Antares
LETRAS E HUMANIDADES ISSN: 1984-4921
Programa de Pós-Graduação em Letras e Cultura
UCS UNIVERSIDADE DE CAXIAS DO SUL

DOI: http://dx.doi.org/10.18226/19844921.v12.n25.08

Linguística de *corpus* como metodologia para o desenvolvimento de glossários especializados: uma experiência com relatórios de sustentabilidade[*]

Corpus linguistics as a methodology for the development of specialized glossaries: an experience with sustainability reports

Carolina Rübesam Ourique[**]

Resumo

O objetivo deste artigo é descrever a metodologia adotada para a construção de um glossário bilíngue português-inglês de termos e fraseologias da área de sustentabilidade com base nos pressupostos da Linguística de Corpus. Apesar da rápida expansão nas últimas décadas, a área ainda carece de obras terminográficas confiáveis que auxiliem adequadamente na realização de tarefas de tradução e redação de textos sobre sustentabilidade, sem comprometer a convencionalidade. Assim, compilou-se um corpus comparável bilíngue de Relatórios de Sustentabilidade, do qual extraíram-se termos e fraseologias equivalentes a partir de palavras-chave simples e compostas levantadas automaticamente pela ferramenta Sketch Engine e selecionadas manualmente. A metodologia descrita possibilitou a compilação de uma obra de referência que leva em consideração os padrões evidenciados pelos textos autênticos e pode ser aplicada a outros domínios.

Palavras-chave

Relatórios de Sustentabilidade. Linguística de Corpus. Tradução. Terminologia bilíngue.

Abstract

This article aims to describe the methodology adopted to build a bilingual Portuguese-English glossary of terms and phraseologies of sustainability. Despite being expanding rapidly in recent decades, and therefore requiring services as writing, translation and version of its reports, the area lacks reliable terminological materials that help professionals to effectively perform translation and writing tasks. Hence, we have compiled a comparable bilingual corpus of Sustainability Reports in order to obtain equivalent terms and phrases from simple and compound keywords, which were extracted automatically by the Sketch Engine tools and further manually selected. The methodology described enabled the compilation of a reference work that takes

[*] Agradeço à professora Dra. Rozane Rebechi pela orientação da minha pesquisa de mestrado, junto ao PPG do Instituto de Letras da UFRGS, e pela revisão minuciosa das primeiras versões deste artigo.
[**] Universidade Federal do Rio Grande do Sul (UFRGS).

into account the patterns raised by authentic texts, and can be applied to other domains.

Keywords

Sustainability Reporting. Corpus Linguistics. Translation. Bilingual Terminology.

Introdução

No Brasil, desde os anos 1990, o uso de termos relacionados à sustentabilidade vem se consolidando por meio da divulgação dos impactos ambientais provocados pelas atividades de empresas e indústrias na forma de um documento conhecido como Relatório de Sustentabilidade (RS) (GRI, 2018). Diante da tendência de crescimento no número dessas publicações em âmbitos nacional e internacional, percebe-se uma oportunidade para tradutores e redatores. O inglês é a língua franca da comunicação internacional, portanto não surpreende que todo esse material transite entre esse idioma e os outros. No entanto, quando consideramos o par de línguas português-inglês, observamos também nessa área escassez de obras de referência produzidas com metodologia de base empírica, conforme os estudos de Machado e Bevilacqua (2018) e Rebechi e Silva (2017).

Em consonância com os estudos sobre padrões da linguagem especializada realizados pelo grupo Termisul [1], o objetivo da nossa pesquisa é explicitar a metodologia adotada na construção de um glossário bilíngue voltado aos profissionais envolvidos na redação e tradução de Relatórios de Sustentabilidade que apresentasse, de forma clara e contextualizada, os termos e as fraseologias recorrentes em português, assim como seus equivalentes funcionais em inglês.

Na seção “Relatórios de Sustentabilidade: uma breve descrição”, apresentamos uma breve definição desse gênero textual. Após, descrevemos a construção dos dois *corpora* utilizados na pesquisa, bem como os critérios e os parâmetros adotados para o levantamento dos dados. A seção seguinte se dedica à análise dos dados coletados: como realizou-se a seleção dos candidatos a termos, quais foram os critérios de limpeza e a consequente exclusão de palavras-chave. A seção final do artigo traz alguns comentários sobre as limitações que as ferramentas

[1] Termisul UFRGS – site: http://www.ufrgs.br/termisul/.

computacionais apresentam e sobre a importância da análise humana para a validação dos dados.

Relatórios de Sustentabilidade e a tradução funcionalista

Segundo Ciapuscio (1998), "texto" constitui um objeto de caráter linguístico e ao mesmo tempo comunicativo, sendo, por essa razão, um objeto complexo. Por ser um produto da atividade intelectual humana, é passível de análise por diferentes perspectivas, como a do processo (a atividade de produzir um texto) ou a do produto (resultado da atividade). Ciapuscio (1998) prevê que os textos se agrupam em classes textuais conforme aspectos macro e microestruturais. A macroestrutura de um texto, segundo a autora, está relacionada com questões de como e para que se emprega tal texto. A microestrutura, por sua vez, se refere aos seus padrões léxicos, morfossintáticos e pragmáticos, elementos que se somam para configurar o caráter de especialidade do texto.

