



O uso de ferramentas computacionais em pesquisas terminológicas

Mirella de Souza Balestero¹
msbalestero@gmail.com
Clotilde de Almeida Azevedo Murakawa²
jtm.jau@uol.com.br

RESUMO:

Desde o avanço da Informática, as pesquisas em Terminologia passaram a ter a contribuição da tecnologia para a realização de suas tarefas. Nesse sentido, objetiva-se demonstrar, aqui, o uso de ferramentas computacionais em atividades terminológicas, já que cada etapa é executada de forma (semi)automática. Ademais, busca-se refletir sobre as contribuições e os entraves dessas ferramentas, fundamentando-se na Teoria Comunicativa de Terminologia, na Linguística Computacional e na Linguística de Corpus. Apresenta-se, então, as etapas de um trabalho em Terminologia baseadas na vertente terminológica descritiva e linguística, além de alguns dos programas que tornam as tarefas mais autênticas, precisas e eficazes.

PALAVRAS-CHAVE:

Terminologia;
Ferramentas
computacionais;
Processamento de
Língua Natural;
Linguística de *Corpus*;
Tarefas (semi)automatizadas.

¹ Doutoranda em Linguística e Língua Portuguesa pela UNESP/FCLAr (Bolsista Univesp).

² Doutora em Letras, Linguística e Língua Portuguesa pela UNESP/FCLAr.

1 Introdução

Antes de falar sobre Terminologia, faz-se relevante mencionar o surgimento da Informática, pois data da década de 60 as primeiras manifestações de sistematização cibernética. No entanto, é apenas na década de 90 que se dá o avanço da Informática, o que, posteriormente, iria colaborar, com mais ênfase, para os estudos em léxico (geral ou especializado).

Sabemos que a tecnologia, sobretudo a internet, contribuiu e ainda contribui muito para o nosso cotidiano, facilitando algumas atividades (como falar com um parente ou amigo em outro país), além de possibilitar uma melhoria em diversas áreas de trabalho e do conhecimento, tais como a Medicina, a Engenharia, a Arquitetura etc.

Assim como nas exatas e nas áreas da saúde, a Informática tem um papel fundamental nas humanidades, especialmente nos estudos de língua e linguagem. Desse modo, traremos algumas contribuições de ferramentas (semi)automatizadas em uma das linhas de pesquisa da Linguística, denominada Estudos do Léxico – mais especificamente, para a Terminologia.

Em países mais desenvolvidos, a Terminologia e a Informática estão unidas desde sempre, “de forma a facilitar o armazenamento e a difusão de dados terminológicos na elaboração de grandes bases de dados especializados” (ALMEIDA, 2006, p. 42). É a partir dessa integração que surge o termo Terminótica, que é a junção entre *Terminologia* e *Informática*. Já no Brasil, essa relação começa a aparecer muito tempo depois, nos anos 90.

Na medida em que os pesquisadores – sobretudo os linguistas – começam a dar maior relevância para a Informática, passa-se a tratar, de forma automática (ou semiautomática), o português brasileiro. Dessa forma, as pesquisas em Léxico (Lexicologia, Lexicografia, Terminologia e Terminografia) aumentam e visam, cada vez mais, tratar computacionalmente as unidades lexicais e unidades terminológicas.

Isso porque a tecnologia dispõe de procedimentos que interferem nas atividades do terminólogo, auxiliando suas atividades ao tornar mais rápida a obtenção de resultados, por exemplo, além de obras terminográficas mais precisas e autênticas (BALESTERO, 2019, p. 20).

Para discutirmos, então, o uso das ferramentas computacionais em pesquisas terminológicas, partimos de pressupostos teórico-metodológicos da Terminologia, principalmente a descritiva de base linguística, embora reconheçamos as contribuições dos pesquisadores prescritivistas, como Wüster. Além disso, valemo-nos dos aparatos da Linguística Computacional (LC), área da Linguística que compreende o tratamento computacional de uma língua natural e que envolve profissionais da Linguística e da Informática.

A Linguística Computacional começou a ser desenvolvida nos anos 50, cuja contribuição está diretamente relacionada à Inteligência Artificial. A partir disso, a área passou a ser impulsionada pelo desenvolvimento de programas de tradução automática.

De acordo com Othero (2006), no artigo *Linguística Computacional: uma breve introdução*, a Linguística Computacional pode ser dividida em dois campos: Linguística de Corpus e Processamento de Língua Natural (PLN). Nesse contexto, nota-se que uma pesquisa terminológica pode se pautar em ambas as áreas, haja vista que o objetivo consiste em, respectivamente, organizar e/ou analisar amostras da língua com alguma finalidade (geralmente, para se estudar algum fenômeno linguístico) e construir programas capazes de identificar, produzir e processar a língua natural.

Além dessa perspectiva, pode-se compreender a Linguística Computacional como sinônimo de Processamento de Língua Natural, configurando-se como uma área multidisciplinar, na qual envolve Linguística, Informática, Inteligência Artificial, além de produzir e utilizar sistemas que reconheçam e tratam a língua natural. Nessa configuração, a Linguística de Corpus seria uma subdivisão do campo, isto é, uma das possibilidades de estudo da LC. Ademais, ela – a Linguística de Corpus – não tem o objetivo de elaborar, diretamente, programas, mas é responsável por realizar atividades de compilação e tratamento de *corpus* (ou *corpora*) por meio do computador.

Neste artigo, portanto, baseamo-nos no conceito apontado por Othero (2006). Sendo assim, buscamos discorrer sobre a importância da Informática nos trabalhos terminológicos e apresentar algumas tarefas e ferramentas que exemplificam seu uso a partir dos fundamentos da Linguística Computacional – entendendo essa área como um domínio maior, que abarca, em geral, os estudos relacionados à Linguística e à Informática –, especialmente os princípios da Linguística de Corpus.

Com ambas as perspectivas, Terminologia e LC, as etapas de um trabalho terminológico podem tornar-se mais eficientes, haja vista que as tarefas serão executadas com o auxílio de ferramentas computacionais e podem facilitar a elaboração de softwares.

2 Embasamentos teórico-metodológicos

2.1 A prática lexicográfica e terminográfica

Na tentativa de se comunicar com outros povos e compreender diversos idiomas, surge a prática lexicográfica. Além desse fator, tem-se a motivação escolar, uma vez que a emergência da escrita suscitou a necessidade da elaboração de materiais (vocabulários e glossários) para o ensino.

Essa prática se deve, principalmente, à necessidade de nomear as coisas, os objetos, as ações, as emoções, etc. Ademais, a necessidade comunicativa configura-se como uma das

maiores motivações do advento dos Estudos do Léxico em geral: Lexicologia, Lexicografia, Terminologia e Terminografia. Nesse sentido, assim como a Lexicologia, a Terminologia é a responsável por nomear as unidades, mas, diferente da primeira, tem foco nos contextos especializados, isto é, visa sistematizar os termos de domínios de especialidade.

Em outras palavras, a Terminologia é a ciência que estuda os termos, ou seja, as unidades lexicais contextualizadas no discurso especializado (DOMÉNECH, 2014, p. 31). Seu principal objetivo é “dar conta do funcionamento das unidades lexicais especializadas em situações comunicativas profissionais, acadêmicas ou científicas” (LORENTE, 2004), de modo que essa comunicação “se realize de forma compreensível e sem ambiguidades em ambientes mono e/ou multilíngues.” (TERMISUL, 2021).

