# WORD SKETCH COMO FERRAMENTA PARA EXTRAÇÃO DE COLOCAÇÕES

## *WORD SKETCH AS A TOOL TO EXTRACT COLLOCATIONS*

Manuela ARCOS[1]
Marine Laísa MATTE[2]

**Resumo**: Neste trabalho, descrevemos métodos de identificação e extração de colocações em *corpora* textuais de língua geral, produzidas por aprendizes de língua inglesa em textos acadêmicos, e de colocações especializadas da área da Conservação e Restauração de Patrimônio Cultural por meio da ferramenta Word Sketch (WS), do *software* Sketch Engine. Ao entendermos colocações como palavras que frequentemente ocorrem juntas em função do seu grau de atração semântica, o objetivo deste trabalho é demonstrar como a ferramenta WS permite a identificação e extração de colocações de uma forma semiautomática, uma vez que, após a extração, é papel do pesquisador levar em conta os demais critérios constituintes de uma colocação. Sejam de língua geral ou de língua de especialidade, as colocações são unidades constituídas por critérios sintático-semânticos, pragmáticos e discursivos. Como aporte teórico-metodológico, apoiamo-nos na Linguística de Corpus e buscamos estabelecer critérios para a extração de colocações através da ferramenta WS. Nossos resultados indicam que a ferramenta WS é eficaz para a tarefa de extração de colocações tanto de escrita acadêmica como de linguagem especializada, pois permite que a identificação das unidades parta de seus critérios de constituição.

**Palavras-chave:** Colocações. Linguística de Corpus. Word Sketch.

**Abstract**: In this paper we describe methods to identify and extract collocations in general language (GL) corpora written by Brazilians in academic English texts, and collocations in specialized language (SL) from Conservation and Restoration of Cultural Heritage through the use of Word Sketch (WS) tool from the software Sketch Engine. Based on our understanding of collocation as words that frequently occur together due to their semantic attraction, the goal of this paper is to demonstrate how the WS tool allows for the identification and extraction of collocations in a semi-automatic way, due to the fact that after the extraction it is the researcher's role to take into consideration further constituent criteria. Whether in GL or SL, collocations are units made of syntactic, semantic, pragmatic and discursive criteria. We rely on Corpus Linguistics as both the theoretical background and methodology, and we try to establish criteria to identify collocations through the WS tool. Our results indicate that the WS tool is useful for extracting collocations from both academic writing and specialized language, as the identification of the units can be based on their constituent criteria.

**Keywords***:* Collocations. Corpus Linguistics. Word Sketch.

1 Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil; arcosmanuela@gmail.com; https://orcid.org/

2 Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil; marine.laisa@gmail.com; https://orcid.org/

## Introdução

O presente artigo é fruto de duas pesquisas de mestrado sobre identificação e extração de colocações de língua geral (LG) (MATTE, 2019) e de língua de especialidade (LE) (ARCOS, 2019) através da ferramenta Word Sketch (WS) do *software* Sketch Engine. Considerando que a identificação e a extração contaram com a mesma ferramenta de análise em ambas as pesquisas, este trabalho tem como objetivo a descrição de tal ferramenta a fim de ilustrar quais métodos são esses e quão produtiva é a ferramenta para o estudo de colocações em *corpora* textuais.

Na pesquisa de identificação e extração de colocações de LG, empreendeu-se uma análise quanti- e qualitativa acerca de como alunos brasileiros estudando em universidades britânicas utilizam colocações acadêmicas em inglês em seus trabalhos acadêmicos. Dessa maneira, os colocados de 125 nódulos foram identificados através da ferramenta WS e posteriormente analisados. Os resultados, além de informarem quantitativa e qualitativamente os usos de colocações acadêmicas em inglês por brasileiros, também fornecem subsídios para o ensino de Inglês para Fins Acadêmicos.

Na pesquisa de identificação de colocações especializadas, foram identificadas e extraídas colocações a partir de 65 núcleos terminológicos da área da Conservação e Restauração de bens materiais em suporte papel. O trabalho se insere dentro de um projeto de pesquisa maior do grupo Termisul (Instituto de Letras, UFRGS), intitulado "A linguagem do Patrimônio Cultural Brasileiro: conservação dos bens culturais móveis" com ênfase na terminologia da conservação de acervos documentais, bibliográficos e arquivísticos gráficos em suporte papel. O objetivo do projeto é compilar uma base de dados terminológica *on-line* multilíngue sobre Patrimônio Cultural voltada para tradutores, revisores e redatores técnicos. A pesquisa, cuja metodologia apresentamos aqui, teve como objetivo identificar e analisar as combinatórias especializadas em língua portuguesa da área da Conservação e Restauração a fim de compor um dos campos das fichas terminológicas da base de dados *on-line*, juntamente com os termos simples e sintagmáticos em português e seus equivalentes em cinco línguas (espanhol, francês, inglês, italiano e russo). A partir da pesquisa, identificamos unidades que expressam processos especializados do âmbito de conhecimento (Conservação e Restauração) que se realizam com os núcleos terminológicos, como *acondicionar acervo/acondicionamento de acervo* ou *planificar papel/planificação de papel,* em que *acondicionar/acondicionamento*, *planificar/planificação* são os processos e *acervo/papel*, os núcleos terminológicos.

Os procedimentos metodológicos com o WS utilizados em ambas as pesquisas são especificados na metodologia e são o cerne do presente trabalho. Para propor tais

procedimentos, primeiramente explicitamos o que é entendido por colocação na perspectiva da LG, seguida pelos estudos da área sobre colocações de LE. Em seguida, são elencados o aporte teórico-metodológico adotado no trabalho bem como os *corpora* analisados. Na seção de metodologia, os procedimentos metodológicos de identificação e extração das colocações são descritos e, na seção seguinte, são apresentados os resultados quantitativos. Por fim, concluímos o estudo com as considerações finais.

## Colocações pela perspectiva da língua geral

Os estudos das colocações foram inaugurados por Firth (1957, p. 11) com sua célebre frase "you shall know a word by the company it keeps". A partir de então, nos estudos da área, as colocações apresentaram uma grande variedade conceitual e denominativa. Tomando como ponto de partida o fato de a língua ser formulaica por natureza, isto é, ser formada por estruturas fixas, recorrentes e com certa estabilidade, o uso adequado e competente de sequências formulaicas garante formulaicidade e confere naturalidade aos usos da língua (DURRANT; SCHMITT, 2009). Dentre tais sequências formulaicas, tem-se as colocações, as quais ganham um lugar de destaque na presente análise.

Na perspectiva da LG, Sinclair (1991) entende-as a partir da probabilidade de duas ou mais palavras ocorrerem juntas. Wray (2000) vincula as colocações sob o termo guarda-chuva "sequência formulaica", que se refere a uma sequência de palavras pré-fabricada, armazenada conjuntamente na memória e empregada com um significado particular. A compreensão de que há sequências fixas de palavras também é defendida por Nesselhauf (2005), para quem uma colocação é composta por duas ou mais palavras com fixidez lexical e sintática. Hill (2000), ao definir colocações com base no critério de frequência, destaca que elas são combinações multi-palavras que constituem grande parte de um texto. A combinação de duas ou mais palavras de forma não aleatória é reconhecida por Hyland (2006) como colocação.