Mediante o exposto, podemos afirmar que os Relatórios de Sustentabilidade (RS) configuram-se como texto especializado (produto), resultante de uma atividade intelectual (a escrita) que apresenta uma macroestrutura característica (isto é, são escritos sob as diretrizes da GRI, para reportar os impactos das empresas) e contêm traços específicos de padrões lexicais, morfossintáticos e, principalmente, pragmáticos (sua temática é a sustentabilidade, e são escritos para empresários e especialistas em um contexto de divulgação de informações). Além disso, os RS contemplam dados textuais e numéricos de todos os setores de uma empresa com a finalidade de divulgar os impactos provocados pelas atividades das corporações.

Trata-se de uma importante ferramenta corporativa, uma vez que auxilia as organizações a aprimorarem seus processos e comunicarem seu desempenho nas esferas econômica, ambiental, social e de governança, definindo objetivos e administrando os impactos de maneira mais eficaz. São compostos por conteúdos de diferentes setores organizacionais (áreas técnicas reescritas sob a ótica da linguagem do marketing) e contêm terminologia diversa. Assim, os RS podem se tornar complexos e, consequentemente, desafiadores para tradutores e redatores. Essa tarefa mostra-se ainda mais laboriosa quando a área não conta com materiais terminográficos bilíngues confiáveis, que não só ajudariam a manter a uniformidade terminológica, mas também a naturalidade dos textos desse gênero em diferentes idiomas.

Nesta pesquisa, apoiamo-nos na teoria funcionalista da tradução, segundo a qual a função de um texto não é uma qualidade inerente a ele, uma vez que a tradução deve considerar a situação de recepção do texto traduzido, incluindo o público-alvo (NORD, 2012). Assim, a tradução é uma comunicação mediada ou translacional: ela precisa cruzar uma barreira tanto da linguagem quanto da cultura para chegar adequadamente ao seu destino. Na nossa visão, o texto especializado é um repositório de terminologias inserido em determinada cultura, a qual determinará a forma como esses textos serão adequadamente transpostos, considerando-se, portanto, questões linguísticas e culturais.

Além dos aspectos translacionais, os novos paradigmas teóricos da área da Terminologia "[...] estabelecem princípios voltados ao uso real dos termos, isto é, a sua utilização em textos especializados" (BEVILACQUA, 2013, p. 12). O uso de textos produzidos com propósito autenticamente comunicativo (informativo, instrucional etc.), e não construídos com vistas ao estudo da linguagem, tem eficácia no levantamento de termos e fraseologias em contexto, que podem servir para a construção de produtos terminográficos, como glossários bilíngues, que auxiliarão redatores e tradutores na produção de textos que fluam com naturalidade. Considerando esse ponto de vista, a pesquisa realizou-se a partir da premissa de utilizar textos autênticos para extração de termos e fraseologias, conforme será explicitado a seguir.

Linguística de corpus como metodologia para a construção do glossário

A pesquisa alinhou-se aos pressupostos de Tagnin (2013) e Koester (2010) no que se refere, respectivamente, à definição de *corpus* e aos critérios a serem adotados no trabalho com *corpora* especializado. Além deles, contamos com alguns conceitos sobre construção de *corpus* referidos em Berber Sardinha (2004), como os critérios em relação à origem e ao formato dos textos, entre outros.

Pareceu-nos bastante adequada a perspectiva de Koester (2010), que considera mais relevante o que o *corpus* contém do que seu tamanho, propriamente. O autor defende que *corpora* especializados não necessitam ser demasiadamente extensos (não tanto quanto os *corpora* de língua geral, por exemplo) para gerarem resultados significativos. O que os torna confiáveis é o cuidado durante sua construção, isto é, os critérios estabelecidos para sua compilação. Além disso, o pesquisador pondera que o léxico especializado e as suas estruturas são mais

propensos a apresentarem padrões regulares do que um *corpus* geral (O'KEEFE et al., 2007 apud KOESTER, 2010).

A representatividade situacional dos *corpora* utilizados nesta pesquisa pôde ser verificada pela seleção de apenas um gênero textual coletado de diferentes fontes, ou seja, de diferentes empresas. Caso as amostras fossem coletadas de uma mesma empresa, por exemplo, não representariam o gênero, mas sim o *modus dicendi* da organização, o que não cumpriria com o propósito da pesquisa.

***Corpus* de estudo: compilação e processamento**

Para a realização do estudo, foi construído um *corpus* comparável composto por (i) RS escritos originalmente em português brasileiro e (ii) RS escritos originalmente em inglês estadunidense. Os textos foram selecionados com base no *ranking* de duas publicações: a edição 2016 do Guia de Sustentabilidade da Revista Exame, por ser a publicação mais reconhecida do Brasil, e a *Corporate Knights* 2017 (CK), publicação anual reconhecida em âmbito internacional em que são eleitas as 100 empresas mais sustentáveis do mundo[2]. Cada uma dessas publicações possui metodologia própria para avaliar as empresas participantes do *ranking*[3], embora seus critérios se assemelhem, já que ambas se baseiam nas diretrizes da GRI[4].

