Criação de um corpus específico: expressões idiomáticas à temática da tourada em notícias da web

KARINA BIBIANO SILVA
Universidade de São Paulo

Resumo: Neste artigo será apresentado uma metodologia para a criação de um corpus personalizado de expressões idiomáticas referentes à temática da tourada. O corpus é formado por textos extraídos de sites de notícias produzidas na Espanha, coletados e organizados por meio da linguística computacional. O objetivo é descrever com detalhes os critérios de criação do corpus, assim como investigar a frequência e a usabilidade das referidas expressões idiomáticas no âmbito de notícias jornalísticas, em sua modalidade escrita no ambiente digital.

Palavras-chave: Expressões idiomáticas, Toro, Corpus, Web, Linguística de Corpus.

Abstract: In this article a methodology will be presented for the creation of a personalized corpus of idioms related to the theme of bullfighting. The corpus consists of texts extracted from news sites produced in Spain, collected, and organized through computational linguistics. The objective is to describe in detail the criteria for creating the corpus, as well as to examine the frequency and usability of the idioms in the scope of news reporting, in its written modality in the digital environment.

Keywords: Idioms, Bull, Corpus, Web, Corpus Linguistics.

1. Introdução

A presente pesquisa baseia-se nos estudos da Linguística de Corpus, a fim de verificar as ocorrências de expressões idiomáticas (doravante EI), relacionadas ao universo da tourada[1], publicadas em textos jornalísticos, no meio digital. O propósito será explicar a criação de um corpus exclusivo, especificando critérios para, assim, disponibilizá-lo para

[1]Pesquisa que faz parte do meu projeto de mestrado no *Programa de Pós-Graduação em Língua Espanhola e Literaturas Espanhola e Hispano-Americana* (USP -FFLCH).

análise. Será apresentado o recorte de uma pesquisa desenvolvida com a intenção de examinar essas expressões, valendo-se da *web* de notícias como corpus, e assim analisar a frequência de uso e os assuntos onde há mais ocorrências.

Para isso, foram coletados textos com o termo *toro* (em português, touro), retirados de sites de notícias da região da Espanha. A escolha em trabalhar com textos jornalísticos da *web* se deve ao fato de que, como as notícias se destinam a um público amplo e buscam uma comunicação imediata e simples com os leitores, há maior possibilidade de que apareçam, nesses tipos de textos, a coloquialidade presente com o uso de alguns elementos linguísticos como no caso, as EI.

Este estudo fará uso da Linguística de Corpus, que estuda a linguagem por meio de evidências vivenciadas de um dado conjunto linguístico, muitas vezes colhidas eletronicamente. De acordo com Viana (2010: 34), trata-se de "uma forma de investigação empírica da linguagem a partir da exploração sistemática de um corpus". Para Berber Sardinha (2004:30), a Linguística de Corpus trabalha com "a visão da linguagem como sistema probabilístico, a qual pressupõe que, embora muitos traços linguísticos sejam possíveis, teoricamente, não ocorrem com a mesma frequência". As informações extraídas do uso real da linguagem fazem com que os resultados pesquisados tenham maior autenticidade, conferindo credibilidade à pesquisa. É importante evidenciar também o uso da *web* para a criação do corpus, um recurso de pesquisa linguística produtivo e, muitas vezes, mal explorado para os estudos fraseológicos e lexicográficos.

2. *Corpora* versus *web como Corpus*

Com o advento da tecnologia, a Linguística de Corpus ganhou destaque e relevância nos estudos linguísticos, pois graças a esse avanço, há diversas ferramentas de pesquisa para a seleção de textos na criação de um corpus. Mas o que seria um corpus? Segundo Tagnin (2013:29),"é uma coletânea de textos, necessariamente em formato eletrônico, compilados e organizados segundo critérios ditados pelo objetivo de pesquisa a que se destina".

É possível encontrar na internet vários *corpora* disponíveis para consulta no idioma espanhol. Um deles é o *CREA*[2], o corpus de referência do espanhol atual, da *Real Academia Española (RAE)*, atualizado e com textos de diferentes gêneros textuais. Da *RAE* ainda temos o *CORPES XXI*[3]–cujo diferencial é estar constituído somente por textos literários do século XXI– e igualmente o *CORDE*[4], que forma o corpus diacrônico. Finalmente, temos o *Corpus del Español*[5], desenvolvido pelo Instituto BYU (*Brigham Young Universtity*), com mais de dois bilhões de palavras de diferentes tipos de textos em sua base de dados.