Além disso, cumpre destacar que uma das motivações para a presente pesquisa é o trabalho de Frankenberg-Garcia (2018), no qual foi realizada uma seleção prévia de candidatos a nódulos de colocações, dentre os quais 125 são substantivos, 38 são verbos e 24 são adjetivos. Para este estudo, os 125 substantivos serviram de base para a extração e identificação de colocações. Para os nódulos que são substantivos, os colocados são de três categorias distintas: modificadores, ou seja, adjetivos que antecedem o substantivo (*difficult + task* e *advanced + technique*), verbos que o acompanham quando é o sujeito da frase (*task + require* e *technique + use*) e verbos que o acompanham quando é o objeto da frase (*execute + task* e *apply + technique*). Nesses exemplos, os nódulos das colocações são *task* e *technique* e os colocados são *difficult, advanced, require, use, execute* e *apply*.

Dito isso, com base nos estudos da área e na variedade conceitual, operamos com a seguinte definição para a presente análise das colocações da LG: *combinação de duas palavras associadas por probabilidades estatísticas de ocorrerem juntas.* A palavra principal da colocação é chamada de nódulo e as associadas a ela são os colocados. Assim, para os fins deste trabalho, a estrutura básica de uma colocação na LG é nódulo + colocado, sendo o nódulo necessariamente um substantivo e o colocado podendo ser de três categorias distintas, conforme classificação de Frankenberg-Garcia (2018), anteriormente mencionada.

## Colocações pela perspectiva da língua de especialidade

Os estudos sobre as unidades sintagmáticas sob a ótica das linguagens de especialidade se desenvolveram principalmente a partir do início dos anos 90. Nos estudos do Léxico, mais especificamente dentro do quadro dos estudos da Fraseologia em interface com os estudos da Terminologia, diferentes autores introduziram denominações diversas para as unidades sintagmáticas de valor especializado embora coincidindo em aspectos conceituais e estruturais. Kjaer (1990) as denomina *fraseologias* e as define como combinações idiossincráticas de palavras que constituem o "ambiente" em que os termos se combinam, por exemplo, *a corrente passa*, em que o elemento *corrente* é o termo. Picht (1990) denomina esse tipo de unidade como *LSP phrase*[3] como uma proposição composta por, no mínimo, dois elementos, em que um deles possui, obrigatoriamente, caráter verbal, enquanto o outro pode ter função sintática de objeto ou de sujeito. São exemplos de *LSP phrases passar um cheque* ou *apertar um parafuso*. Pavel (1993) define as unidades sintagmáticas como *Unidades Fraseológicas* (UF), compostas por uma unidade terminológica (UT) decorrente de uma estrutura conceitual coerente. As UT são consideradas como núcleos de coocorrência usuais nos textos de especialidade e constituem UF quando se manifestam pela combinação de: UT + substantivo (*agregado de células*); UT + adjetivo (*agregado bidimensional*) ou UT + verbo (*absorver um agregado*).

Em propostas mais atuais, L'Homme (2004) descreve as colocações – *combinatórias léxicas especializadas* (CLE), na sua denominação – como "um conjunto de unidades lexicais com as quais os termos se combinam de maneira privilegiada nas frases". Essa combinação de termos (unidades nominais) com outras unidades lexicais não ocorre de forma aleatória, mas em função de afinidades semânticas e de preferências de uso em certos domínios especializados. As combinações podem ser do tipo verbo + nome (*administrar um medicamento*); nome + adjetivo (*prognóstico sombrio*) ou nome + nome (*execução de um programa*).

3 LSP: *language for specific purposes.*

Bevilacqua *et al*. (2012), a partir de um estudo anterior sobre unidades fraseológicas (BEVILACQUA, 2004) e com base na proposta de L'Homme (2004), definem as CLEs como unidades sintagmáticas de uso recorrente em situações comunicativas das áreas temáticas que revelam uma preferência marcante por certas especificidades e convenções próprias do idioma, da área e/ou do gênero textual em que ocorrem. Desse modo, são unidades que resultam de uma seleção restritiva condicionada ao modo de dizer característico de cada âmbito do conhecimento" (BEVILACQUA *et al*., 2012, p. 242).

Em sua proposta, na qual nos baseamos para identificar as colocações especializadas desta pesquisa, Bevilacqua (2012) descreve critérios linguísticos (sintático-semânticos), critérios pragmático-discursivos e critérios quantitativos como constituintes das CLE. O critério linguístico (sintático-semântico) refere-se a sua estrutura morfossintática, que será formada por um núcleo eventivo (NE) de caráter verbal que expressa ações e processos especializados da área de domínio, e por um núcleo terminológico (NT) de caráter nominal que constitui um nó de conhecimento na estrutura conceitual da área. Duas realizações morfossintáticas dessas unidades são possíveis[4]:

1. [NE]V + [NT]N, em que o NE é um verbo e o NT é um substantivo com valor de objeto direto, por exemplo, em *proteger acervo*;

2. $[NE]_{Ndev}$ + $[NT]_{SP}$ , em que o NE assume a forma de um nome deverbal (nominalização) e o NT constitui um sintagma preposicional seguido de um substantivo, como em *proteção de acervo.*

O critério pragmático-discursivo refere-se às especificidades do texto, que incluem não só sua temática, mas também a função comunicativa que desempenham em seu contexto de uso, ou seja, indicar processos e ações relacionados à área de domínio (no caso da pesquisa aqui apresentada, o âmbito da Conservação e Restauração de bens culturais móveis em papel). Já o critério quantitativo refere-se à frequência relevante de aparição das combinatórias no texto especializado, um traço típico dessas estruturas. O critério quantitativo auxilia a identificar como unidades típicas de uma área aquelas que ocorrem significativamente[5] – em termos estatísticos – no *corpus* textual.

---

4 Bevilacqua *et al.* (2012) propõem outras estruturas morfossintáticas de CLE. No entanto, por questões de espaço e pertinência, apresentamos neste trabalho somente as duas estruturas de CLEs que identificamos ao longo da pesquisa.

5 Há um amplo debate sobre significância estatística no âmbito da Linguística de Corpus. Nas duas pesquisas aqui apresentadas, o que julgamos ser estatisticamente significativo tem relação direta com os tamanhos dos *corpora* de estudo. Tanto o *corpus* de LG quanto o *corpus* de LE que apresentamos possuem aproximadamente 1 milhão de palavras. Dessa forma, em ambas as pesquisas foi estabelecido como critério de frequência significativa um mínimo de 10 ocorrências no *corpus* de estudo com *range* (distribuição) em dois ou mais textos.

## Linguística de Corpus e os *corpora* de estudo

Como aporte teórico-metodológico, optamos pela Linguística de Corpus (LC), uma área de estudo que lida com linguagem autêntica (BIBER; DOUGLAS; CONRAD; REPPEN, 1998; MCENERY; WILSON, 1996; MCENERY; HARDIE, 2011; BERBER SARDINHA, 2000), isto é, usos reais de produção linguística armazenados em conjuntos de textos denominados *corpus* (ou *corpora*, no plural). Adotando uma perspectiva probabilística, a LC encara a linguagem pelo viés da probabilidade de determinados elementos linguísticos ocorrerem em certos contextos em detrimento de outros. Além disso, análises baseadas em *corpus*

- são empíricas, analisando padrões reais de uso em textos naturais;
- utilizam uma grande coleção de textos naturais (*corpus*) como base;
- fazem uso extensivo de computador, utilizando tanto técnicas automáticas quanto interativas;
- dependem de técnicas analíticas quantitativas e qualitativas (CONRAD; REPPEN, 1998, p. 4, tradução nossa[6]).