Independente de não sabermos o surgimento dos Estudos do Léxico, sempre foi uma realidade do ser humano organizar e sistematizar as unidades de uma língua. Em vista disso, as atividades lexicográficas e terminográficas são bastante antigas e fundamentais para a língua.

Há apenas uma previsão sobre a prática terminológica a partir dos trabalhos de Lavoisier e Berthold, e de Linné. Seus estudos, voltados para a química, e para a botânica e a zoologia, respectivamente, são um marco para as primeiras reflexões sobre a necessidade de se fazer terminologia (ALMEIDA, 2003).

É possível mencionar, no entanto, que, a princípio, a prática terminológica era feita com base em pressupostos prescritivos, a partir da Teoria da Terminologia Geral (TGT), cunhada por Wüster – um engenheiro austríaco que começou a sistematizar sua área de trabalho na intenção de normalizá-la/padronizá-la. O autor defendeu sua tese em 1930, mas foi apenas em 1970 que ela foi transformada em teoria, buscando erradicar a ambiguidade na sua profissão.

Neste artigo, apesar de não trabalharmos com uma proposta terminológica normativa – que segue imposições de língua, como a padronização dos vocabulários e dos conceitos, sem levar em consideração a variação como um fenômeno inerente à língua –, reconhecemos as contribuições de Wüster na área. Dentre elas, está o conceito de Terminologia para o autor, que a considera uma matéria interdisciplinar, cujas disciplinas envolvidas são a Linguística, a Lógica e a Informática (CABRÉ, 2002).

Em contrapartida com a TGT, surgem as teorias terminológicas de viés descritivo a partir da década de 90, a saber: Socioterminologia, Terminologia Cognitiva, Teoria Comunicativa da Terminologia, Terminologia Textual, entre outras. Nesse contexto, muda-se a prática terminológica ao admitir a variação conceitual e denominativa nos discursos especializados (CABRÉ, 1999).

Por conta disso, valemo-nos da Teoria Comunicativa da Terminologia (TCT) neste trabalho, já que se trata de uma vertente que leva em consideração o funcionamento da língua e a descrição de fenômenos linguísticos. Essa descrição pode se tornar ainda mais

eficaz e eficiente a partir do uso de ferramentas (semi)automatizadas, como mostraremos adiante. Com isso, faz-se fundamental a contribuição da Linguística Computacional, e, particularmente, da Linguística de *Corpus*, nas tarefas do terminólogo, que interferem diretamente na compilação, na organização e na análise dos dados.

As pesquisas em Terminologia seguem algumas etapas (ou atividades), que podem ser reconhecidas como os métodos da pesquisa a serem seguidos. Tais métodos incluem a interação entre ser humano e máquina na medida em que considera os fundamentos práticos das teorias anteriormente mencionadas. Assim sendo, quando falamos de tarefas automatizadas, estamos nos referindo a programas computacionais específicos que auxiliam as atividades), e, ao mencionar tarefas (semi)automatizadas, queremos dizer que algumas tarefas são feitas manualmente, ou seja, pelo terminólogo, profissional que possui conhecimentos aprofundados sobre a língua e, por vezes, sobre a língua especializada³.

Logo, as principais tarefas da atividade terminológica são:

Quadro 1 – Etapas de uma pesquisa terminológica

delimitação do domínio de especialidade
compilação do <i>corpus</i>
seleção, conversão e limpeza do <i>corpus</i>
extração e seleção dos termos
elaboração da estrutura conceitual
incremento da base definicional
elaboração e preenchimento das fichas terminológicas
redação das definições

Fonte: elaborado pelos autores

Dessas tarefas, podem ser elaboradas obras terminográficas, tais como dicionários, glossários, vocabulários, e alguns produtos – ontologias, fichas terminológicas, verbetes e, até mesmo, *corpus/corpora*. À vista disso, é fulcrar ressaltar que cada etapa se configura como um processo que gera dados e, conseqüentemente, resultados. Ademais, dentro de cada etapa constam atividades mais específicas, como o tratamento do *corpus*, que pode incluir a anotação dos textos, a análise dos *corpora*, entre outras.

Nesse sentido, vale a reflexão: se hoje é preciso levar em consideração o tempo de elaboração da pesquisa, os objetivos do trabalho e o tempo de aprendizagem de ferramentas que auxiliem o trabalho em Terminologia, a dificuldade em compilar e analisar dados de determinado campo antes das contribuições da Linguística Computacional era muito maior. Portanto, é inimaginável, atualmente, o desenvolvimento de pesquisas terminológicas (em léxico, em geral), sem o auxílio de computadores.

Já a atuação da Linguística de *Corpus* começa a aparecer nos anos 60, na Universidade de Brown, nos Estados Unidos, com o primeiro *corpus* linguístico, chamado

³ Não nos referimos à língua especializada como um tipo de língua, separada do léxico geral, visão defendido por Wüster. Na verdade, defendemos a ideia de Cabré de que existe uma língua e, em determinados contextos e situações, ativamos algumas unidades – que podem ser gerais ou especializadas.

de *Brown Corpus*, compilado por Henry Kučera e W. Nelson Francis. Logo depois, Sinclair torna-se responsável pelo “trabalho pioneiro na área do léxico que traçou os caminhos da maioria da pesquisa em Linguística de *Corpus* feita até hoje” (SARDINHA, 2000, p. 332). Porém, é apenas em 1987 que se tem a elaboração, também por Sinclair, do primeiro dicionário feito a partir de um *corpus* (computadorizado), denominado COBUILD.

No Brasil, a contribuição da área iniciou-se apenas na década de 90 (OLIVEIRA, 2009b) e ainda é pouca a visibilidade da Linguística de *Corpus* nas pesquisas, sobretudo ao ser considerada, por alguns, apenas como uma abordagem metodológica. A aplicação do campo, no entanto, se dá por um dos maiores linguistas de *corpus* do país, Berber Sardinha. O livro do autor, publicado em 2004, é o primeiro a estudar a abordagem de coleta e tratamento de texto a partir de *corpus*. Sendo assim, é a partir dessa obra que a área passa a ter maior alcance.

A Linguística de *Corpus*, com o auxílio de ferramentas computacionais, proporciona o melhor tratamento do português brasileiro. Nesse contexto, as tarefas referentes a essa abordagem são facilitadas pelo uso de alguns programas no computador para a descrição do léxico, geral ou especializado. No caso desse trabalho, pontuamos que as ferramentas (semi)automatizadas agilizam as atividades terminológicas na medida em que auxiliam desde a elaboração do *corpus* até a organização de um glossário, por exemplo.

Por fim, é importante destacar que a Linguística Computacional (com os fundamentos da Linguística de *Corpus* e auxílio de Informática) é uma área diretamente ligada à Terminologia, pois corrobora com a prática terminológica, facilitando a execução das tarefas mencionadas no quadro 1. A seguir, nos aprofundaremos nessas questões.