Da lista das empresas brasileiras, foram escolhidas aquelas eleitas como as mais sustentáveis em seus respectivos setores (Duratex, Fibria, Klabin, Natura e Votorantim Metais), cujos relatórios estavam disponíveis para download nos seus sites institucionais. Para os textos em inglês, foram selecionadas as empresas estadunidenses com maior pontuação no *ranking* da CK. Esse material foi convertido para o formato .txt. Vale ressaltar que, apesar de a ferramenta utilizada - Sketch Engine[5] (KILGARRIFF et al., 2014) - ser capaz de processar textos em diversos formatos – PDF, DOC etc. -, decidimos armazenar o *corpus* em um formato que pudesse também ser processado por outras ferramentas, que, em geral, requerem os textos sem qualquer formatação.

[2] http://www.corporateknights.com/reports/2017-global-100/2017-global-100-results-14846083
[3] https://exame.abril.com.br/edicoes/guia-de-sustentabilidade-2016
[4] Global Reporting Initiative: https://www.globalreporting.org/information/sustainability-reporting/Pages/gri-standards.aspx
[5] Gerenciador de corpus e software on-line de análise textual, cuja licença pode ser adquirida em www.sketchengine.co.uk.

Os nomes dos arquivos foram padronizados em códigos da seguinte forma: idioma (PT, português; EN, inglês), seguido de sigla com três letras para cada empresa (DTX: Duratex; FBR: Fibria; KLB: Klabin; NTR: Natura; VTM: Votorantim Metais; ALL: Allergan; CGP: Colgate-Palmolive; CIS: Cisco; INT: Intel; e JNJ: Johnson & Johnson). Inserido após RS, iniciais que remetem ao gênero analisado, o ano do *ranking* encerra a nomeação dos arquivos, conforme o exemplo: PT_VTM_RS2015.txt.

Processamento automático do *corpus*

Nossa escolha pelo software Sketch Engine (SE)[6] para análise dos textos foi pautada pela praticidade da ferramenta: a plataforma do programa é *on-line* e processa os textos com agilidade. Além disso, a interface é bastante amigável, possibilitando, por exemplo, a lematização das palavras, a sistematização de colocados, entre outros aspectos positivos. Apesar de outros softwares de análise textual oferecerem diversos recursos gratuitamente – o AntConc é, possivelmente, um dos mais populares –, eles não contam com todas as ferramentas que foram essenciais para este estudo. Três principais fatores determinaram nossa escolha pelo SE, a saber: (i) a ferramenta disponibiliza diversos *corpora* em diferentes idiomas, dispensando a construção de *corpora* de referência; (ii) o SE contém um lematizador, que permite que as formas flexionadas das palavras sejam agrupadas; e (iii) a ferramenta Word Sketch evidencia o comportamento gramatical e colocacional das palavras de busca. Essas e outras funções serão detalhadas adiante.

O programa calculou automaticamente o tamanho dos dois sub*corpora*, os quais estão resumidos nas Tabelas 1 e 2.

Tabela 1 - Composição do sub*corpus* em português (RelSustenta_PT)

Nome do arquivo	Número de *tokens*
PT_DTX_RS2016.txt	33.382
PT_FBR_RS2016.txt	20.648

[6] Disponível em: <the.sketchengine.co.uk>.

PT_KLB_RS2016.txt	8.065
PT_NTR_RS2016.txt	55.769
PT_VTM_RS2016.txt	41.416
Total	**159.280**

Tabela 2 - Composição do sub*corpus* em inglês (RelSustenta_EN)

Nome do arquivo	**Número de *tokens***
EN_ALL_RS2017.txt	12.938
EN_CGP_RS2016.txt	45.687
EN_CIS_RS2016.txt	76.803
EN_INT_RS2016.txt	51.348
EN_JNJ_RS2016.txt	46.446
TOTAL	**235.334**

É possível notar nas Tabelas 1 e 2 que o número de palavras do sub*corpus* em inglês supera o *corpus* em português em cerca de 50%, apesar de ambos terem sido compilados com a mesma quantidade de textos, estes que são provenientes de fontes afins e pertencentes ao mesmo gênero. Segundo Katan (1999, p. 177), essa diferença se explica por questões culturais: "[I]ndivíduos, grupos e culturas (e em diferentes momentos) possuem prioridades divergentes no que tange à quantidade de informação (texto) necessária a ser explicitada para que a comunicação seja bem-sucedida" (tradução nossa)[7]. O estudioso denomina *high context* a cultura cuja comunicação se apoia em grande parte em conhecimento compartilhado, e *low context* aquela para qual o conhecimento comum nem sempre é suficiente e, portanto, requer mais detalhes e explicações. Nessa escala, o autor posiciona os Estados Unidos em direção à base, enquanto o Brasil e outros países da América Latina ocupam posição mais elevada. Pesquisas com *corpora* comparáveis em português brasileiro e em inglês estadunidense compostos por textos de outros gêneros, como sites de hotéis (NAVARRO, 2012) e obituários (REBECHI; SILVA, 2018), corroboram essa hipótese.

[7] No original: *"[I]ndividuals, groups, and cultures (and at different times) have differing priorities with regard to how much information (text) needs to be made explicit for communication to take place*."