Neste trabalho, as colocações de LG decorrem da análise de *corpora* de inglês acadêmico, a saber, o *Brazilian Academic Written English* (BrAWE - SILVA, 2017)[7] na comparação com o *British Academic Written English*[8] (BAWE - ALSOP; NESI, 2009) com o objetivo de verificar as diferenças de usos de colocações na escrita acadêmica de alunos brasileiros e de alunos com elevado destaque acadêmico que estão representados no *corpus* BAWE. Há uma diferença considerável no tamanho dos *corpora* (ver Tabela 1), porém ambos representam as mesmas áreas do conhecimento: *Life Sciences, Social Sciences, Arts and Humanities* e *Physical Sciences*[9]. Além disso, os dois *corpora* contêm textos de 13 gêneros acadêmicos, conforme classificação de Gardner e Nesi (2012), elencados a seguir:

---

6 No original:
"- it is empirical, analyzing the actual patterns of use in natural texts;
- it utilizes a large and principled collection of natural texts (corpus), as the bases for analysis;
- it makes extensive use of computer for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques".

7 Para mais informações acerca do *corpus* BrAWE, sugerimos a leitura de Silva (2017).

8 A descrição do projeto no qual o *corpus* BAWE se insere está disponível em: http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf

9 Em uma aproximação com a classificação da CAPES, essas áreas do conhecimento podem ser entendidas como "Ciências Biológicas", "Ciências Sociais", "Ciências Humanas" e "Ciências Exatas e da Terra", incluindo as Engenharias neste último.

- Case Study
- Critique
- Design Specification
- Explanation
- Exercise
- Essay
- Empathy Writing
- Literature Survey
- Methodology Recount
- Narrative recount
- Problem question
- Proposal
- Research Report

**Tabela 1.** *Corpora* BAWE e BrAWE

| | **BAWE** | **BrAWE** |
|---|---|---|
| **Número de palavras** | 3.312.196 | 768.323 |
| **Número de textos** | 2.761 | 380 |

**Fonte:** Elaboração própria

Já as colocações especializadas deste trabalho decorrem da análise de um *corpus* da área da Conservação e Restauração de bens materiais em suporte papel, o *Corpus* Papel, elaborado pelo grupo de pesquisa Termisul[10]. O *corpus*[11] foi compilado seguindo critérios bem definidos[12], pelos quais os textos deviam:

a) conter as palavras-chave *documento, documentação, conservação, papel, patrimônio, preservação, restauração e restauro*, entre outras;

b) pertencer a 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.

10 O TERMISUL, Grupo Terminológico do Cone Sul, é um grupo de pesquisa de Terminologia numa perspectiva multilíngue do Instituto de Letras da Universidade Federal do Rio Grande do Sul (UFRGS) (www.ufrgs.br/termisul).

11 Neste trabalho descrevemos o *corpus* de língua portuguesa, pois as colocações especializadas que identificamos partiram dos NT em português. No entanto, os mesmos critérios de construção de *corpus* foram replicados para as demais línguas (espanhol, francês, italiano, inglês e russo) para a identificação de equivalentes.

12 Para mais informações sobre compilação de *corpus* especializado, sugerimos a leitura de Pearson (1998).

O *Corpus* Papel conta com 161 textos e, aproximadamente, 38.129 *types* e 967.852 *tokens*[13].

**Tabela 2.** *Corpus* Papel

| | ***Corpus* Papel** |
|---|---|
| **Número de palavras** | 967.852 |
| **Número de textos** | 161 |

**Fonte:** Elaboração própria

Na primeira fase da pesquisa, o grupo Termisul identificou, a partir do *Corpus* Papel, 65 termos da área para compor as entradas da base de dados terminológica *on-line* multilíngue[14]. Neste trabalho, esses termos constituem os NT a partir dos quais identificamos as colocações especializadas.

Apesar de a LG e a LE terem suas especificidades, há também pontos de intersecção no que tange ao estudo das colocações. Dado que a variedade denominativa e conceitual permeia os estudos da área, operamos com a mesma nomenclatura – "colocações" – para ambas as perspectivas, sendo este o primeiro ponto de intersecção. A presença de um núcleo nas colocações é também um elemento em comum entre as abordagens de análise da LG e da LE, sendo que, no caso da análise de LG empreendida aqui, o núcleo é um substantivo e, no caso da LE, o núcleo é um termo. O núcleo da colocação é o ponto inicial para a busca na ferramenta WS, uma vez que é a partir dele que a identificação dos colocados produtivos é feita.

Na seção seguinte, apresentamos a ferramenta WS, seus filtros e índices matemáticos, e explicamos de que forma a utilizamos para a identificação e extração de colocações de LG e colocações especializadas.

## Metodologia

Para a identificação e extração das colocações da LG e da LE, utilizamos a ferramenta Word Sketch (WS) do *software* Sketch Engine (SE) (KILGARIFF *et al.*, 2004). O SE oferece ferramentas variadas para conduzir análises de LC, porém o WS é particularmente útil e ideal para o tipo de pesquisa empreendida aqui pelo fato de apresentar o comportamento

13 O conceito de *types* se refere ao número de palavras diferentes que ocorrem em um *corpus* e *tokens,* ao número total de palavras.

14 Atualmente, a base de dados elaborada pelo grupo Termisul conta com mais de 300 termos-entrada em língua portuguesa e seus equivalentes em espanhol, francês, inglês, italiano e russo.

gramatical e colocacional de uma palavra. Ou seja, por meio do WS os colocados – palavras que se combinam com a palavra principal da colocação (núcleo) – são organizados de acordo com critérios sintáticos de diferentes tipos, os quais foram levados em conta na extração das colocações do presente estudo. Além disso, os colocados podem ser acompanhados por valores estatísticos[15] que descrevem o grau de atração entre as palavras. As medidas de associação mais comuns são a Informação Mútua (MI)[16] (*Mutual Information*, em inglês), que indica o grau de atração entre as palavras, e o logDice, que leva em consideração o índice de frequência com que nódulo e colocados co-ocorrem[17]. Sendo assim, o WS é uma ferramenta adequada para identificar e extrair colocações tanto da LG quanto da LE por permitir que os critérios de constituição das colocações sejam tomados como critérios de busca para sua identificação. Cumpre destacar, também, que há diversos *corpus* de referência acoplados na ferramenta SE[18], o que facilita as pesquisas com LC e não exige que novos *corpora* sejam inseridos para as análises e cruzamentos estatísticos.

Para as colocações da LG, foram considerados os critérios de frequência, sintáticos e pragmáticos. Assim, para a colocação ser contabilizada no presente estudo e ser considerada como tal, é necessário que ela ocorra com frequência mínima de quatro ocorrências e em pelo menos duas ou mais áreas do *corpus* analisado. O primeiro passo é digitar a palavra de busca, nesse caso o nódulo da colocação, na lacuna "lemma"[19] e selecionar "noun"[20], conforme mostra a Figura 1:

15 Para mais informações, sugerimos a leitura de Rychlý (2008), disponível em: https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf

16 O MI calcula a probabilidade de que duas palavras apareçam juntas a partir da probabilidade de sua aparição por separado. O MI ajuda a medir o grau de afinidade semântica entre duas palavras, de modo que, quanto maior for esse valor, maior será a atração entre as unidades (CHURCH; HANKS, 1990).