2.2 A Linguística de *Corpus* e os programas (semi)automatizados

A Linguística contém diversas ramificações. Uma delas é a parte de descrição da língua, que pode ser feita por meio de atividades linguísticas com o auxílio de ferramentas (semi)automatizadas. Com base nisso, temos ciências que auxiliam os trabalhos na área, sobretudo em Estudos do Léxico, tal qual a Linguística de *Corpus*.

De acordo com a literatura, a Linguística de *Corpus* faz parte da Linguística Computacional, assim como o Processamento de Língua Natural (PLN) (OTHERO (2006). Ambas possuem contribuições diretas nas pesquisas, uma vez que se pretende organizar e/ou analisar amostras da língua para alguma finalidade específica. Contudo, há uma diferença entre as subáreas da Linguística Computacional. Primeiramente, na Linguística de *Corpus*, o objetivo é, em geral, organizar textos e identificar e/ou analisar determinado fenômeno linguístico. Já no PLN, tem-se a intenção de construir programas para produzir e processar a língua natural.

Em algumas leituras, pode-se referir à Linguística Computacional como sinônimo de PLN, em que a Linguística de *Corpus* seria um sub-campo, ou seja, uma das diversas áreas de estudo para embasar a pesquisa no campo. No entanto, como a intenção nos trabalhos terminológicos, em sua maioria, não é elaborar novos softwares, focalizaremos a discussão, neste tópico, na Linguística de *Corpus*.

Julga-se importante mencionar que, no site da Abralín⁴, por exemplo, tem-se a Linguística Computacional e a Linguística de *Corpus* como subáreas da Linguística, apenas. Nesse local, estão disponíveis algumas *lives* relacionadas aos campos mencionados.

Ainda pode-se verificar uma outra perspectiva. Quanto à Linguística Computacional (LC), esta se relacionaria à Linguística de *Corpus* por ambas basearem-se no *corpus* para buscar evidências linguísticas, por suas características ligadas à tecnologia, e por focalizarem o uso de linguagem em seus estudos linguísticos. Entretanto, seus objetivos são diferentes, já que a Linguística Computacional:

explora relações entre as áreas de linguística e informática, tornando possível a construção de sistemas com a capacidade de reconhecer e produzir informação apresentada em língua natural (VIEIRA; LIMA, 2001).

Não obstante, reiteramos que, para fins didáticos, amparamo-nos com a divisão feita por Othero (2006):



Cabe dizer, também, que a LC auxilia os trabalhos em Linguística em geral, já que ela dispõe de aparatos tecnológicos que colaboram com as etapas de compilação e tratamento de um *corpus* (objetivos da Linguística de *Corpus*), que podem ser utilizados para qualquer pesquisa. Nesse contexto, vale mencionar que as ciências funcionam como uma via de mão dupla, independente do conceito de cada uma delas – qual o termo hiperônimo e qual o hipônimo, isto é, se uma área pertence ou abarca a outra.

Em outras palavras, como os trabalhos, nos últimos anos, estão sendo elaborados a partir dos princípios teórico-metodológicos da Linguística de *Corpus*, aumentou-se “a construção de programas capazes de interpretar e/ou gerar informações em linguagem natural” (OLIVEIRA, 2009b, p. 55).

Assim, a área da Linguística Computacional também precisou de um olhar mais atento, dado que é a partir das ferramentas de computador que os *corpus* e os *corpora* podem ser tratados com maior eficiência, valendo-se de uma grande quantidade de textos

⁴ Para saber mais, acesse <https://aovivo.abralin.org/lives-por-subarea/linguistica-computacional-corpus/>.

na língua natural para ser estudada. Nesse sentido, afirma-se que a Linguística Computacional é, em geral, uma disciplina resultante da união entre Linguística e Computação que “cria instrumentos de composição, extração, análise, manipulação e processamento de dados linguísticos” (BARROS, 2006, p. 25).

Além disso, determinadas vertentes podem considerar a Linguística de *Corpus* apenas como uma metodologia. Conforme Sardinha (2004), essa visão defende que os métodos podem ser aplicados em vários trabalhos, sem mudar sua orientação teórica. Outras, no entanto, já veem a Linguística de *Corpus* como uma teoria, uma vez que se produz conhecimentos novos. E ainda existe um outro grupo que a considera uma perspectiva teórico-metodológica – visão que tomaremos como base para este trabalho.

Na realidade, pretende-se, com a Linguística de *Corpus*, estudar a língua em uso, isto é, em funcionamento. Por essa razão, fala-se em língua natural, e não em língua artificial. Ademais, por se tratar de grandes quantidades de amostras de textos, os cientistas da linguagem, neste caso, os terminólogos, recorrem às ferramentas informatizadas para descrever um fenômeno linguístico, ou seja, para observar estatisticamente um fenômeno e identificar padrões de uso entre palavras (FINATTO; LOPES; CIULLA, 2015, p. 43).

Consoante a Finatto, Lopes e Ciulla (2015, p. 47):

Enquanto linguistas, nesse trabalho associado ao PLN, valemo-nos de informações recuperadas pelos sistemas automáticos para nossas análises e descrições da língua, especialmente em se tratando de um grande volume de dados, e contribuímos para a melhoria desses sistemas, com a consideração de regras linguísticas que possam ser a eles integradas

As autoras ainda afirmam que primeiro é preciso analisar os fenômenos linguísticos e, posteriormente, as ferramentas computacionais poderão ser desenvolvidas e usadas de forma adequada (FINATTO; LOPES; CIULLA, 2015, p. 44). Outrossim, com relação às pesquisas em Terminologia, vale mencionar que essas ferramentas podem ser úteis ao extrair informações de maneira rápida, como itens lexicais que podem se configurar como termos de um domínio de especialidade (FINATTO; LOPES; CIULLA, 2015, p. 56).

No início de sua descoberta, a Linguística de *Corpus* era entendida como o conjunto de textos impressos (ou digitais, o que era mais raro na época). Entretanto, com o avanço da Informática, começou a ser pouco frequente o uso de textos impressos, haja vista que a internet tornou mais fácil o acesso a materiais on-line. Por conta disso, alguns referenciais teóricos podem apresentar, ainda, concepções não muito adequadas para o tempo em que vivemos, posto que podem restringir a compilação de *corpus* sem textos digitais.

Seguindo nessa discussão, são muitas as definições encontradas para *corpus*, variando de autor para autor. A título de exemplo, os mais consagrados na literatura são: Kennedy (1998), Renouf (1998), Biber et al (1998), Berber Sardinha (2004) e Sinclair (2005).

Em geral, entende-se que *corpus* é um compilado de textos que segue critérios pré-determinados para alguma finalidade de investigação. Nas palavras de Berber Sardinha (2004), por exemplo, um conjunto de textos nem sempre é considerado um *corpus*, já que este precisa ter uma organização prévia, ou seja, obedecer a certos critérios de seleção, coleta, organização e nomeação.

Na mesma direção, Sinclair (2005, p. 23) defende que:

Um *corpus* é uma coleção de amostras da língua em formato eletrônico, selecionados de acordo com critérios externos para representar, na medida do possível, uma língua ou uma variedade da língua como uma fonte de dados para uma pesquisa linguística.⁵

Há, também, uma variação no conceito de acordo com a fundamentação teórica: Linguística ou Linguística de *Corpus*. A seguir, temos um quadro comparativo das definições de *corpus* de acordo com a perspectiva da Linguística, uma vez que já apresentamos o conceito conforme a segunda vertente.

