Calidoscópio
Vol. 14, n. 3, p. 480-488, set/dez 2016
Unisinos - doi: 10.4013/cld.2016.143.11

O uso da Web como corpus em pesquisas fraseológicas: uma prática prejudicial ou um recurso valioso?

Using web as corpus in phraseological researches: A damaging practice or a valuable resource?

Eloísa Moriel Valença[1]
elomoriel@hotmail.com
Universidade Estadual Paulista Júlio de Mesquita Filho

Marilei Amadeu Sabino[1]
amadeusm@ibilce.unesp.br
Universidade Estadual Paulista Júlio de Mesquita Filho

RESUMO – Este artigo pretende apresentar o percurso metodológico seguido na busca de Expressões Idiomáticas de Baixa Dedutibilidade Metafórica (EIBDM) presentes em dicionários de fraseologismos italianos. Por Expressões Idiomáticas de Baixa Dedutibilidade Metafórica (EIBDM) entendem-se aquelas combinatórias que apresentam, em sua constituição, elementos históricos, geográficos, religiosos, mitológicos, meteorológicos (temporais) ou mesmo simplesmente linguísticos, explícitos ou implícitos, que são típicos ou significativos apenas na língua fonte ou língua de origem (no nosso caso, o italiano). Isso quer dizer que seu material semântico não auxilia nem fornece pistas para a tradução na língua alvo ou língua de chegada (no caso desta investigação, o português, na variante brasileira da língua). Discutir-se-á a importância da Linguística de Corpus para os estudos fraseológicos e a relevância da utilização da Web como corpus (WaC) para pesquisas fraseológicas e lexicográficas. Ressalvadas suas limitações, o uso da Web como corpus demonstra-se valioso para atestar a frequência, os contextos de uso, o comportamento sintático e os sentidos que as expressões idiomáticas assumem nos diferentes contextos em que figuram.

Palavras-chave: metáfora, Expressões Idiomáticas, Web como corpus (WaC), Baixa Dedutibilidade Metafórica.

ABSTRACT – This paper intends to present the methodological approach used in the search for Idiomatic Expressions of Metaphorical Low-Deductible (EIBDM) found in Italian dictionaries of phraseologisms. By Idiomatic Expressions of Metaphorical Low-Deductible, we mean those combinatorial expressions in foreign language which presents, in its constitution, historical, geographical, religious, mythological, weather (temporal) or simply linguistic elements, explicit or implicit, that are typical or significant only in the source language or in the native language (in our case, the Italian).This means that its semantic material does not help or provide clues to the translation process in the target language (in the case of this research, the Brazilian variant of the Portuguese). We also discuss the importance of Corpus Linguistics for phraseological studies and the relevance of using the Web as Corpus (WAC) in phraseological and lexicographical research. Despite its limitations, the Web as Corpus is a valuable tool to attest frequency, contexts of use, syntactic and semantic behavior of idiomatic expressions in different contexts.

Keywords: metaphor, idioms, Web as corpus (WaC), Low Metaphorical Deductibility.

Introdução

Este trabalho visa a apresentar o recorte de uma pesquisa desenvolvida em nível de mestrado, cujo objetivo era analisar as metáforas subjacentes a alguns tipos de Expressões Idiomáticas de Baixa Dedutibilidade Metafórica (EIBDM), valendo-se da Web como corpus (*Web as corpus*, doravante WaC) para a verificação da frequência de uso e a extração de exemplos. Esta pesquisa deu origem a um glossário bilíngue (italiano-português) com 72 EIBDM.

Pode-se dizer que o uso da web como corpus é um recurso que ainda gera bastante polêmica em determinados tipos de pesquisas linguísticas, uma vez que alguns pes-

[1] Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Instituto de Biociências, Letras e Ciências Exatas. Rua Cristóvão Colombo, 2265, Jardim Nazareth, São José do Rio Preto - SP, 15054-000.

quisadores consideram a web um corpus "sujo", que não se adequa à visão tradicional de corpus como apresentada pela Linguística de Corpus (doravante LC). Desse modo, procurar-se-á, neste artigo, apresentar duas diferentes visões sobre a WaC, seja aquela que a vê como uma prática prejudicial às investigações linguísticas, seja aquela que a considera um recurso valioso, principalmente no que se refere aos estudos fraseológicos e lexicográficos.

O estudo dos fraseologismos de uma língua é um assunto muito vasto que se expande consideravelmente, quando se torna comparativo entre dois ou mais idiomas (italiano e português brasileiro, neste caso). O objeto de investigação das pesquisas dessa natureza é muito amplo, visto que as línguas naturais possuem milhares de fraseologismos, grande parcela dos quais é diariamente empregada pelos falantes, em seus discursos de rotina.

Para a pesquisa em questão, coletamos algumas expressões idiomáticas (doravante EIs) em dicionários italianos (Lapucci, 1990; Radicchi, 1985; Termignoni, 2010; Quartu e Rossi, 2012) e verificamos a sua frequência seguindo o método, que será explicitado a seguir. Essas combinatórias fraseológicas foram selecionadas com base na presença de elementos culturais em sua constituição, do tipo histórico, geográfico, mitológico, meteorológico, ou outros, que poderiam ser específicos da cultura italiana e que, por essa razão, não seriam encontrados na cultura brasileira, tal como figuram no fraseologismo italiano.