Procedimentos para extração de termos e fraseologias

A seleção de candidatos a termos foi iniciada pela análise de palavras-chave, cujas listas foram obtidas escolhendo-se dois *corpora* de referência para o contraste: ptTenTen 11[8], e o *Open American National Corpus*[9], respectivamente, para o levantamento de palavras-chave em português e em inglês. Entre os diversos *corpora* oferecidos pelo SE, optamos por esses dois por serem compostos apenas de textos escritos (e não orais) em português brasileiro e em inglês estadunidense. Para a extração de palavras-chave, cada sub*corpus* de estudo foi comparado ao *corpus* de referência do respectivo idioma, usando a ferramenta "Keyword/Terms".

Foram realizados testes-piloto com a ferramenta antes de determinarmos os valores finais dos parâmetros. Iniciamos pela configuração da frequência mínima de três ocorrências para os termos ou palavras-chave, e o parâmetro *simple math*[10] foi mantido com o default (1), valor que possibilita que mesmo palavras com poucas ocorrências sejam identificadas como chave. No entanto, o levantamento realizado com essa configuração apresentou muitos resultados irrelevantes, já que mesmo expressões sem sentido completo, como "lealdade cns", "cuidado com", "inclui capão", "rodovia br", entre outras, foram consideradas chave por ocorrerem ao menos três vezes.

Nos testes-piloto subsequentes, aumentou-se o parâmetro *simple math*, mantendo-se o valor de frequência mínima de três ocorrências. Como os dados se mantiveram semelhantes aos resultados do primeiro teste, os valores foram novamente ajustados. Somente obtivemos melhora na qualidade e na consistência dos dados quando elevamos o valor *simple math* para 1000, estabelecemos o número mínimo de ocorrências em cinco e estabelecemos em 1000 o número máximo de palavras. Com esses parâmetros, a ferramenta gerou duas listas: uma de palavras-chave simples e outra de palavras-chave compostas. Ambas as listas foram armazenadas no programa Excel para que pudessem ser classificadas em ordem alfabética, ordem de frequência absoluta ou ordem de chavicidade (que indica a

[8] Disponível em: <https://www.sketchengine.eu/pttenten-portuguese-corpus/>.
[9] Disponível em: <https://www.sketchengine.eu/oanc_masc-corpus/>.
[10] Variável utilizada pelo SE para calcular a extração das palavras-chave simples e compostas do *corpus* de estudo quando comparado a um *corpus* de referência. Ele calcula maior ou menor frequência; quanto maior o valor do parâmetro, a tendência é que sejam selecionadas as palavras de frequência mais alta.

frequência relativa das palavras no *corpus* de estudo em relação ao *corpus* de referência).

Apesar de o SE possuir um etiquetador morfossintático, possibilitando a lematização das palavras e priorizando o levantamento de palavras de conteúdo – substantivos, verbos lexicais etc. –, é importante ressaltar que a lista de termos simples e compostos gerada automaticamente, a partir dos ajustes estabelecidos, não configura uma lista pronta de termos e fraseologias características de determinada área, que pudessem servir de imediato como entradas ou subentradas de um glossário especializado. O olhar do pesquisador sobre essas listas revela a necessidade de limpeza manual e organização dos dados levantados automaticamente. Para facilitar essa seleção, juntamos as duas listas de palavras-chave de cada idioma em apenas uma lista para cada idioma.

Após unificarmos as listas de palavras simples e compostas, organizamos as palavras alfabeticamente no Excel, agrupamento que possibilitou a visualização das expressões com a mesma letra inicial. A limpeza, passo seguinte, consistiu em eliminar candidatos que apresentassem pelo menos um dos seguintes critérios:

- Nomes próprios (pessoas, organizações ou lugares. Por exemplo: Fibria, Ekos, Deca, Cajamar, Niquelândia, Cisco, Ushahidi, Protex etc.);

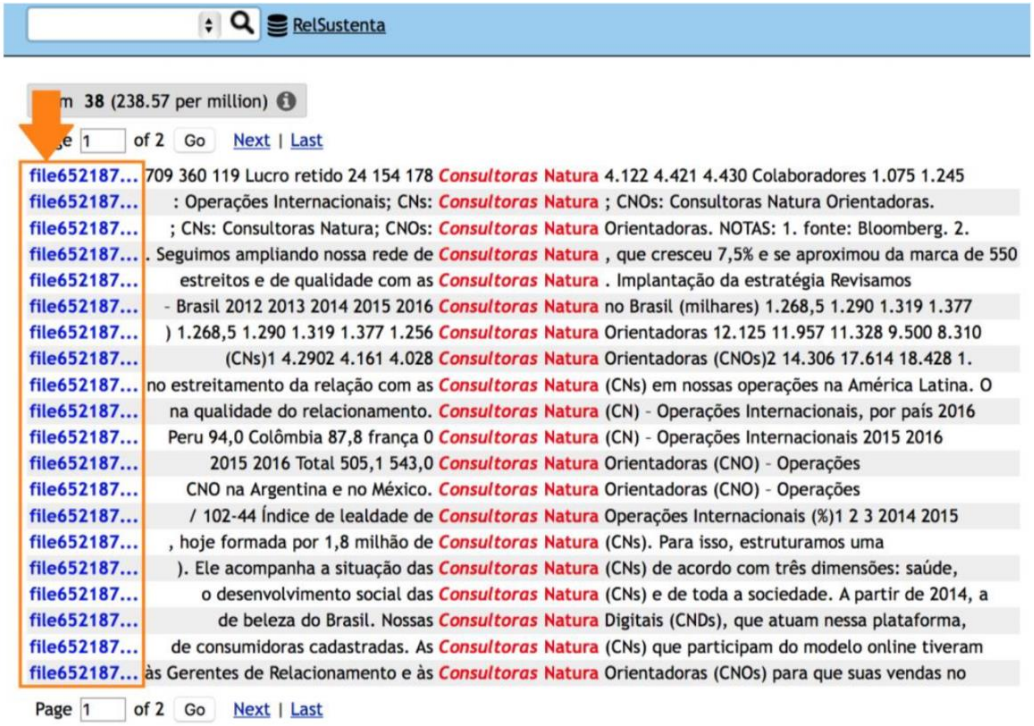
- Números, elementos químicos e símbolos do sistema métrico (por exemplo: m3, CO2, kW, hectares, Zn, etc.);