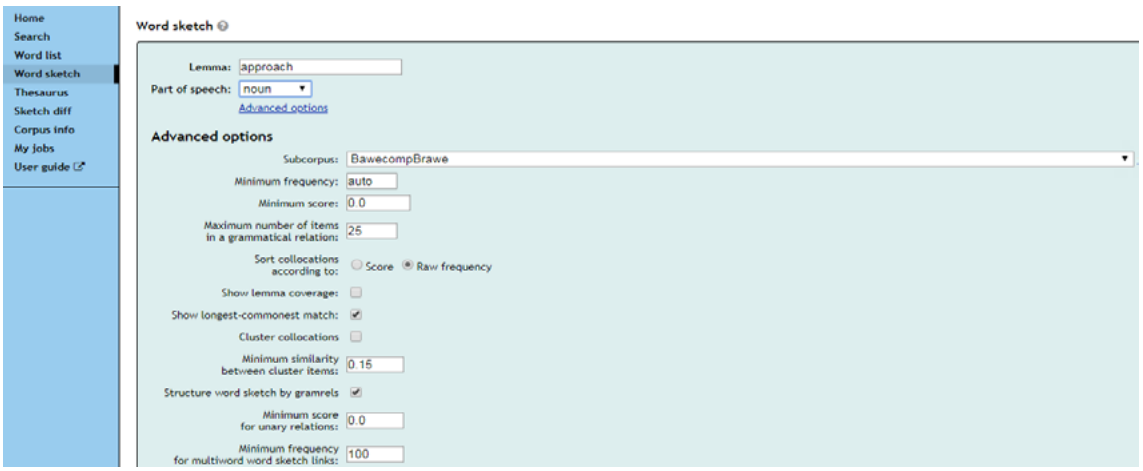
17 Para mais informações, consultar https://www.sketchengine.eu/my_keywords/logdice/

18 Este também se configura como um dos diferenciais do SE, uma vez que outras ferramentas gratuitas não costumam oferecer tal função.

19 No exemplo da Figura 1, *approach* é a palavra de busca, isto é, o nódulo da colocação.

20 Substantivo, em Português.

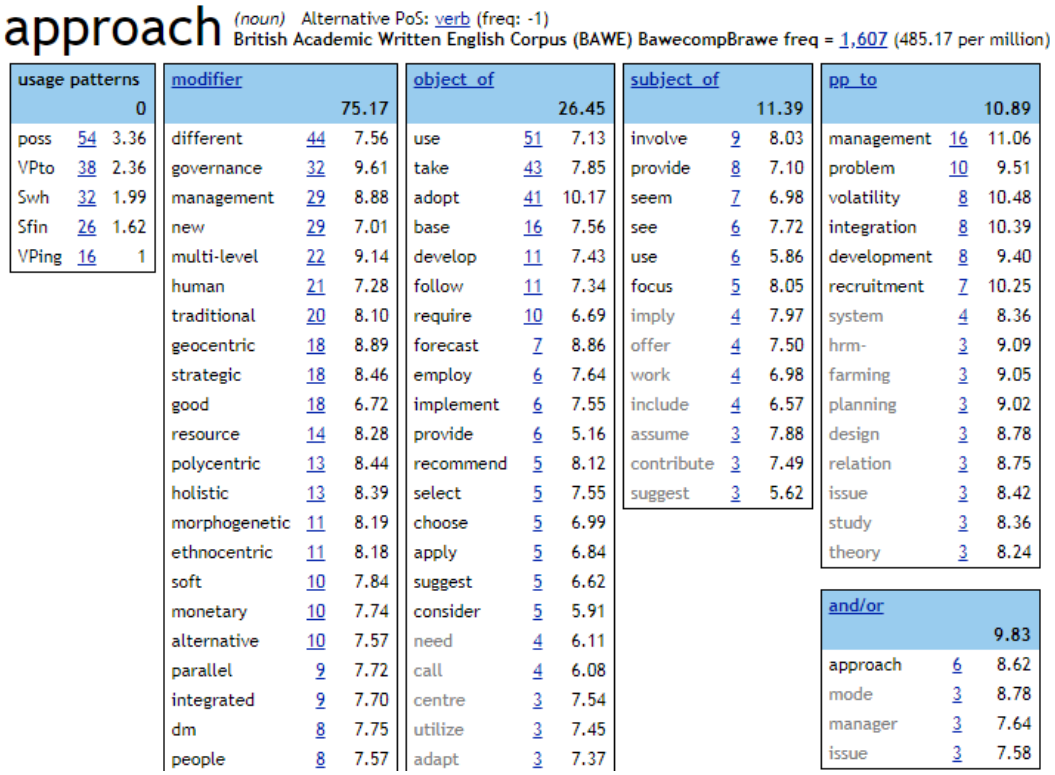
**Figura 1.** Busca das colocações a partir do nódulo na ferramenta Word Sketch




**Fonte:** Word Sketch

Apesar de na Figura 1 a frequência mínima aparecer como *auto*, o mínimo de 10 ocorrências refere-se aos colocados – e não ao nódulo. Além disso, em relação à dispersão da colocação em pelo menos duas áreas ou textos, é necessário observar a etiqueta em cada linha de concordância. Os resultados dos colocados que se combinam com o nódulo podem ser visualizados conforme a Figura 2. Nela, vêem-se cinco categorias de colocados, porém apenas três foram levadas em consideração para os propósitos da análise empreendida, a saber: *modifier*, *object of* e *subject of*, ou seja, adjetivos que antecedem o nódulo e verbos que se combinam com o nódulo quando este é o objeto e o sujeito, respectivamente. Além disso, outra etapa da análise foi verificar se de fato a colocação em questão é utilizada em pelo menos duas áreas distintas a fim de evitar estilo de escrita individual e possibilitar generalizações de uso na escrita acadêmica em inglês por alunos universitários brasileiros.