Quadro 2 – Conceito de *corpus* para a Linguística

conjunto finito de enunciados tomados como objeto de Análise. Trata-se, pois, de uma coleção de documentos quer orais (gravados ou transcritos) quer escritos, quer orais e escritos, de acordo com o tipo de investigação pretendido (GALISSON; COSTE, 1983).
--

conjunto de enunciados a partir do qual se estabelece a gramática descritiva de uma língua (DUBOIS et al. 1993).
--

Fonte: elaborado pelos autores e adaptado de Aluísio e Almeida (2006).

De acordo com as definições mencionados no quadro, nota-se que a diferença principal entre a concepção de *corpus* para a Linguística e para a Linguística de *Corpus* é o formato do texto, ou seja, impressos para a primeira e eletrônicos para a segunda. Contudo, como já discutimos, os textos impressos podem ser considerados, atualmente, para a compilação de um *corpus* na medida em que digitalizamos cada um deles e passamos para o formato digital.

Tomando, então, a Linguística de *Corpus* como ponto de partida, é importante ressaltar as etapas de construção de um *corpus*. Dessa forma, baseado em autores da área, como Berber Sardinha e Sinclair, Souza e Di-Felippo (2010) apresentam as etapas que abarcam a elaboração de um *corpus* ou *corpora*: a) projeção do *corpus*, que consiste na definição do tipo de *corpus* necessário à pesquisa; b) compilação; c) pré-processamento (conversão, limpeza, nomeação, anotação e armazenamento); e d) disponibilização.

Para realizar essas tarefas, deve-se seguir os critérios da Terminologia de base descritiva, a TCT, a fim de elaborar produtos terminográficos – glossários, dicionários – eficientes (ALMEIDA, 2006). Na faceta da Terminologia, busca-se, principalmente, dados reais (baseados no uso), que devem ser elaborados a partir dos parâmetros da Linguística de *Corpus* para que as amostras de textos sejam o mais eficaz possível. São eles:

⁵ “A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.”

Quadro 3 – Parâmetros para elaborar um corpus computadorizado

Autenticidade	escritos em língua natural
Representatividade	conter as características necessárias do projeto de acordo com a língua ou com a variedade
Balanceamento	equilíbrio de gêneros
Amostragem	conter amostra suficiente da língua ou variedade analisada
Diversidade	conter diferentes gêneros de textos
Tamanho	pode variar em tamanho de acordo com o tipo de pesquisa

Fonte: elaborado pelas autoras, baseado em Aluísio e Almeida (2006)

Faz-se relevante mencionar que, em alguns trabalhos, os parâmetros apresentados no quadro acima podem ser distintos. Domínios de especialidades novos ou pouco sistematizados, por exemplo, terão o fator ‘representatividade’ diferente de outras áreas. Com isso, *corpus* que contenham aproximadamente 300 textos pode significar muito para uma pesquisa terminológica devido à falta de obras de referências. Já em outros campos, essa quantia pode ser considerada insignificante para descrever um fenômeno, pois não simboliza, de forma real, o funcionamento do domínio.

Outro caso está relacionado com a autenticidade. Dito isso, o que seria um texto autêntico? Seriam considerados legitimados para a pesquisa apenas artigos científicos e pesquisas acadêmicas em geral? Diante dessas questões, vê-se que o terminólogo é quem dirá se uma fonte pode ou não ser válida para a pesquisa, dado que, em determinados domínios, os diferentes gêneros e meios de circulação abarcam dados relevantes e consistentes – além de autênticos – para o trabalho.

Para finalizar a discussão sobre a Linguística Computacional e a Linguística de *Corpus* e sua importância para o contexto do artigo, é válido pontuar que daremos ênfase ao caráter linguístico dos exemplos dados, já que não se trata de uma discussão direcionada à Computação, mas sim em reconhecer as contribuições das ferramentas computacionais para os trabalhos em Linguística.

3 Prática: tarefas terminológicas (semi)automatizadas

Em uma pesquisa terminológica, é necessário delimitar o domínio de especialidade que irá ser estudado, pois:

Ainda que a tecnologia auxilie nas tarefas terminológicas, é necessário restringir a descrição e a sistematização de determinada área, limitando a elaboração do glossário a um domínio ou subdomínio de especialidade. Isso ocorre devido à extensão do léxico e, portanto, do léxico especializado, já que é impossível registrar todas as unidades em um dicionário devido à complexidade de uma área (BALESTERO, 2019, p. 20).

Para desenvolvermos, então, um trabalho em Terminologia, baseamo-nos em fundamentos da Linguística de *Corpus*. Além disso, dado os avanços na tecnologia, outra teoria serve de base para a pesquisa: a Linguística Computacional. Esta consiste, como já discutido, em um campo que utiliza processos e ferramentas computacionais em trabalhos de linguística, especialmente em pesquisas do léxico.

Sendo assim, percebe-se, ao longo dos últimos dez anos, uma crescente nos trabalhos em Terminologia com o auxílio das ciências mencionadas, pois elas interferem diretamente nas atividades do terminólogo, uma vez que torna mais rápida a coleta dos dados e a obtenção dos resultados, das análises e, conseqüentemente, contribui para a produção de obras terminográficas mais precisas (BALESTERO, 2019).

Igualmente às perspectivas anteriores, na visão de Cabré, a internet é uma vertente que possibilita o uso de tarefas que auxiliam no trabalho com unidades terminológicas, como acesso a dicionários *online*, acesso a *corpora* digitalizados, entre outras. Contudo, há um lado da internet que pode ser adverso, isto é, a grande disponibilidade que esse meio oferece ao indivíduo pode fazê-lo se perder na coleta de dados, bem como no momento de filtrá-los. Isso porque o pesquisador precisa utilizar critérios para selecionar bem os dados (CABRÉ, 2004, p. 18).

Ainda sobre a relevância da internet nos trabalhos terminológicos e das ferramentas nela contidas, Barros (2004, p. 23) afirma que:

De fato, as ferramentas informáticas acompanham todo o trabalho terminológico desde a constituição de *corpora* e seu armazenamento em bancos de dados textuais até a elaboração de dicionários eletrônicos, passando pela recolha e tratamento dos dados.

Partindo, então, dos pressupostos da Teoria Comunicativa da Terminologia (TCT), ressaltamos que essa escolha teórica, segundo Almeida (2006), “implica em acolher determinadas escolhas metodológicas durante todas as etapas de construção de um produto terminológico, seja ele um glossário, dicionário, mapa conceitual, listas de termos ou mesmo uma base de dados.”

Nesse viés, desenvolver pesquisas terminológicas é estudar um domínio de especialidade, sistematizando-o e descrevendo-o. Dessa forma, faz-se necessário, primeiro, selecionar uma área para ser analisada. Alguns exemplos de domínios são: Doenças Respiratórias, Esportes, Editoração, Provas de Proficiência, Peças da Engenharia Naval, entre outras. Em suma, todos os campos podem ser sistematizados, tudo depende do objetivo da pesquisa e do tempo que se tem. Assim, pode-se selecionar uma área mais ou menos ampla, como a Agropecuária ou o Café, respectivamente.