Para verificar as suas ocorrências e frequência, decidiu-se utilizar a web como corpus, uma vez que, nos corpora tradicionais, geralmente não são encontradas evidências suficientes para atestar o uso corrente de fraseologismos. Primeiramente apresenta-se o conceito de Expressões Idiomáticas de Baixa Dedutibilidade Metafórica e as dificuldades de tradução em relação aos elementos culturais idiossincráticos da cultura italiana, depois discutem-se os prós e contras do uso da WaC em pesquisas sobre EIs. O termo "Baixa Dedutibilidade Metafórica" foi cunhado pelas autoras Tonfoni e Turbinati (1995) e, nesta pesquisa, foi aplicado especificamente às Expressões Idiomáticas. Em seguida, estabelece-se o limiar de frequência para as EIs da língua italiana, o qual regulará a inclusão ou não das EIs selecionadas no corpus em tela. Xatara (2008) e Rios (2010) estabeleceram o limiar de frequência de EIs para as línguas francesa e portuguesa, respectivamente, mas para a língua italiana essa informação ainda é praticamente desconhecida. Neste artigo, faz-se, ainda, um levantamento dos corpora italianos existentes e que podem ser relevantes para diferentes pesquisas.

As Expressões Idiomáticas de Baixa Dedutibilidade Metafórica (EIBDM) e as dificuldades de tradução

De acordo com Tonfoni e Turbinati (1995, p. 240), a complexidade do processo de tradução de EIs vem de alguns aspectos típicos dessas expressões: a metaforicidade e a idiomaticidade. Os problemas referentes ao processo de tradução são: os diversos níveis de correspondência entre as expressões pertencentes a uma língua de origem e as expressões existentes na língua de chegada (L2); a falta de expressões equivalentes; e o problema de se manter os eventuais traços contextuais, no caso de referências literais.

Para as autoras, existem três níveis de dedutibilidade do nível pragmático:

(i) Alta dedutibilidade: expressão metafórica com nível pragmático imediatamente dedutível;
(ii) Média dedutibilidade: expressão metafórica com nível pragmático dependente do nível semântico;
(iii) Baixa ou nenhuma dedutibilidade: expressão metafórica sem ligação aparente entre o nível pragmático e o nível semântico.

O que se está denominando dedutibilidade metafórica, aqui, para autores como Zuluaga (1997), Ortíz-Alvarez (2000) e outros, tem a ver com a idiomaticidade do fraseologismo. Segundo Ortíz-Alvarez (2000, p. 153), a idiomaticidade apresenta uma escala, podendo existir em maior ou menor grau em uma expressão. As menos idiomáticas são aquelas em que apenas um elemento é idiomático, ou seja, as expressões metafóricas cuja imagem seja de fácil codificação (alta dedutibilidade metafórica, alta transparência, baixa opacidade). As totalmente idiomáticas (baixa dedutibilidade metafórica, baixa transparência, alta opacidade) são aquelas em que nenhum dos significados

Tabela 1. Exemplos de dedutibilidade, idiomaticidade, transparência e opacidade.
Table 1. Examples of deductibility, idiomaticity, opacity and transparency.

Português	Espanhol	Italiano	Grau
Deitar lenha na fogueira	Poner leña en el fuego	Mettere legna al fuoco	Alta dedutibilidade / Baixa opacidade Baixa idiomaticidade / Alta transparência
Declarar forfait	Declararse forfait / Ganar por forfait / Hacer forfait	Dichiarare forfait	Baixa dedutibilidade / Alta opacidade Alta idiomaticidade / Baixa transparência

de seus constituintes contribui para o significado total da expressão (ver Tabela 1).

Como é possível analisar na tabela acima, a expressão *deitar lenha na fogueira* (retirada de Ortíz-Alvarez, 2000, p. 153) oferece, com o significado individual de cada palavra, pistas para se descobrir o sentido global da expressão; portanto, ela possui um alto grau de dedutibilidade. Já a expressão *declarar forfait* apresenta um baixo grau de dedutibilidade e, consequentemente, alta opacidade semântica, pois seu material linguístico não auxilia na compreensão do significado global da expressão.

As expressões metafóricas de baixa ou nenhuma dedutibilidade fazem referência a elementos ligados à cultura da língua de origem, por isso nem sempre é possível traduzir esses valores para a língua de chegada por meio da tradução literal.

A primeira dificuldade encontrada no processo de tradução é a metaforicidade das expressões, pois, muitas vezes, não existe correspondência entre o nível semântico da expressão e o seu nível pragmático. De acordo com Tonfoni e Turbinati (1995), para se traduzir uma metáfora, não basta traduzi-la sintático-semanticamente; é preciso analisar o nível pragmático em questão. O primeiro passo é identificar a mensagem presente na expressão metafórica, para poder transpô-la para a língua de chegada. Na medida do possível, é melhor utilizar a mesma metáfora da língua de partida (L1), mas quando não há essa possibilidade, é necessário transformar a "imagem metafórica" de maneira que o conteúdo comunicado seja igual tanto na língua de partida quanto na língua de chegada (Tonfoni e Turbinati, 1991, p. 240).

A segunda dificuldade encontrada pelas autoras é a idiomaticidade. O caráter idiomático indica uma particularidade. As expressões em questão podem pertencer a certo sistema linguístico, assim como podem ser compartilhadas por sistemas linguísticos diferentes. Ao contrastarmos o sistema linguístico do italiano com o do português observamos que muitos fraseologismos são comuns às duas línguas, mas uma parcela é específica de um ou de outro sistema. Tanto a cultura italiana como a brasileira fazem parte da cultura ocidental, além de serem línguas latinas, de existir uma forte ligação entre Itália e Brasil devido à migração, às artes, o que justifica, pelo menos em parte, os fraseologismos comuns às duas culturas.