**Figura 2.** Colocados que se combinam com o nódulo da colocação



**approach** *(noun)* Alternative PoS: verb (freq: -1)
British Academic Written English Corpus (BAWE) BawecompBrawe freq = 1,607 (485.17 per million)

| usage patterns | | 0 |
|---|---|---|
| poss | 54 | 3.36 |
| VPto | 38 | 2.36 |
| Swh | 32 | 1.99 |
| Sfin | 26 | 1.62 |
| VPing | 16 | 1 |

| modifier | | 75.17 |
|---|---|---|
| different | 44 | 7.56 |
| governance | 32 | 9.61 |
| management | 29 | 8.88 |
| new | 29 | 7.01 |
| multi-level | 22 | 9.14 |
| human | 21 | 7.28 |
| traditional | 20 | 8.10 |
| geocentric | 18 | 8.89 |
| strategic | 18 | 8.46 |
| good | 18 | 6.72 |
| resource | 14 | 8.28 |
| polycentric | 13 | 8.44 |
| holistic | 13 | 8.39 |
| morphogenetic | 11 | 8.19 |
| ethnocentric | 11 | 8.18 |
| soft | 10 | 7.84 |
| monetary | 10 | 7.74 |
| alternative | 10 | 7.57 |
| parallel | 9 | 7.72 |
| integrated | 9 | 7.70 |
| dm | 8 | 7.75 |
| people | 8 | 7.57 |

| object_of | | 26.45 |
|---|---|---|
| use | 51 | 7.13 |
| take | 43 | 7.85 |
| adopt | 41 | 10.17 |
| base | 16 | 7.56 |
| develop | 11 | 7.43 |
| follow | 11 | 7.34 |
| require | 10 | 6.69 |
| forecast | 7 | 8.86 |
| employ | 6 | 7.64 |
| implement | 6 | 7.55 |
| provide | 6 | 5.16 |
| recommend | 5 | 8.12 |
| select | 5 | 7.55 |
| choose | 5 | 6.99 |
| apply | 5 | 6.84 |
| suggest | 5 | 6.62 |
| consider | 5 | 5.91 |
| need | 4 | 6.11 |
| call | 4 | 6.08 |
| centre | 3 | 7.54 |
| utilize | 3 | 7.45 |
| adapt | 3 | 7.37 |

| subject_of | | 11.39 |
|---|---|---|
| involve | 9 | 8.03 |
| provide | 8 | 7.10 |
| seem | 7 | 6.98 |
| see | 6 | 7.72 |
| use | 6 | 5.86 |
| focus | 5 | 8.05 |
| imply | 4 | 7.97 |
| offer | 4 | 7.50 |
| work | 4 | 6.98 |
| include | 4 | 6.57 |
| assume | 3 | 7.88 |
| contribute | 3 | 7.49 |
| suggest | 3 | 5.62 |

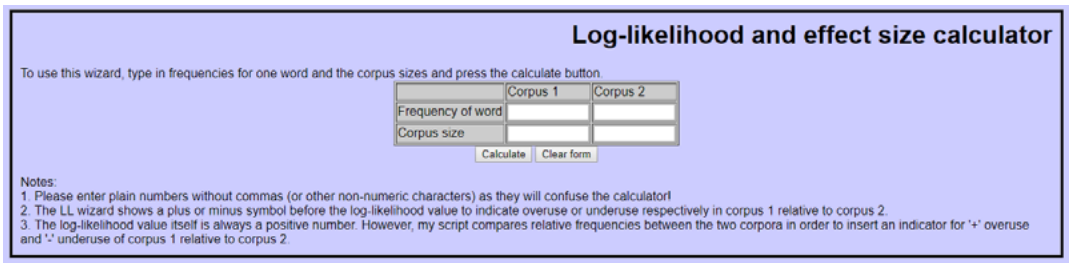
| pp_to | | 10.89 |
|---|---|---|
| management | 16 | 11.06 |
| problem | 10 | 9.51 |
| volatility | 8 | 10.48 |
| integration | 8 | 10.39 |
| development | 8 | 9.40 |
| recruitment | 7 | 10.25 |
| system | 4 | 8.36 |
| hrm- | 3 | 9.09 |
| farming | 3 | 9.05 |
| planning | 3 | 9.02 |
| design | 3 | 8.78 |
| relation | 3 | 8.75 |
| issue | 3 | 8.42 |
| study | 3 | 8.36 |
| theory | 3 | 8.24 |

| and/or | | 9.83 |
|---|---|---|
| approach | 6 | 8.62 |
| mode | 3 | 8.78 |
| manager | 3 | 7.64 |
| issue | 3 | 7.58 |

**Fonte:** Word Sketch

Após a contagem dos colocados no BAWE e no BrAWE, utilizou-se a calculadora estatística Log-Likelihood para verificar se a diferença de frequência absoluta entre os dois *corpora* é estatisticamente significativa (RAYSON, 2002).

**Figura 3.** Log-Likelihood



**Log-likelihood and effect size calculator**

To use this wizard, type in frequencies for one word and the corpus sizes and press the calculate button.

| | Corpus 1 | Corpus 2 |
|---|---|---|
| Frequency of word | | |
| Corpus size | | |

Calculate Clear form

Notes:
1. Please enter plain numbers without commas (or other non-numeric characters) as they will confuse the calculator!
2. The LL wizard shows a plus or minus symbol before the log-likelihood value to indicate overuse or underuse respectively in corpus 1 relative to corpus 2.
3. The log-likelihood value itself is always a positive number. However, my script compares relative frequencies between the two corpora in order to insert an indicator for '+' overuse and '-' underuse of corpus 1 relative to corpus 2.

**Fonte:** Word Sketch

Por convenção, se o resultado for igual ou maior que 6,63, a diferença de frequência é estatisticamente significativa, podendo ser um resultado positivo ou negativo. No primeiro caso, o resultado positivo indica um sobreuso no *corpus* dos alunos brasileiros, ao passo que o segundo caso aponta para um subuso. Por exemplo, a colocação *new*

*system* possui um valor de LL de -18.78, o que significa que ela é subutilizada pelos alunos brasileiros representados no *corpus* BrAWE.

No caso das colocações de LE, para sua identificação e extração a partir do WS, baseamo-nos em seus critérios de constituição: critério sintático e quantitativo[21]. Por exemplo, a partir de *acervo,* podemos recuperar estruturas em que o termo tem função sintática de objeto direto de um verbo (NE). Assim, a ferramenta recupera estruturas como ***abrigar*** *acervo* e ***preservar*** *acervo*, em que *abrigar* e *preservar* requerem um objeto direto (*acervo*). Do mesmo modo, é possível recuperar estruturas em que o NT *acervo* constitui um sintagma preposicionado, e a ferramenta recupera as formas nominais que co-ocorrem com o sintagma preposicionado *de acervo*, como ***preservação*** *de acervo,* ***conservação*** *de acervo,* ***guarda*** *de acervo.* Portanto, seguimos as seguintes etapas para cumprir cada critério:

a) Critério sintático:

- Partir do NT (termo de busca) para identificar os NE (coocorrentes);