Para realizar a pesquisa terminológica, então, são seguidas algumas etapas em conformidade com a TCT. Dentre as tarefas, pode-se relatar:

- seleção de um domínio de especialidade;
- definição dos objetivos da pesquisa;
- elaboração de um *corpus* da área;
- extração de candidatos a termos;
- seleção dos termos do campo;
- validação dos termos pelo especialista;
- elaboração de uma estrutura conceitual (ou ontologia);
- incremento de uma base definicional para cada termo;
- elaboração de uma ficha terminológica;
- preenchimento das fichas terminológicas;
- elaboração das definições terminológicas;
- análise das definições e, conseqüentemente,
- elaboração do verbete (e do glossário).

Para nos auxiliar na organização e na sistematização dos dados, fundamentamo-nos na Linguística de *Corpus*. Além disso, essas e outras atividades podem ser facilitadas com o uso de ferramentas computacionais, como defende a Linguística Computacional.

Diante das tarefas que o pesquisador precisa executar, é imprescindível pensar em alguns critérios a fim de alcançar os objetivos da pesquisa com maior eficiência. Em primeiro lugar, o profissional precisa delimitar a área que será trabalhada, e isso pode ser feito principalmente de duas formas: (i) intuição do terminólogo de acordo com sua percepção da necessidade de sistematização do campo e (ii) interesse e necessidade dos especialistas em reorganizar o domínio. Depois, é o momento para estruturar o projeto, delimitando os objetivos da pesquisa.

Igualmente, deve-se definir, por exemplo, de que forma os textos serão coletados e organizados. Nessa perspectiva, são muitos os meios para compilar um *corpus*, como a busca de artigos científicos em buscadores do Google, na *World Wide Web*, por meio de palavras-chave do campo, o escaneamento de textos impressos de forma que eles sejam digitalizados, a identificação e seleção de textos acadêmicos em repositórios de universidades brasileiras, etc.

Em geral, busca-se textos na *Web*, porque, segundo Berber Sardinha (2003, p. 195), trata-se de uma “uma coleção de arquivos de computador em rede, que faz parte da Internet”, dispondo de um acesso facilitado aos materiais que já estão em formato eletrônico e, muitas vezes, gratuitos. Em algumas pesquisas, o *corpus* é compilado automaticamente. Um exemplo é o programa e-Termos, que possui, em uma das etapas, a possibilidade de buscar textos a partir de filtros, como palavras-chave, idioma, quantidade de palavras, entre outros exemplos.

O e-Termos é um ambiente colaborativo gratuito que, respaldado nos fundamentos da orientação descritiva e linguística da Terminologia, auxilia as atividades terminológicas. Além disso, diz respeito a uma ferramenta computacional, elaborada por linguistas e profissionais da computação pela Embrapa Informática Agropecuária (CNPTIA) em parceria com a UFSCar e a USP-São Carlos, cujo propósito é descrever, organizar, sistematizar e analisar um domínio de forma semiautomática (ALMEIDA; OLIVEIRA; ALUÍSIO, 2006, p. 70).

Selecionados os textos, é hora de organizá-los, convertê-los, nomeá-los e tratá-los. Para tal, há algumas ferramentas (gratuitas e pagas) que convertem os textos em diferentes formatos. Alguns deles respeitam mais ou menos as configurações do arquivo original, mantendo, por exemplo, notas de rodapé, figuras, etc. Entretanto, existem

programas, como o Abbyy, que são pagos e não mantêm, nos arquivos convertidos (geralmente em textos editáveis), todas as informações do primeiro texto.

Por esses motivos, é importante que o terminólogo teste alguns programas e veja qual é o mais adequado para a sua pesquisa. Vejamos, a seguir, descrições de algumas ferramentas já testadas:

Quadro 4 – Ferramentas de Conversão

Programas	Vantagens	Desvantagens
Abbyy	Converte vários textos ao mesmo tempo.	Pago. Precisa baixar o programa. Não converte o texto de forma fiel. Corta algumas partes no texto convertido.
OCR online	Gratuito e rápido. Mantém as informações no local correto (exatamente como no original).	Converte uma página por vez.
Calibri	Gratuito. Converte o PDF inteiro.	Quebrou as linhas na conversão.
Convertio	Gratuito e rápido. Converte tudo sem cortar palavras, exatamente como no original.	Limite de 10 minutos por dia (mas, na verdade, o limite também é referente a 10 textos). Alguns textos maiores, o programa leva mais tempo para converter.
NewOCR	Gratuito e rápido.	Converte uma página por vez e 10 textos por dia. Dificuldade em converter, no lugar correto, as notas de rodapé, misturando-as.
SmallPDF	Gratuito e rápido. Converte todas as informações do original.	De forma gratuita, converte apenas para word.
Online Convert	Gratuito e rápido. Converte todas as informações do original. Converte o texto que está em imagens.	Alguns arquivos não são convertidos. Ao fazer o download, o documento está em branco.
Zamzar	Gratuito e rápido. Converte todas as informações do original. É possível converter mais de um arquivo ao mesmo tempo.	Duas conversões gratuitas por dia.

Fonte: elaborado pelas autoras

Convém lembrar que, embora as ferramentas auxiliem na tarefa de conversão de um texto, ainda é necessário que o terminólogo analise os dados gerados, haja vista que algumas conversões podem apresentar erros. Na verdade, pode ocorrer menos ou mais falhas no processo, o que implica maior ou menor nível de dificuldade para limpar os textos. Nessa perspectiva, alguns desvios são facilmente corrigidos, comparando o arquivo convertido com o original.

Na etapa seguinte, de limpeza do *corpus*, são corrigidas as falhas de conversão cometidas na etapa anterior, além de filtradas as informações que foram delimitadas pelo pesquisador. Assim, tudo aquilo que não é relevante para a pesquisa é filtrado, ou seja, identificado e removido. Para isso, empregam-se ferramentas computacionais disponíveis,

como editores de textos (*word, libreoffice, bloco de notas, notepad++, editpad*) e expressões *regex* – expressões que identificam um padrão no texto.

Em suma, trata-se de uma tarefa demorada e, muitas vezes, cansativa, dado que o terminólogo precisa corrigir os erros para que eles não prejudiquem as atividades seguintes. Neste momento, então, começa-se a realizar o tratamento do *corpus*, que abarca diversas atividades, tais quais limpar, filtrar, anotar, nomear, etc.

A título de exemplificação, pode-se pensar em uma pesquisa na qual o objetivo é identificar os termos da Covid-19. Nesse domínio de especialidade, é sabido que algumas unidades apresentam números (códigos e fórmulas), por isso, não é correto eliminar do *corpus* qualquer tipo de numeração. Já em outros trabalhos, como a variação dos termos da economia, é um pouco mais raro o uso de números. Portanto, pode-se usar *regex* para remover esse elemento dos textos.

Uma outra tarefa é selecionar os termos da área. Para isso, precisamos realizar duas ações, a saber: extrair uma lista de candidatos a termos automaticamente e validá-las pelo especialista. Antes disso, porém, o terminólogo faz uma limpeza, retirando as unidades que, segundo a sua intuição e o seu conhecimento, não pertencem ao domínio. Assim, fica mais fácil a validação por parte do profissional da área.