A idiomaticidade das expressões limita as opções de tradução para o idioma de chegada. Quando se traduz uma EI que apresenta elementos culturais em sua constituição, o grupo de correspondentes na língua de chegada torna-se restrito, pois muitas vezes essas expressões fazem referência a elementos intrinsecamente ligados à cultura de origem que não podem ser retomados pela tradução literal na língua de chegada.

Como observado pelas autoras (Tonfoni e Turbinati, 1995), outra dificuldade na tradução são os diferentes graus de correspondência entre L1 e L2. O caso mais simples de correspondência é aquele no qual podemos encontrar diretamente uma expressão na língua de chegada que possua a mesma estrutura sintática da expressão de L1, podendo reproduzir essa estrutura exatamente igual em L2. Ou seja, ocorre o que muitos autores nomeiam *equivalência total*. É o caso da expressão *calcanhar de Aquiles/tallone d'Achille*, em que o grau de "tradutibilidade" é alto, uma vez que na língua de chegada pode-se encontrar uma expressão equivalente em todos os níveis textuais (sintático/semântico/pragmático).

Contudo, em muitos casos, não é possível encontrar uma correspondência exata de L1 em outra língua, já que, muitas vezes, não existe uma expressão capaz de transmitir a mesma mensagem. Em outros casos, tem-se que o fraseologismo de L1 possui mais de um correspondente em L2, ou então, que a EI em L1 tem mais de um significado, sendo necessário encontrar um correspondente em L2 para cada um dos seus sentidos. Essa diferença quanto à extensão de sentidos das EIs torna-se um dos grandes desafios de tradução.

A título de exemplo, cita-se a expressão idiomática *fare (essere) il (un) cireneo*, para a qual Quartu e Rossi (2012, p. 96) atribui os seguintes sentidos:

> 1. Ajudar alguém em um trabalho cansativo; 2. Assumir as penas de outro sem tê-las merecido; 3. Dividir penas e sofrimentos de alguém na tentativa de oferecer alívio. Com esse nome ficou conhecido, na história, Simão o cireneu, que segundo o Evangelho (Lucas, 26; Mateus 32; Marcos, 21) ajudou Cristo a carregar a cruz na subida ao Gólgota. Não fica claro se essa ajuda foi dada espontaneamente ou se ele foi obrigado a ajudar Cristo (tradução nossa).

A passagem do cireneu aparece na bíblia, nos evangelhos de Lucas, Mateus e Marcos. Enquanto Jesus carregava a cruz em direção ao local de crucificação, Simão, o cireneu, foi "escolhido ao acaso" pelos soldados para levar a cruz de Jesus até o Gólgota. Sem escolha, ele a carregou. Embora o cristianismo, e, consequentemente, a bíblia sagrada, tenham tido ampla difusão na cultura ocidental (Itália e Portugal/Brasil), a expressão "fazer como o cireneu" ou "ser um cireneu" não foi convencionalizada na cultura brasileira. Itália e Brasil possuem uma forte tradição católica/cristã, mas apesar dessa tradição religiosa em comum, a EI *fare il cireneo* surgiu e se cristalizou apenas no italiano. Contudo, embora essa expressão não tenha sido institucionalizada na cultura brasileira, devido à capacidade que todo ser humano possui de compreender as metáforas (em seu sentido mais amplo) que permeiam a linguagem, torna-se possível (embora não livre de esforços) a compreensão da referida expressão italiana em português.

Percebe-se, contudo, que, para o segundo sentido da expressão, seria possível propor como correspondente idiomático a expressão *pagar o pato*, que carrega o sentido de "levar a culpa", "assumir as penas no lugar de

outra pessoa". Entretanto, a expressão *pagar o pato* não apresenta uma extensão de sentido de modo a abranger todos os significados presentes também nas duas outras acepções da EI italiana. Assim, pode-se dizer que se trata de um caso de equivalência parcial, de acordo com Felber (1984). Ao procurar por um correspondente idiomático para o primeiro e terceiro significados da expressão, seria possível oferecer como possibilidade a expressão *ajudar a carregar a cruz* – correspondente este que possui tanto sentido literal quanto figurado, em ambos os idiomas. Todavia, haveria a perda do culturema *cireneo*, embora a metáfora de *ajudar a carregar a cruz* continue remetendo à história bíblica.

Assim, o tradutor consegue depreender o sentido dessa expressão após percorrer um longo caminho, o qual não é simples, nem tampouco imediato. Primeiro, seria necessário recuperar a informação de quem foi este cireneu, e, ao descobrir que a metáfora subjacente a essa expressão está em uma figura bíblica, saber de quem se trata, qual papel desempenhou no evangelho, dentre outros. Somente após percorrer esse caminho, ele será capaz de compreender que *fare il cireneo* significa ajudar alguém a "carregar uma cruz". Porém, o tradutor deve saber que, pelo fato de este não ser o único correspondente possível para a expressão, ele não será adequado a todos os contextos de uso, exatamente por apresentar extensões de sentido diferentes.

A importância da Linguística de Corpus para os estudos fraseológicos

Desde o lançamento do primeiro corpus linguístico eletrônico em 1964, o *Brown University Standard Corpus of Present-day American English*, com um milhão de palavras, percebe-se que a Linguística de Corpus avançou muito, especialmente nos últimos tempos. Segundo Sardinha (2000), o corpus Brown impulsionou o desenvolvimento da Linguística de Corpus. Isso não quer dizer que não existiam corpora anteriormente, ou que não existiam estudos baseados em corpora até então, mas a tecnologia utilizada no *Brown* foi uma inovação para os estudos da área, possibilitando que a Linguística de Corpus se estabelecesse como uma nova área de estudos.