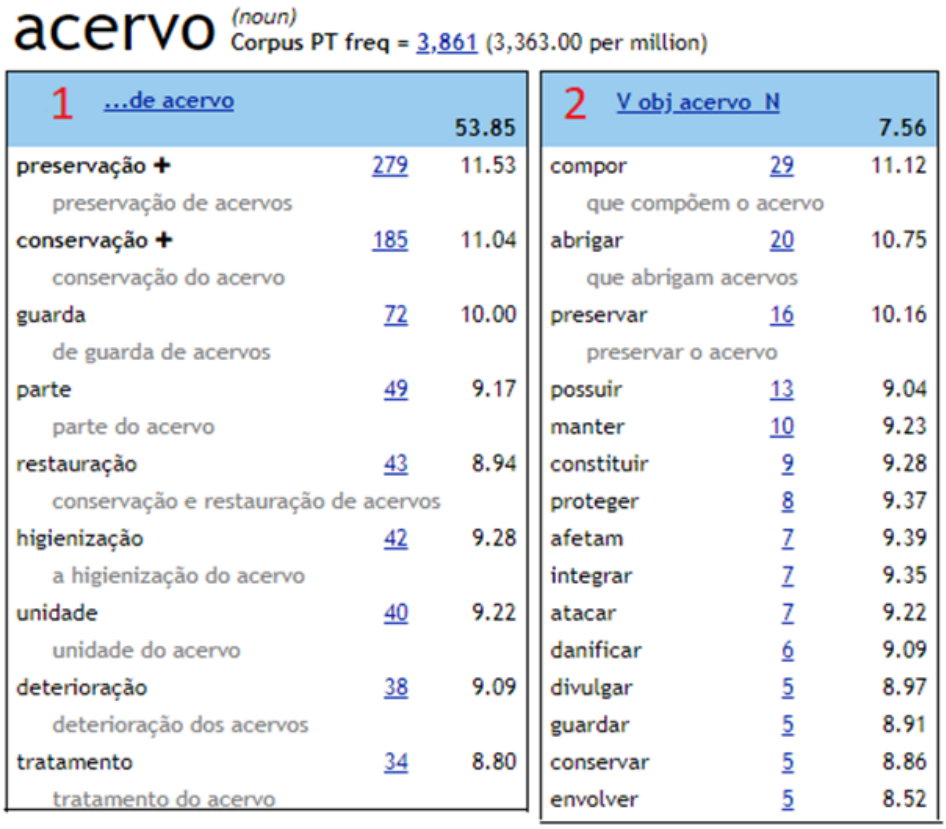
- Identificar a estrutura $[NE]_V + [NT]_N$ onde o colocado será um verbo cujo objeto direto é o NT pesquisado, o que permitiu identificar combinatórias como *conservar o acervo*.

- Identificar o coocorrente do sintagma preposicionado "de + NT", que será um nome deverbal, o que permitiu identificar combinatórias do tipo *conservação de acervo*;

b) Critério quantitativo: o corte de frequência mínima para cada estrutura candidata à colocação especializada deve ser igual ou superior a 10 ocorrências.

21 O critério pragmático-discursivo está respaldado pelos critérios de constituição do *corpus* de estudo, conforme explicamos anteriormente, que garantem que os dados sejam representativos da linguagem especializada do âmbito de estudo.

**Figura 4.** Word Sketch do NT acervo



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:** Word Sketch

A opção de busca 1 "... de acervo" aponta para a estrutura $[NE]_{Ndev}$ + $[NT]_{SP}$, na qual o NE (co-ocorrente identificado pela ferramenta) é uma ação ou processo especializado da área, expresso por um nome deverbal (*conservação*), e o NT é um sintagma preposicionado (*do acervo*), como em *conservação do acervo*. Já a opção de busca 2 "V obj_acervo_N" indica unidades de estrutura: $[NE]_{V}$+ $[NT]_{N}$ (*abrigar acervo, preservar acervo*), isto é, unidades formadas por um verbo (NE) e um objeto (NT).

Num segundo momento, após a extração automática dos dados, faz-se necessária uma análise manual. Para as duas opções de busca, foi preciso confirmar quais unidades conformavam ações ou processos especializados da área. Para isso, foi necessário verificar as concordâncias de cada um dos resultados. Essa etapa permitiu descartar combinações como *compor acervo* ou *parte do acervo/tipo do acervo*, em que *compor* não se refere a uma ação ou processo especializado da área (mas a uma característica do acervo de ser composto por determinados tipos de documentos), tampouco os substantivos *parte* e *tipo* são nomes deverbais. Assim, pudemos selecionar unidades como *abrigar acervo*, *preservar acervo*, etc., que são verbos eventivos próprios da área, bem como *preservação/salvaguarda/restauração de acervo*.

Em resumo, o WS oferece, a partir de apenas uma busca, as estruturas candidatas a colocações especializadas. Essa totalidade dos dados se deve ao fato de que a ferramenta lematiza automaticamente os *corpora* textuais nela 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 ocorre com as unidades que se combinam com o NT, ou seja, a ferramenta também recupera todas as variações morfossintáticas dos colocados, como *preservação de/do(s)/deste(s)/do(s) acervo(s)* a partir de uma busca somente.

Outro aspecto positivo da ferramenta é o *span*[22][23] que aplica em suas análises automáticas. Nesse sentido, a ferramenta reconhece combinatórias mesmo quando os dois elementos apresentam entre eles outras unidades inseridas. Por exemplo, a colocação *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 os elementos da colocação não estejam imediatamente juntos, a ferramenta os reconhece e os contabiliza nos resultados da busca[24].

Cabe ressaltar que, muito embora a ferramenta WS ofereça as unidades candidatas a colocações especializadas a partir de uma única pesquisa – o que foi um ponto extremamente positivo para uma pesquisa como a nossa, que visou recuperar um grande número de dados –, por outro lado, os resultados também podem recuperar ruído em função do critério semântico dessas unidades (como nos casos de *compor acervo* ou *parte do acervo*), reforçando a necessidade da análise manual.

## Discutindo os resultados

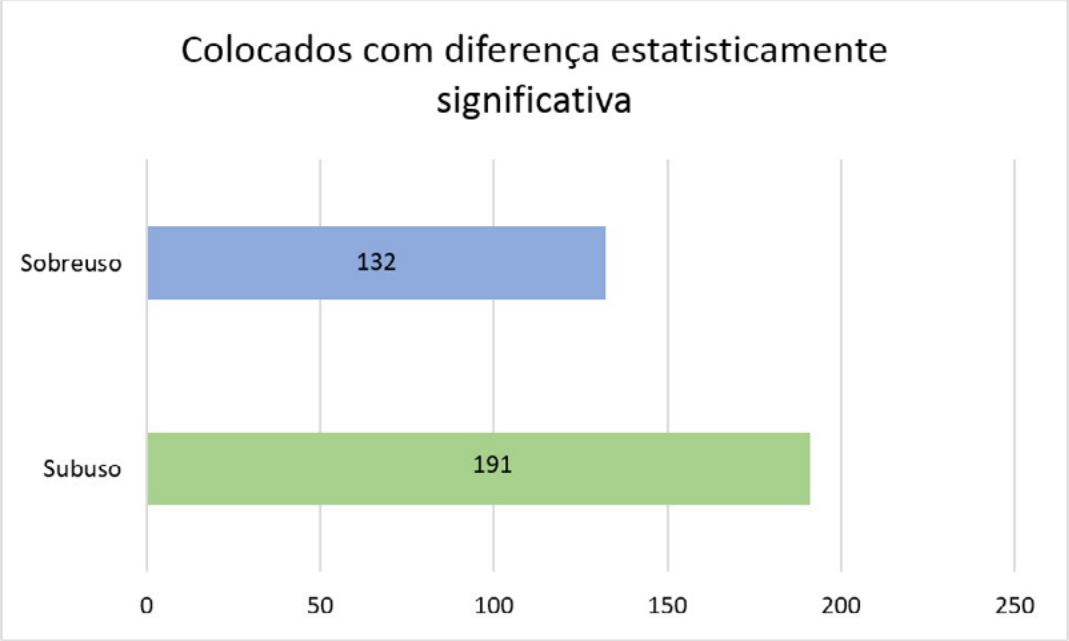
A análise quantitativa das colocações na escrita acadêmica em inglês a partir da perspectiva da LG revelou 125 nódulos e 2522 colocados. A partir da análise comparativa entre os dois *corpora*, verificou-se que, dentre os nódulos, 89 possuem diferença estatisticamente significativa, dos quais 49 são subutilizados por brasileiros e os demais 40 são sobreutilizados, conforme Gráfico 1.