No entanto, para este estudo, como serão analisadas EI com o termo *toro* nos textos, faz-se necessário criar um corpus exclusivo para que esse critério da pesquisa fosse contemplado, posto que anteriormente, pesquisando *corpora* já existentes, notou-se que havia poucas ocorrências no gênero jornalístico. Por isso, foi utilizada a *web* para coletar textos escritos, de sites de notícias mais populares da Espanha para constituir o corpus da pesquisa.

E para uma breve exemplificação, foi feita a comparação da busca entre o *Corpus del Español*, o *CREA* e a ferramenta de busca do Google, focalizando somente a região da Espanha, exposto na tabela 1, com a frequência da expressão, *a toro passado* escolhida aleatoriamente, para o exemplo.

	Corpus del Español	***CREA***	***Google***
a toro pasado	514	27	64.500

Tabela 1:Ocorrências referente à expressão "*a toro pasado."*

A consulta no Google (sistema mais popular de busca na *web*), foi feita no modo "Pesquisa avançada", na barra de configuração da ferramenta, e incluído a expressão a *toro pasado*, no campo de busca por expressão, no idioma espanhol e a região Espanha. Foi encontrado um número superior de casos (64.500) comparado aos outros dois *corpora*. Como no *Corpus del Español* e sobretudo no *CREA*, o número de ocorrências é expressivamente menor do que no Google, pode-se supor,

[2]Disponível em http://corpus.rae.es/creanet.html.

[3] Disponível em http://web.frl.es/CORPES/view/inicioExterno.view.

[4] Disponível em http://corpus.rae.es/cordenet.html.

[5] Disponível em http://www.corpusdelespanol.org/x.asp.

precipitadamente, que essa expressão idiomática é pouco utilizada e irrelevante na linguagem. No entanto, segundo afirma Rios (2010: 70) "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". Logo, com essa constatação renunciou-se aos *corpora* existentes e estendida a busca para a *web*.

Por esse meio, obteve-se resultados mais significativos de casos. O que faz presumir que, mesmo a *web* possuindo um vasto conteúdo, de diferentes tipos de gêneros, continua sendo um bom meio para se fazer pesquisas relativas à linguagem, dado que é de livre acesso, com uma ampla variedade de textos e com enumeras fontes nacionais e internacionais de pesquisa, podendo filtrar a busca de acordo com o foco do estudo.

Além da *web* ser utilizada como um corpus, é permitido também enxergá-la como geradora de conteúdo para criar um corpus próprio. Nesse processo de criação, é importante se ater aos próprios critérios estipulados para a pesquisa, que serão determinantes para moldar o perfil do corpus, e ter maior controle dos textos introduzidos na base de dados.

3. A representação do touro

Apesar de ser considerada por muitos uma tradição polêmica, a tauromaquia, duelo de habilidade e resistência levados até a morte entre touro e toureiro, possui uma forte influência na cultura espanhola ainda hoje, embora essa prática esteja sendo extinta por alguns governos locais. Um pouco menos controversa que as corridas de touros, quiçá seja mesmo a sua origem. Segundo alguns estudiosos como Alcantud (1999:67), a tradição foi originária dos romanos com os sacrifícios dos touros em honra aos escravos mortos, logo substituídos por gladiadores. O imperador Constantino foi quem ordenou que incluíssem as feras para enfrentarem os homens, transformando o evento em um espetáculo de caça. Segundo o mesmo autor, há uma relação também com os árabes e o paganismo na Antiguidade, mas ainda não há uma posição oficial da Igreja sobre isso.

Analisando a História, há registros ainda mais antigos de homens enfrentando touros nos primórdios da civilização em pinturas rupestres de

aproximadamente 13.000 a.C[6], na era paleolítica, encontradas em cavernas da França, como por exemplo, a caverna de Lascaux[7], descoberta em 1940 e que possui em seu interior várias gravuras. Na câmara principal está a sala chamada "Sala dos Touros" com cerca de 130 figuras de animais, entre eles um grande touro que se destaca com o nome de "Primeiro Grande Touro". Outra gruta cujas pinturas rupestres foram encontradas com representações de touros está localizada em Altamira, na Espanha, também pré-histórica (11.000 a.C.), e descoberta em 1879. Enfim, o touro é um animal com muita representatividade histórica ao longo da vida humana.