- Abreviaturas (por exemplo: fem., masc., un., FY16, CEO, etc., com exceção das siglas que representam termos da área, como ODS – objetivos do desenvolvimento sustentável);

- Ruídos (por exemplo: button, nd, title, number etc., pois são indicações de página, cabeçalhos, rodapés, informações que constam na configuração dos arquivos PDF convertidos para acesso *on-line*, e alguns itens de linguagem de programação, como códigos de programação que a ferramenta incluiu);

- Candidatos que ocorreram em apenas um dos textos foram desconsiderados por falta de representatividade, pois poderia indicar alguma particularidade de determinada empresa. Exemplo disso é o agrupamento “Consultoras Natura”, que ocorreu 38 vezes no sub*corpus* em português, e foi selecionado como termo pela ferramenta. A Figura 1 mostra as linhas de concordância dessa expressão de busca.

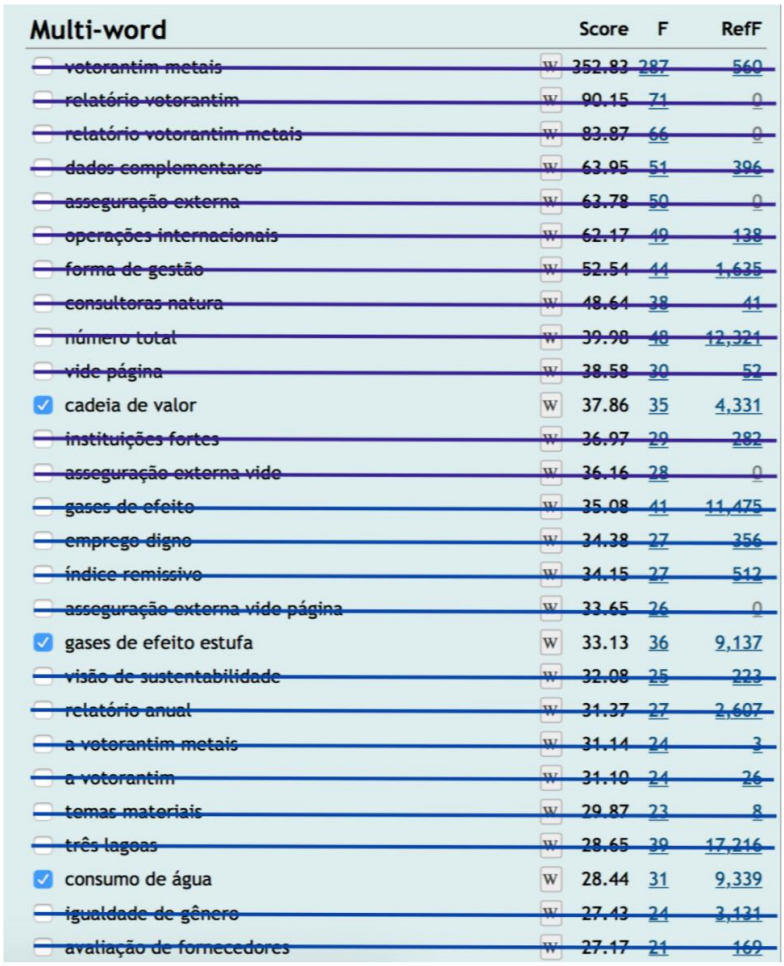
Figura 1 - Amostra das ocorrências de “Consultoras Natura”.



Como foi possível confirmar, trata-se de expressão que ocorre exclusivamente nos relatórios da Natura (cujo texto está identificado com o código file511907), não configurando, portanto, termo da área da sustentabilidade.

Contudo, uma simples lista de palavras isoladas e descontextualizadas, alinhadas aos seus (possíveis) equivalentes, não é suficiente para que o redator ou tradutor produza um texto que flua com naturalidade. Um glossário especializado que vise a esse fim deve abranger, também, o entorno dessas palavras, ou seja, seus colocados, trazendo as colocações e fraseologias características da área. Entre as primeiras palavras-chave compostas em português observamos “consumo de água”, com 31 ocorrências, conforme pode ser visualizado na Figura 2 que apresenta os itens eliminados – tachados -, assim como os validados – ticados -, da lista em português após a limpeza.