A LC encara a linguagem como um sistema probabilístico e a língua em uma abordagem empirista, ou seja, acredita que o conhecimento se origina da experiência, nesse caso, da experiência com a linguagem. A invenção do computador, bem como a ampla aquisição dos computadores pessoais, de fato contribuíram para o rápido desenvolvimento da Linguística de Corpus enquanto disciplina.

Nos estudos lexicológicos e lexicográficos atuais, é tendência o lexicógrafo trabalhar a partir de corpora, preferencialmente os mais extensos e variados possíveis. As informações semânticas e pragmáticas de qualquer Unidade Lexical (UL) encontrarão fundamentação nas informações extraídas de seu uso real, o qual poderá ser encontrado nos corpora de textos autênticos e originais. De acordo com Sanchez, corpus é

> um conjunto de dados linguísticos (presentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (Sanchez, 1995, p. 8-9, *in* Sardinha, 2004, p. 338).

Os quatro pré-requisitos apresentados por Sardinha (2004) para a formação de um corpus computadorizado são: 1. O corpus deve ser formado de textos autênticos, em linguagem natural; 2. Quando se fala em autenticidade, subentendem-se textos escritos [e de discurso falado] por falantes nativos; 3. O conteúdo do corpus deve ser escolhido criteriosamente; e 4. O corpus deve ser representativo, ou seja, conter um conjunto que represente uma variedade linguística. A representatividade de um corpus é um critério bastante problemático, pois é particular a cada tipo de pesquisa desenvolvida. Foi esse quarto critério apresentado por Sardinha o que mais pesou e nos levou a optar pelo uso da *Web* como corpus.

Os corpora tradicionais vs. o corpus Web

Em uma busca na rede por corpora italianos, encontramos no *site* da *Enciclopedia Treccani* (s.d.) uma lista dos principais corpora de língua italiana. O primeiro corpus de referência da língua italiana foi o trabalho pioneiro do padre Roberto Busa, que nos anos 1950 usou o computador para indexar a obra de Tomás de Aquino, para a igreja católica, São Tomás. Anos mais tarde, em 1971, esse corpus levou à publicação do *Lessico di frequenza della lingua italiana contemporanea* (LIF, de Bortoloni, Tragliavini e Zampolli).

O *Corpus e lessico di frequenza dell'italiano scritto* (CoLFIS, s.d.), é um corpus de referência, com cerca de 3.800.000 palavras. O CoLFIS foi elaborado respeitando o equilíbrio da quantidade de textos de diferentes gêneros textuais (jornais, revistas, livros), baseado nas leituras dos italianos segundo o ISTAT (*Italian National Institute of Statistics*), uma organização pública de pesquisa presente na Itália desde 1926.

Nos anos 2000 surge o CORIS (2017), *Corpus di italiano scritto*, um corpus geral de referência do italiano escrito. O CORIS contém 130 milhões de palavras e é constituído por textos autênticos e recorrentes, em formato eletrônico. O corpus está disponível online, com algumas limitações. Segundo a Enciclopedia Treccani (s.d.), o corpus de maior dimensão atualmente disponível é o *itWaC* (Baroni *et al*., 2009, disponível em http://wacky.sslmit.unibo.it).

Outros corpora importantes são os de língua italiana falada, dentre eles, o LIP (*Lessico di frequenza dell'italiano parlato* – http://languageserver.uni-graz.at/badip/), o CLIPS (*Corpora e lessici di italiano parlato e scritto* – http://www.clips.unina.it/), o C-ORAL-ROM (http://lablita.dit.unifi.it/coralrom/), o CiT (*Corpus di italiano televisivo* – http://www.sspina.it/cit), dentre outros.

Como se pode observar, já existe uma grande variedade de corpora italianos, atualmente. O grande problema encontrado, contudo, é que para muitos deles o acesso à sua totalidade é restrito. Além disso, como será possível observar, a seguir, a frequência dos fraseologismos presentes nesses corpora é relativamente baixa, o que torna esses ambientes de pesquisa menos relevantes que a web, para pesquisas fraseológicas.

Observe-se um exemplo do número de ocorrências em um corpus tradicional, comparado com as ocorrências encontradas em uma pesquisa realizada na web. Pesquisando a expressão "fare le cose alla carlona" no corpus italiano CORIS/CORDIS, percebe-se que a frequência desse fraseologismo é muito baixa. O número total de ocorrências foi 11. É importante ressaltar que esse tipo de corpus é do italiano escrito e sabe-se que as EIs ocorrem mais frequentemente na linguagem oral, principalmente em registro coloquial.

Poder-se-ia pensar que, já que não ocorrem muitas vezes nos corpora tradicionais, os fraseologismos são irrelevantes para o uso cotidiano da língua. Como afirma Rios, "o fato de os idiomatismos terem baixa frequência relativa nos *corpora*, ao invés de indicar que eles são pouco empregados na língua corrente, pode indicar que eles ainda não estão suficientemente presentes nesses bancos de dados textuais" (2010, p. 70). O fraseólogo encontra um problema particular quando trabalha com corpora tradicionais gigantescos: de um lado, ele poderá ter nos corpora a confirmação de que os fraseologismos buscados existem, mas por outro, se ele for fazer um estudo da sua frequência, pode achar que, por ser de baixa recorrência, esse fraseologismo não seja relevante para a pesquisa na língua em questão.