Em adição a essas funções, o terminólogo pode levar em consideração a frequência das unidades terminológicas no *corpus* que foi compilado. Tem-se, desse modo, informações mais precisas sobre a relevância da unidade para determinado campo. No entanto, é importante pontuar que, em alguns casos, os termos de alguns domínios, como a Revisão de Textos, apresentam uma baixa frequência de uso em razão da pouca visibilidade que a profissão tem e das raras obras de referência que discutam a nomenclatura da área (sobretudo em materiais digitalizados).

Consideramos relevantes, antes de prosseguir com as etapas terminológicas, apresentar alguns programas que corroboram para que essas atividades sejam mais eficientes. A maior parte delas é voltada para gerar e analisar listas de frequência de determinado texto, como, por exemplo, e-Termos, WordSmith Tools e AntConc.

Nesses programas, é necessário subir o arquivo final do *corpus*, isto é, junta-se todos os textos selecionados e tratados e um arquivo zipado é formado, o qual será analisado posteriormente. A partir daí, utiliza-se alguma das ferramentas apresentadas anteriormente para extrair uma lista de candidatos a termos. Depois, os termos podem ser categorizados e organizados de acordo com sua frequência, seu contexto de uso, sua posição, entre outros critérios.

Importa destacar que o e-Termos é uma das ferramentas mais completas para pesquisas terminológicas, uma vez que foi elaborada para tal. Além disso, ele dispõe de todas as tarefas, desde a compilação do *corpus* até a divulgação dos verbetes. Já o

WordSmith Tools, é um programa pago, que disponibiliza apenas alguns recursos de uma pesquisa que se baseia em Linguística de *Corpus*, como o *concord*, a *wordlist* e a *keyword*.

Com relação ao AntConc, é importante dizer que ele auxilia as tarefas terminológicas na medida em que contém lista de concordanciador, lista de palavras, palavras-chave, entre outras ferramentas. Nota-se, portanto, que o AntConc e o WordSmith Tools, assim como outros programas, não têm maior ou menor valor comparado com o e-Termos, por exemplo, porém, este se configura como um ambiente colaborativo gratuito no qual visa atender a todas (ou quase todas) as necessidades do terminólogo.

Dessa forma, em alguns casos, mostra-se mais adequado usar uma ou outra ferramenta computacional, ou até mais de uma. Em resumo, não há uma regra para isso, e nem um software melhor que o outro. Neste trabalho, a ideia é apenas apresentar os possíveis programas e discorrer sobre suas contribuições em pesquisas terminológicas.

Dando continuidade às etapas da pesquisa, aproveita-se dos recursos computacionais, após gerar e delimitar os termos da área, para a elaboração de uma estrutura conceitual, ou árvore de domínio, ou ontologia, dependendo da vertente teórica.

Pierozzi Junior, Oliveira e Souza (2010, p.4) definem ontologia como:

um modelo comum ou estrutura conceitual sistematizada e de consenso que permite não só armazenar, mas também buscar e recuperar a informação sobre um determinado domínio do conhecimento. Mais especificamente, uma ontologia define os termos e as relações básicas para a compreensão de uma área do conhecimento.

Para elaborar uma estrutura conceitual, é preciso identificar os traços semânticos dos conceitos, cujos termos serão definidos. A partir desses traços, cria-se uma árvore com as unidades de um domínio, organizando-as a partir de relações hierárquicas e não hierárquicas entre elas.

De acordo com a Linguística de *Corpus*, tanto a seleção dos termos da área quanto a elaboração da estrutura conceitual seguem alguns critérios, como a frequência das unidades terminológicas. Entretanto, sabemos que, na Terminologia, nem sempre esse critério é suficiente, já que termos pouco frequentes no *corpus* também pode ser os mais usados naquele domínio.

Com base nisso, podem ser levantadas algumas questões: o *corpus* está pequeno; há problemas na seleção das fontes; os textos não são relevantes para a pesquisa; as ocorrências são baixas porque o domínio é novo; a pouca frequência se dá pela ausência de pesquisas que abordem tais problemáticas; há poucos trabalhos digitalizados ou publicados em meio digital; etc. Portanto, é bastante delicado tomar uma decisão terminológica baseando-se apenas na frequência das palavras, deixando de lado todos os fatores que podem influenciar na pesquisa.

Nesse sentido, o terminólogo, que goza de conhecimentos da língua, do domínio de especialidade em questão e dos princípios e métodos de variadas ciências – Linguística Computacional, Estudos do Léxico, Terminologia e Terminografia, Linguística de *Corpus* – é quem dita, a partir, é claro, dos dados obtidos, o percurso da pesquisa.

A ontologia é, então, a sistematização e a organização terminológica de um domínio, mais especificamente, das unidades terminológicas, expondo as relações semânticas entre cada uma delas. Em geral, parte-se de um nó (termo hiperônimo) para relacionar os outros termos (outros nós). Em prol de facilitar essa tarefa, pode-se usufruir de algumas ferramentas automáticas: e-Termos, Protégé, ONTOLP, Onto4All.

Posteriormente, é elaborada a ficha terminológica de cada termo, além de incrementar sua base definicional. Ambas as tarefas podem ser feitas simultaneamente, uma vez que o repositório de informações de uma unidade terminológica pode ser alimentado enquanto são criadas e preenchidas as fichas.

Sobre a base definicional importa dizer que se trata de um conjunto de dados, extraídos automaticamente pelo e-Termos, por exemplo, consistindo em trechos do próprio *corpus*, ou seja, é o contexto em que o termo ocorreu. Apesar de ser gerado de forma automática, o terminólogo precisa revisá-lo e preenchê-lo na medida em que se torna necessário buscar informações que nos façam entender o que aquela unidade significa.

É importante lembrar que esses excertos são o ponto de partida para as tarefas seguintes, em especial, as definições terminológicas, visto que analisamos os traços semânticos mais pertinentes de cada termo para nos auxiliar a redigir as DTs.

Automaticamente, também é elaborada, no ambiente e-Termos, por exemplo, a ficha terminológica. Nela, constam, segundo Oliveira (2009a, p. 80), “o registro completo e organizado de informações referentes a um dado termo, consistindo em um verdadeiro dossiê que contém todos os dados e elementos pertinentes ao objetivo da pesquisa.”

Para criá-las, são definidos alguns campos de acordo com as necessidades da pesquisa, exemplos: termo, classificação morfológica, definição, data de elaboração da definição, data da validação pelo especialista, termos sinônimos, termos antônimos e informação enciclopédica.

Por último, temos a redação das definições e, caso seja necessário, sua análise e difusão. Nesse momento, são observados os excertos definitórios contidos na base definicional e as relações semânticas da ontologia. A partir disso, o terminólogo irá elaborar as definições terminológicas, pautando-se também, na literatura da área.

Após apresentar detalhadamente todas as etapas de uma pesquisa em Terminologia, interessa-nos apontar algumas problemáticas que podem surgir com o uso de programas computacionais. Antes disso, vale salientar que toda ferramenta pode gerar erros, dado que se trata de uma máquina. Contudo, pesquisadores também podem cometer falhas (partindo-se da premissa de que todos erram).