Figura 2 - Amostra da parte superior da tela do SE exibindo a lista de (candidatos a) termos compostos fornecida pela ferramenta “Keywords/Terms”.



Multi-word	Score	F	RefF
votorantim metais	352.83	287	560
relatório votorantim	90.15	71	0
relatório votorantim metais	83.87	66	0
dados complementares	63.95	51	396
asseguração externa	63.78	50	0
operações internacionais	62.17	49	138
forma de gestão	52.54	44	1,635
consultoras natura	48.64	38	41
número total	39.98	48	12,321
vide página	38.58	30	52
cadeia de valor	37.86	35	4,331
instituições fortes	36.97	29	282
asseguração externa vide	36.16	28	0
gases de efeito	35.08	41	11,475
emprego digno	34.38	27	356
índice remissivo	34.15	27	512
asseguração externa vide página	33.65	26	0
gases de efeito estufa	33.13	36	9,137
visão de sustentabilidade	32.08	25	223
relatório anual	31.37	27	2,607
a votorantim metais	31.14	24	3
a votorantim	31.10	24	26
temas materiais	29.87	23	8
três lagoas	28.65	39	17,216
consumo de água	28.44	31	9,339
igualdade de gênero	27.43	24	2,131
avaliação de fornecedores	27.17	21	169

A partir da lista ‘limpa’, passamos, então, à análise detalhada dos itens a fim de validar sua seleção como termos ou fraseologias que comporão o glossário. Para auxiliar na tarefa, utilizamos a ferramenta Word Sketch, que processa os colocados das palavras-chave e apresenta um resumo do comportamento gramatical e colocacional da palavra, esquematizado na forma de um quadro, conforme mostra a Figura 3.

A ferramenta “Word Sketch” faz a busca a partir de um núcleo, e neste caso, usamos o núcleo de “consumo de água”, isto é, “água”. Conforme a ferramenta “Keywords/Terms” já havia indicado, as 31 ocorrências de “consumo de água” também aparecem no quadro resumido do Word Sketch (ver Figura 3, em que a expressão aparece circulada) como sendo a combinação mais frequente com a palavra-chave “água”:

Figura 3 - Tela do SE apresentando esquema sintético do comportamento do termo "água", gerado pela ferramenta Word Sketch.

A ferramenta apresenta os dados de duas formas, que separamos em [1] e [2]: a parte [1] mostra as ocorrências em que o substantivo "água" é modificado por outras palavras (*n_modifier*), por exemplo, "limpo", "novo", "reutilizado" etc. Vale ressaltar que as palavras são apresentadas nas formas canônicas (dicionarizadas), pois a ferramenta as lematiza – verbos conjugados são unidos à forma infinitiva, substantivos e adjetivos são apresentados no masculino singular etc. No entanto, clicando-se na frequência (número sublinhado), é possível ver a forma flexionada da ocorrência no texto (que, neste caso, é "água limpa"). A parte [2] mostra os casos em que a palavra de busca modifica outras (*modifies*). Assim, observa-se que a palavra "água" modifica "consumo", formando a combinação "consumo de água". As demais combinações (por

exemplo, “água limpa e saneamento” e “água nova”) foram desconsideradas por se enquadraram em algum dos critérios de exclusão.

Essa ferramenta (Word Sketch) possibilitou a análise dos candidatos visualizando-se as colocações, ou *clusters* (agrupamentos), mais frequentes. A partir da frequência absoluta, pudemos analisar as linhas de concordância. Esse foi o momento em que se verificou a origem, isto é, em qual texto do *corpus* a expressão foi encontrada, e também o seu contexto. Além disso, foi possível alinhar as palavras de busca à esquerda, para que pudessem ser analisadas as palavras à esquerda da expressão em ordem alfabética; à direita, para analisar as palavras à direita da expressão em ordem alfabética; e também centralizada, conforme apresentada na Figura 4.

Figura 4 - Tela do SE exibindo as linhas de concordância com as ocorrências para o termo “água” e os colocados à sua esquerda.

A análise das linhas de concordância (Figura 4) revelou que, das 31 ocorrências de “consumo de água”, seis apresentam o co-ocorrente verbal “reduzir” à esquerda (“reduzir o consumo de água”) e cinco, o co-ocorrente nominal “redução” à esquerda (“redução do consumo de água”). Concluímos, portanto, que essas colocações devem ser consideradas na obra terminográfica.

Nesta seção, mostramos que a análise manual das listas de palavras simples e compostas, assim como o uso da ferramenta “Word Sketch” para validar as colocações e fraseologias características com os termos, permite um levantamento da terminologia padrão da área da sustentabilidade que não seria possível a partir da leitura sequencial dos textos que compõem o *corpus*. Em seguida, detalharemos o processo de busca de equivalentes funcionais para as entradas e subentradas em português.

Identificação de equivalentes funcionais

A busca pelos equivalentes procurou estabelecer uma correlação funcional por meio da comparação entre as listas de termos simples e compostos geradas pela ferramenta. A partir dos termos e fraseologias do português, identificamos termos e fraseologias em inglês que desempenhassem similar função nos respectivos textos. Essa análise foi realizada com o uso concomitante das ferramentas “Concordanciador” e “Word Sketch” a partir da lista de palavras-chave.