Assim, como foi apontado anteriormente, ao se fazer a busca no corpus do italiano escrito, a frequência de alguns fraseologismos foi relativamente baixa; contudo, consultando a web, obteve-se uma frequência muito maior, como se verá adiante. Do mesmo modo, nos outros corpora citados, abertos para consulta, também não foi verificada frequência significativa dos fraseologismos pesquisados. Assim sendo, diante de todos os desafios encontrados tanto para se ter acesso a essas bases de dados de pesquisa, quanto no que diz respeito à frequência dessas combinatórias nos corpora, acredita-se que a utilização da Web como corpus, apesar das limitações e obstáculos, ainda representa uma escolha acertada para pesquisas fraseológicas. Essa escolha se sustenta principalmente pelas vantagens de ser de acesso livre, apresentar uma imensidão de dados que não é contemplada em nenhum outro corpus já criado, além de representar tanto a linguagem oral como a escrita, nos registros culto e coloquial, uma vez que possui uma gama variada de gêneros textuais que apresentam diversas características da linguagem oral, como os blogs, twitters e etc.

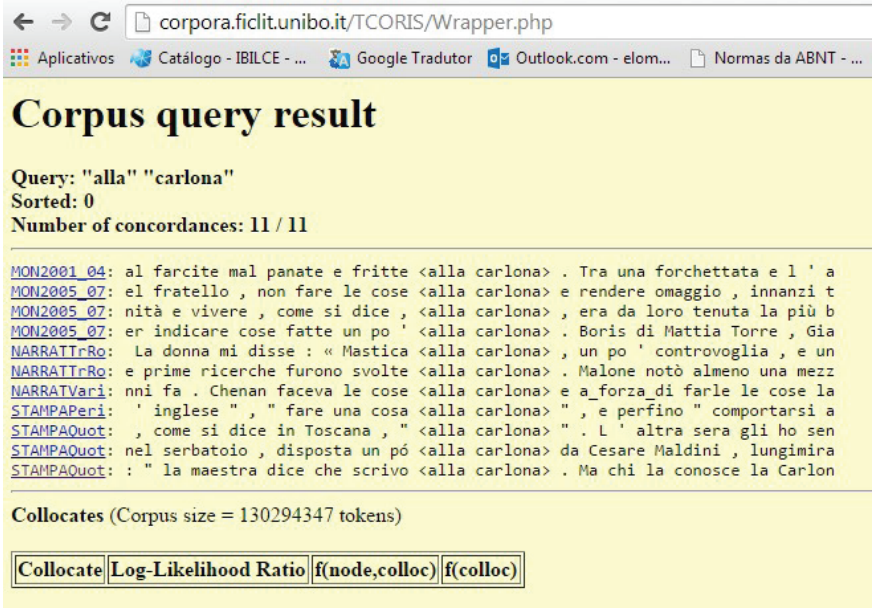

Figura 1. Número de ocorrências em um corpus tradicional CORIS/CORDIS.
Figure 1. Number of occurences in a tradicional corpus CORIS/CORDIS.

A web como corpus (WaC)

Colson (2007), ao discutir o uso da web como corpus, afirma que a web não é um corpus no sentido estrito de "conjunto de dados linguísticos coletados criteriosamente para serem objeto de pesquisa linguística", já que os textos ali presentes não foram compilados por linguistas cautelosos que levam em consideração as variantes regionais, o estilo, a linguagem falada e a linguagem escrita, a diversidade de fontes, etc. Na definição do autor (Colson, 2007), um linguista de corpus deveria ter total controle sobre o seu corpus, o que não acontece utilizando a World Wide Web (WWW). Outra objeção trazida pelo autor quanto ao uso da web como corpus é a de ter que lidar com o caráter artificial da linguagem da internet. Essa linguagem é de um tipo particular e representa um gênero que fica entre o registro oral e o registro escrito. Segundo ele, essas duas maiores objeções devem ser consideradas e a precaução e atenção do pesquisador devem ser redobradas ao escolher trabalhar com a WaC. Dentre as desvantagens de se utilizar a Web como *corpus* também estão: a presença de "erros" que podem ser ortográficos, gramaticais, de vocabulário, entre outros, além da falta de instrumentos eficazes de busca linguística.

Contudo, é importante ressaltar a diferença de se utilizar a web para a formação de um corpus (Web for Corpus – WfC) e a web como um corpus, de acordo com Schryver (2002 *in* Rios, 2010, p. 68). Quando se utiliza a web para a constituição de um corpus, se usa um corpus compilado a partir de páginas retiradas da web; ao contrário, quando se utiliza a web como corpus, analisa-se o corpo de documentos disponível *on-line*, acessado de maneira direta e gratuita como um corpus.

Fletcher (2005, p. 4), em seu artigo *Concordancing the Web: promise and problems, tools and techniques*, apresenta as vantagens da utilização da web como corpus: atualidade e espontaneidade; completude e escopo; diversidade linguística; custo e conveniência; e representatividade.

De acordo com Kilgarriff e Grefenstette (2003), os cientistas da linguagem e tecnólogos estão se voltando cada vez mais ao uso da web como fonte de dados, seja por causa do seu tamanho, seja porque é a única fonte disponível para o tipo de linguagem em que eles estão interessados, ou simplesmente porque é de graça e instantaneamente acessível. Esse aumento na busca pela Web como fonte de dados suscita um questionamento: a Web é um corpus? Segundo Kilgarriff e Grefenstette (2003), se corpus for considerado como uma coleção de dados, a resposta é afirmativa.