22 Conforme Sinclair (1991), o *span* consiste na distância que há entre a base da colocação e seus colocados.

23 O Sketch Engine não explicita o valor exato do *span* que a ferramenta Word Sketch é capaz de processar. Contudo, a partir dos exemplos que extraímos e que apresentamos neste trabalho, julgamos ser um valor de *span* eficiente, uma vez que a ferramenta identificou, automaticamente, colocados com até cinco palavras intercaladas.

24 Entendemos que estes sejam aspectos positivos da ferramenta WS, uma vez que outros *softwares* que também permitem a identificação e extração de colocações nem sempre oferecem tais recursos. Para mais informações sugerimos a leitura de Arcos e Bevilacqua (2018).

**Gráfico 1.** Colocados com diferença estatisticamente significativa




**Fonte:** Elaboração própria

Dos 2522 colocados, 323 apresentaram diferença estatisticamente significativa, isto é, a comparação das frequências dos colocados nos dois *corpora* analisados (BAWE e BrAWE) é estatisticamente significativa. Dentre esses 323, 191 são subutilizados pelos alunos brasileiros representados no BrAWE e 132 são sobreutilizados, como aponta o Gráfico 1.

Por restrições de espaço, apenas os cinco nódulos mais frequentes do *corpus* de referência BAWE serão apresentados. Conforme a Tabela 3, *system, result, value, figure* e *process* compõem os cinco nódulos mais frequentes no BAWE. O número de palavras que se colocam com tais nódulos, isto é, os colocados, é sempre inferior no BrAWE na comparação com o BrAWE, o que indica que as produções escritas em inglês acadêmico por alunos brasileiros possuem uma menor variedade de colocações.

**Tabela 3.** Colocados em ambos os *corpora* da LG

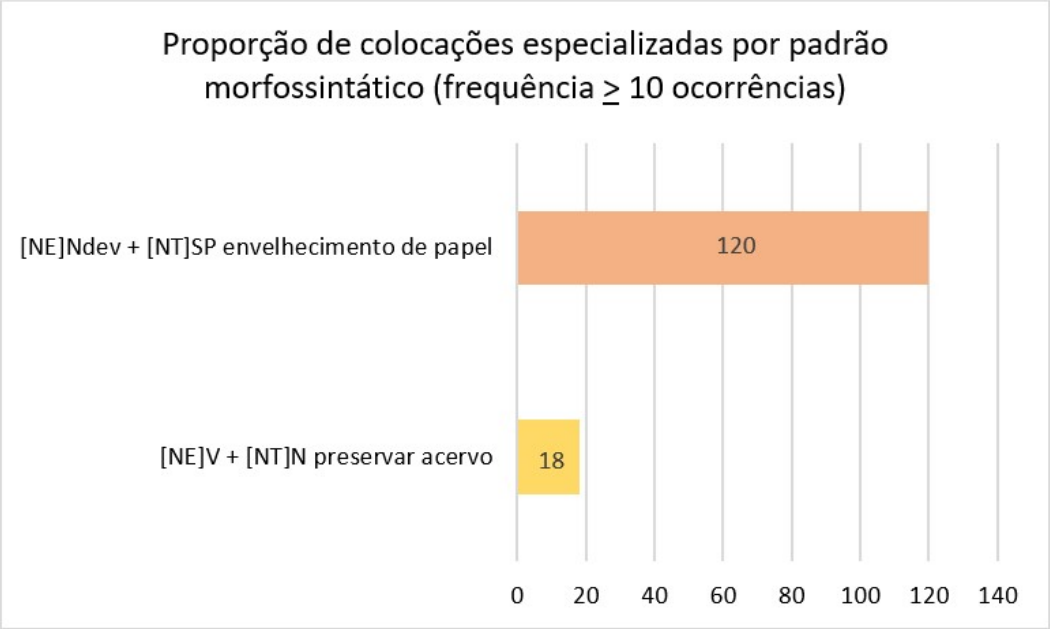
| Nódulo | Número de colocados | |
|---|---|---|
| | BAWE | BrAWE |
| **System** | 48 | 30 |
| **Result** | 56 | 44 |
| **Value** | 52 | 33 |
| **Figure** | 15 | 3 |
| **Process** | 56 | 31 |
| **TOTAL** | 227 | 141 |

**Fonte:** Elaboração própria

Para fins de explicitação do fenômeno observado na Tabela 3 e anteriormente comentado, observa-se que, dentre as colocações *control system*, *new system*, *reward system*, *production system*, *whole system*, *current system*, *computer system*, *communication system* e *complex system*, apenas *control system*, *production system*, *whole system* e *complex system* são utilizadas no *corpus* BrAWE. Ou seja, há uma grande variedade de colocados junto do nódulo *system* no *corpus* BAWE, ao passo que os colocados produzidos pelos alunos brasileiros representados no BrAWE é menor.

Quanto à análise quantitativa das colocações especializadas do *corpus* de Conservação e Restauração de bens em suporte papel (*Corpus* Papel), seguindo os critérios sintáticos (estruturas de padrão morfossintático $[NE]_{V}$ + $[NT]_{N}$ e $[NE]_{Ndev}$ + $[NT]_{SP.}$) e quantitativos (unidades com frequência de 10 ou mais ocorrências), foi possível identificar um total de 138 colocações a partir de 65 NT com a ferramenta WS. Das 138 colocações, 18 foram de estrutura morfossintática $[NE]_{V}$ + $[NT]_{N}$ e 120 de estrutura $[NE]_{Ndev}$ + $[NT]_{SP.}$, conforme ilustra o Gráfico 2.

**Gráfico 2.** Proporção de colocações especializadas por padrão morfossintático (frequência ≥ 10 ocorrências)




**Fonte:** Elaboração própria

A partir dos resultados quantitativos, foi possível analisar qualitativamente as colocações especializadas da área da Conservação e Restauração de bens em suporte papel. Dentre elas, destacamos, para este trabalho, a proporção numérica das unidades de estrutura morfossintática de nominalização $[NE]_{Ndev}$ + $[NT]_{SP.}$ que representou mais de 80% dos resultados totais, indicando uma preferência notável do uso de estruturas nominalizadas na linguagem especializada do âmbito em detrimento das estruturas verbais de padrão $[NE]_{V}$ + $[NT]_{N.}$

## Considerações finais

O presente artigo revelou que a ferramenta WS do *software* Sketch Engine é eficaz para desempenhar as funções de identificação e extração de colocações de LG e de LE, uma vez que tais etapas são facilitadas pela disponibilização de informações acerca do comportamento gramatical e colocacional do nódulo de forma sistematizada. Nesse sentido, apesar de outros programas de análise de *corpora* também oferecerem a possibilidade de estudo de colocações, o WS se mostrou eficaz ao expor os resultados de busca de forma satisfatória e semiautomatizada.

Entendemos que a eficácia da ferramenta se deve a fatores comentados ao longo das metodologias, entre os quais destacamos:

- a lematização automática do *corpus;*

- a busca por padrões morfossintáticos;

- a identificação de colocações com itens intercalados (*span*);

- o uso de índices estatísticos.