Tomemos como exemplo a palavra ‘água’, posicionada entre os primeiros termos simples levantados automaticamente. Não é necessário ser especialista na área da sustentabilidade para saber que seu equivalente *prima facie* em inglês é *water*, também uma das primeiras palavras-chave levantadas pelo utilitário “Keyword/Terms” do SE no sub*corpus* em inglês, excetuando-se aquelas que se encaixam nos critérios de exclusão. Contudo, a fim de identificar um equivalente funcional para a combinação “redução do consumo de água”, aplicamos a ferramenta “Word Sketch” para o núcleo “*water*”. Foram encontrados os seguintes dados, listados com as respectivas frequências (F) no sub*corpus* em inglês:

- “*water conservation*” (F = 43);
- “*water use*” (F = 39);
- “*water risk*” (F = 18);
- “*water stress*” (F = 17);
- “*water consumption*” (F = 17);
- “*water conservation awareness*” (F = 13);
- “*water usage*” (F = 12);
- “*water scarcity*” (F = 9);
- “*water stewardship program*” (F = 6);
- “(*manufacturing*) *water intensity*” (F = 6);

 “*reducing water*” (F = 6).

Nessa lista, observam-se as ocorrências de “*water use*”, “*water consumption*” e “*water usage*”. A priori, os três candidatos poderiam ser apresentados no verbete “água” como equivalentes da subentrada “consumo de água”, já que dicionários de língua geral encadeiam os substantivos “*use*”, “*usage*” e “*consumption*” de forma que o consulente os interpretaria como sinônimos (USE..., 2020; USAGE..., 2020; CONSUMPTION..., 2020). No entanto, no âmbito da Sustentabilidade, essas combinações não são intercambiáveis. Reig (2013) explica que “*water use*” e “*water usage*” são expressões usadas para descrever o total de água retirada de sua fonte a ser utilizada nos processos industriais, enquanto “*water consumption*” se refere à quantidade de água usada que não retorna à fonte original depois de ser retirada. O consumo ocorre quando há perda de água por evaporação ou pela incorporação a produtos (por exemplo, plantas), tornando-se indisponível para reúso. Com base nessa distinção, o equivalente mais adequado para “consumo de água”, portanto, é “*water consumption*”, já que uma obra terminográfica confiável da área deve considerar essa diferença.

No modelo de glossário que propomos, oferecemos ao consulente não só as entradas, formadas por termos simples, mas também as combinações usuais com esses termos – denominados Fras. (fraseologias) –, assim como seus equivalentes funcionais (Eq.), identificados em textos do mesmo gênero na língua de chegada. Incluímos, também, fraseologias maiores, ou seja, aquelas formadas por seus co-ocorrentes, a que denominamos COFras. Além disso, incluímos exemplos de uso, que auxiliam o redator e o tradutor na produção de textos convencionais na língua de chegada (FRANKENBERG-GARCIA, 2018). Quando julgado conveniente, o verbete inclui também remissivas, na forma de ‘Ver também’, a fim de complementar a compreensão do consulente, se assim desejar.

A Figura 5, a seguir, mostra um modelo de verbete, utilizando como exemplo para a entrada ‘água’.

Figura 5 - Modelo de entrada do glossário[11].

ÁGUA. Eq. ***WATER.*** **Fras.** consumo de água. **Eq. Fras.** *water consumption.* **Ex.** "*We have made great improvements in our ability to track* ***water consumption****, but this continues to be a challenge [...]*" (EN_CIS_RS2016.txt). **COFras.** redução do consumo de água. **Eq.COFras.** *reduce water consumption.* **Ex.** "*We have* ***reduced water consumption*** *through production efficiency improvements* [...]." (EN_ALL_RS2017.txt). **COFras.** reduzir o consumo de água. **Eq.COFras.** *reduce water consumption.* **Ex.** "[...] *we understand the importance of* ***reducing water consumption*** *as much as we can in our operations* [...]" (EN_CIS_RS2016.txt). Ver também: consumo de **energia**.

Considerações finais

Com relação à metodologia adotada, podemos afirmar que a LC exerceu um papel primordial na construção de um modelo de glossário de sustentabilidade, pois permitiu o levantamento de padrões a partir da análise de textos autênticos da área. Diante da grande oferta de ferramentas computacionais disponíveis para o processamento automático de textos, a escolha deve levar em consideração os propósitos da pesquisa a ser realizada e os benefícios que podem trazer, como reduzir o tempo de processamento e fornecer dados mais precisos. Para esta pesquisa, o SE mostrou-se adequado, já que possibilita a sistematização de colocados com a palavra de busca, de forma a auxiliar na seleção das entradas e subentradas do glossário.

Conforme mostramos aqui, para o tradutor ou redator de textos especializados, não basta uma lista de termos se estes estiverem descontextualizados, ainda que alinhados aos seus equivalentes na língua estrangeira. Além disso, a equivalência é bastante relativa, conforme mostrado no exemplo de "*use*" e "*usage*", em que a primeira expressão demonstrou ser muito mais utilizada na área do que a segunda. Isso só foi possível descobrir com o uso de um programa analítico. Por isso, defendemos o uso de ferramentas e programas que possam calcular as frequências (uma vez que seria impossível calcular as ocorrências em todos os textos através da leitura sequencial) concomitante a uma análise com base metodológica feita por especialistas em linguagens especializadas

[11] Siglas: Eq = equivalente; Eq. Fras. = Equivalente da fraseologia; Ex. = Exemplo; COFras. = Co-ocorrente da fraseologia.