Encontram-se algumas dificuldades ao buscar informações acerca do tamanho da web, como a falta de informações completas e alguns dados desencontrados. Uma possível explicação para esse fato é o constante aumento de informações nela realizado diariamente, o que faz com que o seu tamanho chegue a ser imensurável. Por essa razão, os dados sobre a sua dimensão são apenas aproximativos e não precisos. Em levantamento feito em janeiro de 2003, os autores Kilgarriff e Grefenstette calculavam existir 172 milhões de endereços registrados na internet.

Para justificar o uso da Web como corpus em pesquisas, Sardinha (2000) aponta três critérios, no que diz respeito a sua extensão. O primeiro critério para justificar o uso da Web nessas pesquisas refere-se ao número de palavras que este tipo de corpus possui, uma vez que "quanto maior o número de palavras, maior será a chance de o corpus conter palavras de baixa frequência" (2000, p. 344). Segundo o site *Statistic Brain*, (http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/), em 2014, o Google possuía cerca de 67 bilhões de páginas indexadas. Se a Web possui mais de 60 bilhões de páginas de internet, o número de palavras torna-se quase imensurável. O segundo critério é o número de textos, ou seja, "um número de textos maior garante que este tipo textual, gênero ou registro [ao qual se refere o autor] esteja mais adequadamente representado" (Sardinha, 2000, p. 344). E a terceira dimensão seria o número de gêneros, registros ou tipos textuais. Na web há uma quantidade variada de gêneros: textos jornalísticos, blogs, textos de divulgação científica, contos, narrativas e tantos outros. O "número maior de textos de vários tipos permite uma maior abrangência do espectro genérico da língua" (Sardinha, 2000, p. 344). Portanto, a web é um corpus representativo da heterogeneidade da língua.

Acreditamos que, para o estudo dos fraseologismos, sejam eles expressões idiomáticas, provérbios, colocações, frases feitas, entre outros, o uso da WaC é bastante relevante, já que em corpora tradicionais a ocorrência de fraseologismos é muito baixa, menos de 1 a cada milhão de palavras, de acordo com Colson (2007). O autor afirma que os fraseologismos estão presentes tanto na língua falada como na língua escrita, mas ao procurar, por exemplo, a expressão de língua inglesa, *to spill the beans* no corpus *Bank of English* (com 211 milhões de palavras), a frequência foi de 0.56 por milhão de palavras (PMW). Isso indica que seria necessário um corpus gigantesco para se conseguir estudar os fraseologismos, e a World Wide Web é a alternativa encontrada para solucionar essa questão.

Contudo, apesar de todos os inconvenientes e limitações do uso da web em pesquisas, Colson (2007) chama a atenção para aspectos que são positivos:

> O tamanho do corpus é tão grande (de 1 até 50 bilhões de palavras para as línguas europeias) que a probabilidade de se tirar conclusões erradas é muito limitada, embora não possa ser totalmente excluída. Além do mais, no caso da fraseologia, nenhum corpus existente, seja lá em qual for o idioma, consegue dar conta de abarcar tantas combinatórias como a World Wide Web (Colson, 2007, p. 1072).[2]

[2] "the size of the *corpus* is so big (from one to fifty billion words for European Languages) that the probability of drawing wrong conclusions is very limited, although it cannot be totally excluded. In the case of phraseology, besides, no existing *corpus* in any language can claim to include as many set phrases as the World Wide Web. (Colson, 2003, p. 1072).

Nesses termos, apesar das desvantagens apresentadas acima, muitos autores, como Kilgarriff e Grefenstette (2003), Colson (2003), Fletcher (2005), Xatara (2008) e Rios (2010), defendem o uso da web como base textual de grande utilidade para a realização de pesquisas, principalmente as fraseológicas, pois, apesar de não ser exatamente uma base de dados linguísticos, no sentido restrito do termo, é uma fonte válida para a obtenção de dados que podem atestar o uso real da língua, proporciona ao pesquisador acesso a um corpus gigantesco, formado por mais de 3 bilhões de páginas, e possui uma valiosa natureza multilíngue.

Assim, na pesquisa de EIDBDM que propusemos realizar, optamos por também utilizar a web como corpus e o Google como motor de busca, por ser este último um gerenciador capaz de procurar informações em mais de 4,28 bilhões de páginas de texto, em apenas alguns segundos. Desse modo, estabeleceu-se o limiar de frequência dos fraseologismos italianos seguindo a metodologia proposta por Xatara (2008) e Rios (2010).

A pesquisa fraseológica tendo a web como corpus e o estabelecimento de limiar de frequência

O grande problema dos lexicógrafos é estabelecer, com maior cientificidade, o limiar de frequência de determinada unidade fraseológica (UF) que possa garantir sua presença em um dicionário. Como dito anteriormente, de acordo com Colson (2007), os idiomatismos em corpora tradicionais ocorrem com pouca frequência. O que acontece é que se um pesquisador procura por um idiomatismo em particular, em determinado corpus, a frequência relativa não será muito alta. Os corpora linguísticos oferecem um padrão útil para medir a frequência dos idiomatismos: número de ocorrências por milhão de palavras (PMW). O autor afirma ainda que, em algumas línguas europeias, como o francês e o alemão, por exemplo, idiomatismos verbais correspondem à frequência de menos de 1 PMW.

Considerando-se a estimativa de ocorrência dos fraseologismos na rede a que chegou Colson (2003) (1 ocorrência por milhão de palavras - 1 PMW), e levando-se em conta o número de páginas em francês presentes na web (aproximadamente 200 milhões), esse pesquisador chegou ao limiar de frequência de 200 ocorrências nesse idioma.