À vista disso, o papel do terminólogo é minimizar os problemas, que não podem prejudicar o objetivo da pesquisa. Nesse sentido, vamos mencionar alguns contratemplos

reais, mas que podem ser facilmente solucionados e/ou refletidos para aprimorarmos nossas tarefas.

Primeiro, é interessante mencionar o tempo de aprendizado de uso de uma ferramenta computacional, pois isso interfere diretamente no prazo da pesquisa. Determinados programas são mais intuitivos que outros, porém, alguns podem ser de difícil manuseio e complicar ainda mais as tarefas. Posterior a isso, é salutar expor que os entraves da Terminótica podem ser verificados:

1) **na compilação do corpus:** caso os textos não forem convertidos adequadamente para o formato legível da ferramenta computacional, alguns dados podem não ser gerados; informações irrelevantes para a pesquisa, que não foram limpas na etapa de limpeza, aparecerão nos dados; alguns programas não convertem adequadamente os textos, retirando ou misturando informações essenciais para o trabalho.

2) **na extração dos termos:** as ferramentas compreendem o conceito ‘palavra’ de forma diferente. Geralmente, considera-se que a palavra é a unidade separada por espaços em branco. No entanto, cada programa pode gerar listas de palavras diferentes devido às regras nas quais foi construído (alguns termos podem não aparecer na lista⁶ ou apresentar uma porcentagem distinta). O importante é levar isso em consideração, deixando claro qual ferramenta está sendo usada, analisando, se preciso, os contextos de uso de cada unidade terminológica.

3) **na elaboração da estrutura conceitual:** muitas são as estruturas conceituais elaboradas antes de se chegar à estrutura final. Isso ocorre devido à sistematização do campo, que vai sendo construída na medida em que o terminólogo o descreve. Assim, o tempo e o objetivo da pesquisa são considerados para elaborar a quantidade de relações suficientes (nós). Esses nós são alterados enquanto o pesquisador vai entendendo o funcionamento do domínio, organizando os termos de acordo com o que eles significam e representam em comparação com os outros. Desse modo, criar e alterar nós implica dominar a ferramenta computacional na qual está sendo elaborada a ontologia. Porém, em algumas situações, o programa pode travar o trabalho, gerar falhas, e até mesmo não salvar as modificações feitas.

4) **no preenchimento das fichas terminológicas e no incremento da base definicional:** nem todas as ferramentas permitem alterar os campos criados da ficha terminológica, ou gerar, novamente, os excertos definitórios automáticos (do corpus). Além disso, inserir trechos definitórios na base definicional também pode ser um problema em algumas ferramentas.

Apesar das dificuldades para executar as tarefas em programa computacionais, estas são de grande valia para o trabalho terminológico em virtude de facilitar e otimizar as tarefas, bem como os resultados. Nessa perspectiva, a tecnologia se mostra, atualmente, importante e essencial em uma pesquisa na área de Linguística, sobretudo em estudos voltados para o léxico – geral ou especializado. Logo, é possível contornar as falhas que os

⁶ Sobre a extração de (candidatos a) termos, no e-Termos e no AntConc é possível dar o comando para que o programa gere listas com um ou mais elementos (n-gramas).

programas geram, já que os benefícios de seu uso na prática terminológica, principalmente, são mais significativos.

4 Considerações finais

Após as discussões feitas neste artigo, não nos restam dúvidas de que a internet possui um papel fundamental nos trabalhos científicos, sobretudo terminológicos. Isso pode ser comprovado pelas pesquisas realizadas ao longo dos anos, cujo aparato teórico e metodológico está, na maioria das vezes, respaldado na tecnologia.

Dessa forma, é relevante fundamentar-se nos princípios e métodos da Linguística de *Corpus* para a elaboração de *corpus* e *corpora* mais eficientes, bem como usufruir de ferramentas computacionais para tornar mais rápida a obtenção de dados e, conseqüentemente, mais precisa a análise dos resultados.

Nessa direção, baseado nas ideias de Sinclair (1994), Oliveira (2009b, p. 49) afirma que:

(...) análises feitas com auxílio de programas de computador podem também levar a novas descobertas sobre aspectos linguísticos até então não considerados como relevantes pelos pesquisadores, visto que evidências não esperadas podem emergir dos dados.

Assim, busca-se ferramentas computacionais nos trabalhos em Terminologia para que as tarefas sejam feitas em um prazo menor, a partir de interfaces computacionais amigáveis, capazes de gerar informações antes inimagináveis de forma manual. Com isso, pode-se manusear os dados obtidos com maior precisão, como gerar uma lista de frequência das palavras de um texto.

A estruturação e a sistematização de um domínio de especialidade, por exemplo, tornam-se mais eficazes na medida em que usamos programas que nos auxiliam desde a elaboração do *corpus* até a elaboração e difusão de verbetes. No entanto, tais programas apresentam dois lados: primeiro, eles permitem que as tarefas, neste caso, terminológicas, sejam otimizadas; segundo, é preciso levar em consideração o tempo de aprendizado de uma ferramenta, bem como os possíveis problemas que podem ocorrer, como falhas na compilação do *corpus*, erros na elaboração da ontologia, e outros.

Com relação, então, aos entraves dos programas computacional, faz-se importante mencionar que, em alguns casos, pode haver, por exemplo, “a falta de integração com outros sistemas e recursos, o baixo grau de automação recurso e habilidade humana e a falta de *corpus* automatizados disponíveis” (CABRÉ, 1993).

Apesar dos obstáculos encontrados em uma tarefa terminológica, a Linguística Computacional possibilita potencializarmos cada etapa. Além disso, ao partir dos fundamentos da Linguística de *Corpus*, a pesquisa pode ser mais bem elaborada desde o início, já que a compilação, o tratamento e a análise do *corpus* são (semi)automatizadas.

Nas palavras de Tagnin (2018, p. 14), a Linguística de *Corpus* vislumbra áreas linguísticas e não linguísticas, contribuindo significativamente para as tarefas, haja vista que permite o uso ou a criação de ferramentas que lidam com dados mais objetivos e confiáveis. Diante disso, a Linguística Computacional disponibiliza as ferramentas

necessárias para cada tipo de tarefa, como conversores de textos, compiladores, extratores de listas de palavras, ferramentas que elaboram (ou ajudam a elaborar) ontologias, entre outros exemplos.

Convém destacar, portanto, que a relação entre Terminologia – em especial à Teoria Comunicativa da Terminologia – e a Informática é relevante, e mais que isso, indispensável nos dias atuais. A partir dessa união, são aprimoradas a automatização de tarefas lexicais e terminológicas, mas não só: é possível elaborar tradutores automáticos ou semiautomáticos, desenvolver bancos de dados, sistematizar hierarquicamente domínios de especialidade, etc.

Em suma, o uso de ferramentas computacionais em trabalhos terminológicos pode (i) facilitar a elaboração dos produtos terminográficos; (ii) tornar mais rápida e prática a disseminação dos produtos; (iii) tornar mais eficaz a compilação e análise dos dados; (iv) gerar dados mais precisos e autênticos; (v) analisar dados maiores; e (vi) otimizar as etapas terminológicas. Sendo assim, este trabalho configura-se como uma obra basilar para demonstrar o interesse dos pesquisadores (linguistas) nas tecnologias e sua pertinência para as pesquisas em Terminologia.