Em relação ao tamanho, é preciso reconhecer que o desenho do nosso *corpus* de estudo apresentou certas limitações. Seria necessário aumentar o seu tamanho para conseguir identificar expressões com mais recorrência. A inclusão desse parâmetro nos obrigaria, portanto, a aumentar o *corpus*.

Ainda assim, em estudos com *corpora* especializados, todas as ocorrências dos itens de alta frequência podem ser examinadas, o que se mostra vantajoso; além disso, nesses estudos, fica mais evidente a relação entre o *corpus* e os contextos em que se inserem os textos (KOESTER, 2010). Os textos selecionados compreendem um recorte nesse âmbito de especialidade da Sustentabilidade. Como nosso propósito não foi incluir toda a gama de variações possíveis, esperamos poder oferecer um caminho viável e objetivo para a análise dos padrões da linguagem em Relatórios de Sustentabilidade.

Referências

BERBER SARDINHA, T. Lingüistica de Corpus. Barueri, SP: Manole, 2004.

BEVILACQUA, C. R. Por que e para que a Linguística de Corpus na Terminologia. In: Tagnin, S; Bevilacqua, C. *Corpora na Terminologia*. São Paulo: Hub Editorial. 2013. p. 11-27.

BOURIGAULT, D.; SLODZIAN, M. Por uma terminologia textual. *Cadernos de Tradução*, n. 17, Porto Alegre: IL/UFRGS, p. 101-108, 2004.

CIAPUSCIO, G. E. La terminología desde el punto de vista textual: selección, tratamiento y variación. *Organon*, Porto Alegre, n. 26, v. 12, p. 43-65, 1998.

CONSUMPTION. In: MERRIAM-WEBSTER Dictionary. Merriam-Webster, 2020. Disponível em: <https://www.merriam-webster.com/dictionary/consumption>. Acesso em: 16 mar. 2020

FRANKENBERG-GARCIA, A. Dicionários e exemplos de apoio à produção linguística. Trad. Luísa Rabaldo. *Cadernos de Tradução*, Porto Alegre, n. 43, jul./dez., p. 63-86, 2018.

GRI. GRI Standards Download Center. 2018. Disponível em: <https://www.globalreporting.org/standards/gri-standards-download-center/>. Acesso em: 25 fev. 2020.

KATAN, D. *Translating Cultures, An Introduction for Translators, Interpreters and Mediators*. Manchester, St. Jerome Publishing, 1999.

KILGARRIFF, A. et al. The Sketch Engine: ten years on. *Lexicography*, n. 1, 2014, p.

7-36.

KOESTER, A. Building small specialised corpora. In: O'Keeffe, A.; McCarthy, M. *The Routledge Handbook of Corpus Linguistics.* London/New York: Routledge, 2010, p. 66-79.

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

Merriam-Webster.com. 2020. Disponível em: https://www.merriam-webster.com. Acesso em: 16 Mar. 2020.

NAVARRO, S. *Glossário bilíngue de colocações de hotelaria: um modelo à luz da Linguística de Corpus*. 2011. Dissertação (Mestrado em Estudos Linguísticos e Literários em Inglês) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2012.

NORD, Christiane. *Quo vadis, functional translatology*? Amsterdam: Target, 2012, p. 26-42.

REBECHI, R. R.; SILVA, M. M. Brazilian recipes in Portuguese and English: the role of phraseology for translation. In: MITKOV, Ruslan. (Org.). *Computational and Corpus-based Phraseology*. Berlin: Springer, 2017, p. 102-114.

REBECHI, R. R.; SILVA, M. M. Obituaries in translation: a corpus-based study. *Cadernos de Tradução*, n. 3, v. 38, p. 238-318. Florianópolis: UFSC, 2018. Disponível em: <https://periodicos.ufsc.br/index.php/traducao/article/view/2175-7968.2018v38n3p298/37398>. Acesso em: 25 fev. 2020.

REIG, P. What's the Difference Between Water Use and Water Consumption? *World Resources Institute*, 12 mar. 2013. Disponível em: <https://www.wri.org/blog/2013/03/what-s-difference-between-water-use-and-water-consumption>. Acesso em: 25 fev. 2020.

SINCLAIR, J. *Trust the text*. London/New York: Routledge, 2004.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. In: TAGNIN, S. O.; BEVILACQUA, C. R. (org.) *Corpora na Terminologia*. São Paulo: HUB Editorial, p. 215-219, 2013.

USAGE. In: MERRIAM-WEBSTER Dictionary. Merriam-Webster, 2020. Disponível em: <https://www.merriam-webster.com/dictionary/usage>. Acesso em: 16 mar. 2020

USE. In: MERRIAM-WEBSTER Dictionary. Merriam-Webster, 2020. Disponível em: <https://www.merriam-webster.com/dictionary/use>. Acesso em: 16 mar. 2020.

Recebido em: 16 mar. 2020

Aceito em: 18 abr. 2020