Assim, em um estudo sobre unidades fraseológicas do francês e do português, Xatara (2008) estabeleceu o limiar de frequência para elas tanto com base nos estudos realizados por Colson (2003), quanto nos desenvolvidos por Kilgarrif e Grefenstette (2000, 2004), Evans *et al*. (2004) e União Latina (2006). Estimando a existência de 200 milhões de páginas em francês na internet (segundo afirma Colson) e de 70 milhões de páginas em português, das quais 56 milhões seriam em português brasileiro, a autora determina o seguinte:

> [...] como normalmente cada EI ocorre uma vez em cada página da web, definiu-se o limiar de frequência em 56 ocorrências para o português e 200 para o francês. Essas ocorrências equivalem, na realidade, aos resultados oferecidos por intermédio do buscador Google, que na prática funciona como um gerenciador de texto (Xatara, 2008, p. 772).

No que se refere à nossa pesquisa em italiano, para descobrir o número de páginas italianas (.it) existentes na web, foi necessário consultar o site http://www.nic.it/, que registra o número de domínios italianos desde 1987. Em 2005, os domínios italianos alcançaram a marca de 1 milhão; cinco anos depois, em 2010, o número dobrou para 2 milhões de domínios registrados; em 2013, esse número passou a 2,5 milhões de domínios italianos. Em novembro de 2014, esse site registrou 2.745.877 páginas; e agora, em 2016, esses números chegaram a 2.948.872 de páginas com a extensão .it. Nessa esteira, o limiar de frequência do italiano seria de aproximadamente 3 ocorrências. Percebe-se que o limiar da língua francesa é muito mais alto em relação à língua italiana, pois há que se considerar que a quantidade de países que têm como língua oficial o francês (França, Bélgica, Canadá, alguns países da África) é bem maior do que aqueles que falam italiano.

Assim, para garantir que as EIs selecionadas para a pesquisa em pauta fossem realmente frequentes, decidiu-se trabalhar com uma margem probabilística maior, elevando nosso limiar de frequência para 10 ocorrências. Essa marca pautou o limite mínimo para a escolha dos fraseologismos de baixa dedutibilidade metafórica frequentes na língua, os quais passaram a constituir o corpus da pesquisa.

É fato que a falta de lematização do corpus dificulta a pesquisa dos idiomatismos. Para tanto, a estratégia utilizada foi digitar o núcleo da EI (sintagma nominal principal), por exemplo, "alla carlona", sempre entre aspas e, a fim de evitar uma busca muito restrita, um dos termos da expressão foi substituído por um asterisco ("* alla carlona"), em grande parte das vezes, o verbo. Por exemplo, para a EI "fare le cose alla carlona", obteve-se:

• "fare le cose alla carlona": 1.120 resultados.

Em geral, a expressão completa (com o verbo, inclusive), entre aspas, restringe mais o número de ocorrências.

• "le cose alla carlona": 2.310 resultados.

Digitar a expressão sem a presença do verbo costuma aumentar o número de ocorrências.

• "alla carlona": 40.800 resultados.

A presença apenas do núcleo da expressão também tende a aumentar o número de ocorrências.

• "* alla carlona": 45.000 resultados.

Acrescentar um asterisco, antes do núcleo da expressão, possibilita o aparecimento apenas de expressões

que contenham algum elemento antes do núcleo, porém, permite ainda o aparecimento de outros verbos além daqueles que frequentemente acompanham a expressão.

Como se pode observar, quando a busca acontece apenas pelo núcleo da EI, a quantidade de resultados é maior. Esse recurso utilizado ao se substituir uma palavra pelo asterisco é um artifício que visa a suprir a falta de ferramentas linguísticas de pesquisa na web, já que não existe um motor de busca específico para isso.

No caso de se desejar analisar mais de uma língua, aconselha-se a utilização da página do Google específica do idioma em questão (por exemplo: *.it* – para o italiano; *.fr* – para o francês; *.pt* – para o português). Além disso, o Google também possui algumas ferramentas de pesquisa muito úteis, como a seleção de páginas escritas apenas em uma variante de um determinado idioma (como *.pt* – para o português de Portugal; e *.br* – para o português do Brasil).

Desse modo, com o limiar de frequência estabelecido em 10 ocorrências (número que provém do número de 10.000 resultados encontrados por meio do gerenciador de buscas Google), começou-se a pesquisar as expressões italianas. A primeira expressão que não atingiu o limiar de frequência foi *andare a Babborivegoli*, que significa "morrer". Foram encontrados na web 4.900 resultados (4,9 ocorrências), mas como o limiar foi estipulado em 10, essa expressão não foi incluída no glossário proposto. Outra expressão que não foi selecionada é *essere l'ombra di Banco*, que significa "ser a recordação obsessiva de uma má ação, de uma culpa". Por meio da busca, foram encontrados 5.740 resultados (5,7 ocorrências), ficando, também, aquém do limite estabelecido. Já expressões como *l'Achille sotto la tenda* (que significa "abster-se de um ato de maldade, que traria danos a si próprio") e *essere/fare un ambaradan* (que significa "criar um grande tumulto e confusão" ou "ser uma confusão, conjunto caótico", refere-se também a situações e grupos de pessoas) alcançaram, respectivamente, 73.900 (73 ocorrências) e 119.000 (119 ocorrências). Após esse percurso metodológico, foi criado um glossário bilíngue com 72 EIBDM. Cada verbete desse glossário possui informações como: origem, correspondente idiomático em português, frequência e contextos de uso.

Considerações finais

Pode-se concluir que a web como corpus, além de ser uma ferramenta para verificação da frequência e da ocorrência de expressões, é uma fonte quase inesgotável de extração de exemplos, principalmente no que se refere ao estudo de EIBDM, uma vez que essas EI não figuram em grande quantidade nos corpora tradicionais. Caberá, contudo, ao pesquisador saber interpretar os dados conseguidos por meio dessa ferramenta, bem como identificar sua utilidade.