A lematização automática permite que as buscas sejam feitas a partir das estruturas morfossintáticas que o pesquisador procura identificar, sem a necessidade de realizar diferentes buscas para cada padrão morfossintático dos diferentes tipos de colocações. Outro aspecto positivo da ferramenta é a identificação de colocações no *corpus* que extrapolam os colocados imediatos. Assim, a possibilidade de se identificar combinações de palavras a partir de um *span* maior permite que unidades intercaladas por outros itens lexicais também sejam contabilizadas como colocações, sem a necessidade de um comando específico no momento da busca. Da mesma forma, o uso de índices matemáticos como o *Mutual Information* e o *logDice*, bem como a possibilidade de estabelecer cortes de frequência para as unidades extraídas, possibilitam que os critérios de constituição das colocações sejam contemplados no momento de sua identificação, tornando a tarefa mais objetiva e produtiva.

Por fim, apesar dos diferentes recursos oferecidos pela ferramenta WS que facilitam a identificação e a extração de colocações, reiteramos a necessidade de uma análise manual para este trabalho, assim como para qualquer investigação que se valha da Linguística de *Corpus*. Entendemos que nenhuma ferramenta é capaz de cumprir todos os requisitos que devem ser considerados para identificar unidades que refletem a idiossincrasia e as particularidades tanto da língua geral como das linguagens especializadas, como é o caso das colocações. Nesse sentido, as ferramentas computacionais continuam com a função de tornar o trabalho de identificação mais produtivo, sem nunca substituir a necessidade da análise manual dos dados por parte do linguista.

## Referências

ARCOS, M. **Identificação e análise de UFE eventivas na área da conservação e restauração de bens culturais móveis em suporte papel.** 2019. Dissertação (Mestrado em Lexicografia, Terminologia e Tradução) — Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

ARCOS, M.; BEVILACQUA, C. R. Metodologias para a extração e identificação de unidades fraseológicas especializadas eventivas em corpora textuais. **Guavira Letras**, n. 27, p. 75-95, 2018.

ALSOP, S.; NESI, H. Issues in the development of the British Academic Written English (BAWE) corpus. **Corpora**, v. 4, n. 1, p. 71-83, 2009.

BERBER SARDINHA, T. B. Lingüística de corpus: histórico e problemática. **DELTA: Documentação e Estudos em Linguística Teórica e Aplicada**, v. 16, n. 2, p. 323-367, 2000.

BEVILACQUA, C. R. *et al*. CLEs da linguagem jurídica: as combinatórias discursivas do texto legislativo brasileiro *In*: ALVAREZ, M. L. O. **Tendências atuais na pesquisa descritiva e aplicada em fraseologia e paremiologia**. v. 2. Campinas: Pontes Editores, 2012. p. 241-253.

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

BIBER, D.; DOUGLAS, B.; CONRAD, S.; REPPEN, R. **Corpus linguistics:** Investigating language structure and use. Cambridge: Cambridge University Press, 1998.

CHOI, S. Processing and learning of enhanced English collocations: An eye movement study. **Language Teaching Research**, v. 21, n. 3, p. 403-426, 2017.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational linguistics**, v. 16, n. 1, p. 22-29, 1990.

DURRANT, P.; SCHMITT, N. To what extent do native and non-native writers make use of collocations? **IRAL-International Review of Applied Linguistics in Language Teaching**, v. 47, n. 2, p. 157-177, 2009.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. **Studies in linguistic analysis**, 1957.

FRANKENBERG-GARCIA, A. Investigating the collocations available to EAP writers. **Journal of English for Academic Purposes**, v. 35, p. 93-104, 2018.

GARDNER, S.; NESI, H. A classification of genre families in university student writing. **Applied linguistics**, v. 34, n. 1, p. 25-52, 2012.

HILL, J. Revising priorities: From grammatical failure to collocational success. **Teaching collocation**, p. 47-69, 2000.

HYLAND, K. **English for academic purposes:** An advanced resource book. Routledge, 2006.

KILGARRIFF, A.; BAISA, V.; BUŠTA, J.; JAKUBÍČEK, M.; KOVÁŘ, V.; MICHELFEIT, J. The Sketch Engine: ten years on. **Lexicography**, v. 1, n. 1, p. 7-36, 2014.

KJAER, A. L. Phraseology research: state-of-the-art: methods of describing word combinations in language for specific purposes. **Terminology science and research:** Journal of International Institute for Terminology Research, v. 1, n. 1-2, p. 3-20, 1990.

L'HOMME, M. C. **La terminologie:** príncipes et techniques. Montreal: Paramètres, 2004.

MATTE, M. L. **A corpus-based study on the use of academic collocations in English by Brazilian students.** 2019. Dissertação (Mestrado em Linguística Aplicada) — Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

MCENERY, T.; HARDIE, A. **Corpus linguistics:** Method, theory and practice. Cambridge: Cambridge University Press, 2011.

MCENERY, T.; WILSON, A. **Corpus Linguistics**. Edinburgh: Edinburgh University Press, 1996.

NESSELHAUF, N. **Collocations in a learner corpus.** Amsterdam: John Benjamins, 2005.

PAVEL, S. A fraseologia na língua de especialidade: metodologia de registro nos vocabulários terminológicos. *In*: FAULSTICH, E.; ABREU, S. P. **Linguística aplicada à terminologia e à lexicologia:** cooperação internacional: Brasil e Canadá. Porto Alegre: UFRGS, Instituto de Letras, NEC, 2003.

PEARSON, J. **Terms in context**. Amsterdam/Philadelphia: John Benjamins, 1998.

PICHT, H. LSP phraseology from the terminological point of view. **Terminology science & research:** Journal of International Institute for Terminology Research, v. 1, n. 1-2. Viena: International Network for Terminology, 1990. p. 33-48.

RYCHLÝ, P. A Lexicographer-Friendly Association Score. **Proceedings of Recent Advances in Slavonic Natural Language Processing**, RASLAN 2008, p. 6-9, 2008.

SILVA, L. G. Compilation of a Brazilian academic written English corpus. **Revista e-scrita:** Revista do Curso de Letras da UNIABEU, v. 8, n. 2, p. 32-47, 2017.

SINCLAIR, J. **Corpus, concordance, collocation**. Oxford: Oxford University Press, 1991.

RAYSON, Pl. **Matrix:** A statistical method and software tool for linguistic analysis through corpus comparison. 2002. Tese (Doutorado em Ciência da Computação) – Lancaster University, Lancaster, 2002.

RYCHLY, P. A Lexicographer-Friendly Association Score. *In*: PETR, S.; ALES, H. **Proceedings of Recent Advances in Slavonic Natural Language Processing**. Brno: Masaryk University, 2008. p. 6-9.

WRAY, A. Formulaic sequences in second language teaching: Principle and practice. **Applied linguistics**, v. 21, n. 4, p. 463-489, 2000.

---

COMO CITAR ESTE ARTIGO: ARCOS, Manuela; MATTE, Marine Laísa. Word Sketch como ferramenta para extração de colocações. **Revista do GEL**, v. 17, n. 2, p. 61-81, 2020. Disponível em: https://revistadogel.gel.org.br/

DOI: http://dx.doi.org/10.21165/gel.v17i2.2771

Submetido em: 13/10/2019 | Aceito em: 30/07/2020.

---