Referências

ALMEIDA, G. M. de B. A Teoria Comunicativa da Terminologia e a sua prática. **Alfa**, São Paulo, v. 50, n. 2, p. 85-101, 2006. Disponível em: <https://periodicos.fclar.unesp.br/index.php/alfa/article/viewFile/1413/1114>. Acesso em: 12 abr. 2019.

_____. O percurso da Terminologia: de atividade prática à Consolidação de uma disciplina autônoma. **Tradterm**, 9, 211-222, 2003. Disponível em: <https://doi.org/10.11606/issn.2317-9511.tradterm.2003.49087>. Acesso em: 27 ago. 2021.

ALMEIDA, G. M. B.; OLIVEIRA, L. H. M. de; ALUÍSIO, S. M. A Terminologia na era da Informática. **Ciência e Cultura**, Campinas, v. 58, n. 2, p. 42-45, 2006. Disponível em: http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200016. Acesso em: 12 abr. 2019.

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Calidoscópico**, São Leopoldo, v. 4, n. 3, p. 156-178, set./dez. 2006. Disponível em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002>. Acesso em: 12 abr. 2019.

BALESTERO, M. de S. **Definições terminológicas da Revisão de Textos**: estudos iniciais para a elaboração de um glossário. Orientadora: Profa. Dra. Clotilde de Almeida Azevedo Murakawa. 2019. Dissertação (Mestrado em Linguística e Língua Portuguesa) – Faculdade de Ciências e Letras, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Araraquara, 2019.

BARROS, L. A. Aspectos epistemológicos e perspectivas científicas da terminologia. **Cienc. Cult.**, São Paulo, v. 58, n. 2, p. 22-26, June, 2006. Disponível em: http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200011&lng=en&nrm=iso. Acesso em: 27 ago. 2021.

- _____. **Curso Básico de Terminologia**. 1. ed. São Paulo: Edusp, 2004.
- BERBER SARDINHA, T. **Linguística de Corpus**. São Paulo: Manole, 2004.
- _____. Que tipo de corpus é a Web? **Revista da ANPOLL**, Belo Horizonte, v. 1, n. 15, p. 191-220, jul./dez. 2003. Disponível em: <https://revistadaanpoll.emnuvens.com.br/revista/article/view/430>. Acesso em: 12 abr. 2019.
- _____. Linguística de Corpus: histórico e problemática. **DELTA**, São Paulo, v. 16, n. 2, p. 323-367, 2000. Disponível em: <http://www.scielo.br/pdf/delta/v16n2/a05v16n2.pdf>. Acesso em: 12 abr. 2019.
- BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics: Investigating language structure and use**. Cambridge: Cambridge University Press, 1998.
- CABRÉ, M.T. Prefácio. In: BARROS, L. A. **Curso Básico de Terminologia**. 1. ed. São Paulo: Edusp, 2004.
- _____. Terminología y Lingüística: la Teoría de las Puertas. **Estudios de Lingüística del Español (Elies)**. España, v. 16, 2002. Disponível em: <http://elies.rediris.es/elies16/Cabre.html>. Acesso em 22 fev. 2019.
- _____. **La terminología: representación y comunicación – elementos para una teoría de base comunicativa y outros artículos**. Barcelona: Institut Universitari de Lingüística Aplicada, 1999.
- DOMÉNECH, E. V. **Patrones de la definición terminológica en el ámbito de la cerámica: una contribución a la automatización de las definiciones**. 2014. Tesis (Doctorat) – Universitat Jaume I, Castellón, 2014.
- FINATTO, M. J. B.; LOPES, L.; CIULLA, A. Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: uma parceria bem-sucedida. **Domínios de Lingu@gem**, v. 9, n. 5, 2015. Disponível em: <https://core.ac.uk/download/pdf/293601526.pdf>. Acesso em: 27 ago. 2021.
- KENNEDY, G. **An Introduction to Corpus Linguistics**. London; New York: Longman, 1998.
- LORENTE, M. A lexicologia como ponto de encontro entre a gramática e a semântica. In: ISQUERDO, A. N. e KRIEGER, M.G. **As ciências do léxico**, vol. II. Campo Grande: Editora UFMS, 2004.
- OLIVEIRA, L. H. M. **e-Termos: um ambiente colaborativo web de gestão terminológica**. 2009. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2009a. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-19012010-150638/pt-br.php>. Acesso em: 12 abr. 2019.

OLIVEIRA, Lúcia Pacheco de. *Linguística de Corpus: teoria, interfaces e aplicações. Matraga*, 2009b, v. 16, n. 24. Disponível em: <https://www.e-publicacoes.uerj.br/index.php/matraga/article/view/27796>. Acesso em: 27 ago. 2021.

Othero, G. de Ávila. (2006). *Linguística Computacional: uma breve introdução. Letras De Hoje*. Porto Alegre: PUCRS, v. 41, nº 2, p. 341-351. Recuperado de <https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/605>.

PIEROZZI JUNIOR, I.; OLIVEIRA, L. H. M. de; SOUZA, K. X. S. de. Construindo ontologias de domínio: o (re)conhecimento da intensificação agropecuária no Brasil. In: SEMINÁRIO DE PESQUISA EM ONTOLOGIAS NO BRASIL, 3., 2010, Florianópolis. *Anais [...]* Florianópolis: UFSC, 2010. p. 100-108.

RENOUF, A. (ed.). *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 1998.

SINCLAIR, J. 2005. "Corpus and Text - Basic Principles" in *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Disponível: http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf. Acesso em: 27 ago. 2021.

SINCLAIR, J. Trust the text. In: COULTHARD, M. (ed.), *Advances in written text analysis*. London: Routledge, 1994. p. 12-25.

SOUZA, J. W. C. ; FILIPPO, A. Um exercício em Linguística de Corpus no âmbito do projeto TerminoNet. *NILC (ICMC-USP) - Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional*, 2010.

TAGNIN, S. Prefácio. In: *Linguística de corpus: perspectivas*. Porto Alegre: Instituto de Letras - UFRGS, 2018.

TERMISUL. Disponível em: <http://www.ufrgs.br/termisul/>. Acesso em: 27 ago. 2021.

VIEIRA, R.; LIMA, V. L. S. *Linguística computacional: princípios e aplicações*. In: *IX Escola de Informática da SBC-Sul*. Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul, 2001.



The use of computational tools in terminology research

ABSTRACT:

It is known that, since the advance of Informatics, the research in Terminology began to have the contribution of technology for the accomplishment of its tasks. In this sense, this article aims to demonstrate the use of computational tools in terminological activities, since each stage is executed in an (semi)automatic way. In addition, it seeks to reflect on the contributions and obstacles of these tools, based on three theoretical-methodological perspectives: the Communicative Theory of Terminology, Computational Linguistics and Corpus Linguistics. It presents, then, the stages of a work in Terminology based on descriptive and linguistic terminology, besides some of the programs that make the tasks more authentic, accurate and effective. Thereby, this work corroborates the terminological studies and the Natural Language Processing, which aims to treat natural languages by using computer programs.

KEYWORDS:

Terminology. Computational Tools. Natural Language Processing. Corpus Linguistics. (Semi)Automated tasks.