Assim, apesar das ressalvas apresentadas neste artigo quanto ao uso da WaC, não há dúvida de que a sua utilização representa um avanço e benefício a diversas disciplinas, inclusive à Fraseologia. Espera-se que, em um futuro bem próximo, seja dada a atenção devida à web, de modo que (novas) ferramentas de pesquisa mais eficientes e refinadas sejam criadas e desenvolvidas, na expectativa de torná-la um instrumento de investigação fraseológica com maior eficiência.

Referências

COLSON, J.P. 2003. Corpus linguistics and phraseological statistics: a few hypotheses and examples. *In*: H. BURGER; A. HÄCHI BUHOFER; G. GRÉCIANO (eds.), *Flut von texten – vielfalt der kulturen. Ascona 2001 zu Methodologie und kulturspezifik der phraseologie*. Baltmannsweiler, Schneider Verlag Hohengehren, p. 47-59.

COLSON, J.P. 2007. The World Wide Web as a corpus for set phrases. *In*: H. BURGER; D. DOBROVOL'SKIJ; P. KÜHN; N. NORRICK (eds.), *Phraseologie / Phraseology*. Berlim/New York, Mouton de Gruyter, 2007, p. 1071-1077.

CORPUS E LESSICO DI FREQUENZA DELL'ITALIANO SCRITTO (CoLFIS). [s.d.]. Disponível em: http://www.istc.cnr.it/material/database/colfis/. Acesso em: 06/02/2017.

CORIS. 2017. Disponível em: http://corpora.dslo.unibo.it/coris_eng.html. Acesso em: 06/02/2017.

ENCICLOPEDIA TRECCANI. [s.d.]. Disponível em: http://www.treccani.it/enciclopedia/corpus_(Enciclopedia-Italiana)/. Acesso em: 25/08/2013.

EVANS, D.; GREFENSTETTE, G.; VAN GENT, J.; VOSSEN, P. 2004. *The multi-lingual web*. Disponível em: http://www.infonortics.com/searchengines/sh04/04pr. Acesso em: 02/03/2014.

FELBER, H. 1984. *Terminology Manual.* Paris, Unesco/Inforterm, 426 p.

FLETCHER, W. 2005. Concordancing the web. *In*: M. HUNDT *et al., Corpus Linguistics and the Web.* Amsterdam, Rodopi, p. 1-22.

KILGARRIFF, A.; GREFENSTETTE, G. 2003. Introduction to the Special Issue on the Web as a Corpus. *Computational Linguistics*, **29**(3):333-347. https://doi.org/10.1162/089120103322711569

LAPUCCI, C. 1990. *Il dizionario dei modi di dire ed espressioni idiomatiche*. Milão, Vallari, 520 p.

ORTÍZ-ALVARE. 2000. *Expressões idiomáticas do português do Brasil e do espanhol de Cuba: Estudo contrastivo e implicações para o ensino de português como língua estrangeira*. Campinas, SP. Tese de Doutorado. Universidade Estadual de Campinas, 334 p.

RADICCHI, S. 1985. *In italia: modo di dire ed espressioni idiomatiche.* 4ª ed., Roma, Bonacci, 208 p.

QUARTU, M.; ROSSI, E. 2012. *Dizionario dei modi di dire della lingua italiana.* Milano, HOEPLI, 528 p.

RIOS, T.H.C. 2010. *A descrição de idiomatismos nominais: proposta fraseográfica português-espanhol.* São José do Rio Preto, SP. Tese de Doutorado. Universidade Estadual Paulista, 243 p.

SARDINHA, T.B. 2000. Linguística de Corpus: Histórico e Problemática. *D.E.L.T.A*, **16**(2):323-267. https://doi.org/10.1590/s0102-44502000000200005

SARDINHA, T.B. 2004. *Linguística de Corpus.* Barueri, Manole, 410 p.

SCHRYVER, G.M. 2002. Web for/as Corpus: a Perspective for the African Languages. *Nordic Journal of African Studies*, **11**(2):266-282.

TERMIGNONI, S. 2010. *Fare come l'asino del pentolaio.* Porto Alegre, EDIPUCRS, 100 p.

TONFONI, G.; TURBINATI, L. 1995. Visualizzazione dei processi di traduzione: i proverbi e le espressioni idiomatiche. *In*: AA.VV., *La Traduzione*. Saggi e Documenti II. Roma, Ministero per i Beni Culturali e Ambientali, Divisione Editoria, p. 239-252.

UNIÃO LATINA. 2016. *Lenguas y culturas en la red 2005*. Disponível em: http://dtil.unilat.org/LI/2005/index_es.htm. Acesso em: 03/09/2013.

VALENÇA, E.M. 2014. *A tradução de expressões idiomáticas de baixa dedutibilidade metafórica: contribuições aos estudos fraseológicos bilíngues.* São José do Rio Preto, SP. Dissertação de Mestrado. Universidade Estadual Paulista, 190 p.

XATARA, C.M. 2008. A web para um levantamento de freqüência. *In*: J.S. MAGALHÃES; L.C. TRAVAGLIA (orgs.), *Múltiplas perspectivas em lingüística.* Uberlândia, EDUFU, p. 770-777.

ZULUAGA, A. 1997. *Introducción al estudio de las expresiones fijas.* Madrid, Studia Romanica et Linguistica.

Submetido: 28/07/2016
Aceito: 24/11/2016