + $[NT]_{SP}$	20	896	$[NE]_{Ndev}$ + $[NT]_{SP}$	9	573

Pela análise dos dados e dos métodos de identificação e extração de UFE eventivas a partir das duas ferramentas, julgamos que essa diferença nos resultados se deva aos fatos seguintes:

– o SE reconhece *span* maiores, recuperando unidades fraseológicas mesmo com outros itens inseridos, como em *degradação **contínua** dos acervos* ou *organização **física** de **seus** acervos;*

– o SE recupera unidades que apresentam variação morfossintática sob um mesmo resultado, uma vez que lematiza o *corpus*;

– os resultados que se obtêm com o *AntConc* são limitados em função dos critérios de busca estabelecidos na ferramenta *clusters/n-gramas*, que permitem uma análise manual plausível. Caso reduzíssemos o critério de frequência para 1 e, por conseguinte, aumentássemos enormemente o número de *clusters* para identificar as variações morfossintáticas dos verbos, e se aumentássemos o tamanho do *cluster* para mais de 3 itens a fim de identificar *span* maiores, obteríamos os mesmos dados que o *Word Sketch* oferece.

Considerações finais

O propósito inicial deste trabalho foi traçar uma metodologia de identificação de dois tipos de UFE eventivas, $[NE]_V$ + $[NT]_N$ e $[NE]_{Ndev}$ + $[NT]_{SP}$, por meio das ferramentas SE e *AntConc* a fim de compará-las. A partir dessa comparação, separamos nossas conclusões em pontos positivos e pontos negativos de cada programa.

[14] Não estamos considerando a unidade fraseológica *compor acervo* por não conformar uma UFE eventiva (ver seção 2.3).

Sketch Engine

Pontos positivos:

– lematiza automaticamente o *corpus* inserido. Com isso, todas as variações morfológicas de uma palavra são recuperadas, uma vez que são reconhecidas sob a forma de lema;

– recupera, a partir de uma única busca do NT, todos os colocados (em todas as suas variações morfossintáticas em função da lematização) que serão candidatos a conformar UFE eventivas

– reconhece unidades fraseológicas que apresentam outros itens inseridos entre o NE e NT, isto é, reconhece as unidades fraseológicas mesmo com *span*.

Pontos negativos:

– é um *software* pago, cuja licença aumenta de valor conforme o tamanho do *corpus* inserido também aumenta. Portanto, quanto mais dados para analisar, mais cara é a assinatura;

– exige conexão à rede de internet para realizar as buscas;

– por lematizar o *corpus*, recupera ruído nas buscas de colocados por função sintática, seja por recuperar diferentes classes gramaticais, seja por confundir em alguns casos a lematização.

AntConc

Pontos positivos:

– é um *software* gratuito, de livre acesso e que não exige conexão à rede de internet para ser usado;

– oferece resultados com menos ruído quando a unidade pesquisada se encaixa nos critérios de frequência e de tamanho de *cluster* para a busca;

– sua ferramenta *concordanciador* permite ordenar os resultados em ordem alfabética, o que pode ajudar na identificação de certas unidades (ver Fig. 5), e iluminar as palavras em contexto à esquerda e à direita da palavra pesquisada (*kwic – keywords in context*).

Pontos negativos:

– a não lematização do *corpus* para as buscas de *clusters/n-gramas* gera um alto índice de silêncio nos resultados, isto é, muitas unidades não são recuperadas, pois o *software* não reconhece as variações morfológicas de uma palavra (no caso, do NE) sob um mesmo lema;

– para a identificação de cada estrutura morfossintática de UFE eventiva é necessário um tipo diferente de busca, isto é, com diferentes critérios.

– em função da não lematização, para recuperar a totalidade dos dados de cada unidade (isto é, reconhecendo as variações morfológicas), é necessário fazer mais de um tipo de busca com o uso das *wildcards* (*), ou diminuindo o critério de frequência para 1 e aumentando o tamanho dos *clusters.* Em ambos os casos, essas buscas gerarão um grande número de unidades para análise;

– pelos problemas citados acima, raramente será possível identificar e extrair de forma eficiente a totalidade dos dados do *corpus*, sendo muito provável a não identificação de determinadas unidades.

As conclusões positivas e negativas às quais chegamos reforçam a necessidade da análise manual dos dados por parte do linguista, uma vez que a pesquisa feita com a Linguística de Corpus tem sempre um caráter semiautomático, independentemente do *software* utilizado. Dessa forma, nenhuma ferramenta por si só é perfeita, nem tampouco cumpre com todas as necessidades do pesquisador.

No que se refere à análise qualitativa dos dados extraídos, acreditamos que eles reforçam os critérios de caracterização e reconhecimento das UFE eventivas no que se refere:

– a sua conformação estável por um NE e um NT, apesar da sua variação morfossintática;

– ao NE, que parece ser mais produtivo na sua realização como nominalização, o que se justifica pela maior frequência de UFE eventivas de padrão morfossintático $[NE]_{Ndev}$ + $[NT]_{SP}$ que as de padrão $[NE]_{V}$+ $[NT]_{N}$;

– à correspondência morfossintática que nem sempre ocorre entre os diferentes tipos de UFE eventivas, de modo que alguns NE se realizam somente como nominalizações, e outros somente como verbos (por exemplo, ***abrigar*** *acervo* que não tem a forma nominalizada correspondente ao verbo no *corpus*, ou ***aquisição*** *de acervo*, que não tem a ocorrência da forma verbal correspondente no *corpus*).

Por fim, acreditamos que o trabalho permitiu um avanço no estabelecimento de metodologias para o reconhecimento e extração de UFE eventivas em *corpora* textuais por meio de ferramentas computacionais. Assim, o trabalho propõe como direcionamentos futuros a testagem com um maior número de NT, e a aplicação das metodologias propostas para a identificação de UFE eventivas equivalentes nas línguas estrangeiras incluídas no projeto.

REFERÊNCIAS

BEVILACQUA, C.R. *Unidades Fraseológicas Especializadas Eventivas: descripción y reglas de formación en el ámbito de la energía solar*. (Tese de Doutorado) – Instituto Universitário de Lingüística Aplicada (IULA), Universidade Pompeu de Fabra, Barcelona. 2004.

BOURIGAULT, D; SLODZIAN, M. *Por uma terminologia textual*. Cadernos de Tradução, nº 17 – out/dez 2004. p. 101-108.

CABRÉ, M. T.*Terminologia y linguística: la teoría de las puertas*. Estudios de lingüística del español, V. 16. 2002.

CABRÉ, M.T.; LORENTE,M.; ESTOPÀ, R. *Terminología y Fraseología*. In: Actas del V Simposio de Terminología Iberoamericana, 1996.

LORENTE, M; BEVILACQUA, R; ESTOPÀ, R. *El análisis de la fraseología especializada mediante elementos de la lingüística actual.*In: CORREA, M (Org.). *Terminologia, Desenvolvimento e Identidade Nacional*. VI Simpósio Ibero-americano de Terminologia. Lisboa: Colibri, 2002. p. 647-665.

SINCLAIR, J. Collocation. In: SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford/New York: Oxford University Press, 1991. pp. 109-121.

Recebido em 30/06/2018
Aprovado em 28/08/2018
Publicado em 20/11/2018