A corrida de touro é considerada uma festividade e está incluída no calendário de festas anuais na Espanha, atraindo muitos aficionados e turistas de todo o mundo. E apesar de ser um tema polêmico, está associada a alguns santos religiosos. Uma delas homenageia São Firmino[8], na famosa e mais tradicional festa espanhola que ocorre no mês de julho em Pamplona, no norte da Espanha, atraindo um grande público. Anteriormente a esta data, no domingo de Páscoa, se inicia a *Feria de Abril*[9] em Sevilha, onde a temporada taurina é inaugurada com a corrida de *Domingo de Resurrección* em homenagem à Virgem de Macarena, padroeira dos toureiros, na qual participam as grandes celebridades taurinas.

E, apesar de estar associada à Espanha, a tourada tem tradições em diversos países como França, Portugal, México, Colômbia, Guatemala, Peru e Venezuela. E em cada um desses países existe algumas diferenças na técnica, mas no geral são muito similares; porém é na Espanha que, conforme Fuentes (2001:16), a figura do touro ganha o emblema de estereótipo nacional; de fato a influência na língua espanhola também é diferente segundo o pais (Pamies 2020).

E, para adentrar-se na linguagem da tourada, é pertinente entender e conhecer o que se passa durante o ritual, que possui seu próprio

---

[6]Disponível emhttps://hav120151.wordpress.com/2015/04/06/o-primeiro-grande-touro-de-lascaux/. Acesso em 20/04/2020.

[7] Disponível em https://www.donsmaps.com/lascaux.html. Acesso em 20/04/2020

[8]São as festas em honra a São Firmino, que se realizam em Pamplona, no norte de Espanha. (Fonte: https://pt.wikipedia.org/wiki/Festas_de_S%C3%A3o_Firmino).

[9]A Feira de Abril ocorre em Sevilha, na Espanha e a temporada taurina é inaugurada com a corrida de *Domingo de Resurrección*, na qual participam as grandes figuras do momento. (Fonte: https://www.andalucia.org/es/conoce-andalucia/arte-cultura-y-tradiciones/feria-de-abril/la-feria-taurina)

regulamento. Segundo Luque Duran e Manjón Pozas (1998a:55), em uma única corrida de touros, três toureiros afrontam seis touros, sendo dois para cada um. Mas antes do início, há uma espécie de procissão na arena formada por todos os que diretamente participarão da corrida: cavaleiros, toureiros, membros de cada quadrilha formada por banderilheiros e picadores, e ao final os moços e os mulas de arrastre, responsáveis em retirar o touro morto da arena. Na ponta da arena está a presidência que entrega as chaves aos cavaleiros, simbolicamente, para abrir a porta de onde saem os touros, e assim iniciar o espetáculo. O toureiro mais velho começa e a corrida se divide em três partes chamadas "tercios". No primeiro tercio, o toureiro utiliza o capote, uma espécie de capa de cor vermelha e amarela para enganar e tourear o animal. No segundo tercio, é a parte em que o toureiro coloca três pares de banderilhas, uma espécie de vara de madeira com uma lança na ponta, que é cravada nas costas do touro. E no terceiro último tercio, o toureiro, utilizando uma muleta de tecido vermelho na mão, para driblar e atrair o touro, precisa finalmente, cravar a espada no coração do touro, para matá-lo de uma vez diante do público. Se o toureiro conseguir matar o touro, o público na arquibancada acena com lenços brancos, pedindo à presidência que o premie, normalmente, com as orelhas do touro e até mesmo o rabo. Caso a presidência seja pouco generosa, o público pode protestar. O toureiro finaliza a corrida com uma volta na arena, sendo aplaudido e levando seu troféu, e no caso de êxito, sai pela porta principal carregado nos ombros pela multidão presente.

Na tauromaquia, a figura do touro ganha uma alegoria de prestígio, desafiado pelo toureiro até a sua morte, e desse modo, está associado à força e ao perigo. O "espetáculo" envolvendo touro e toureiro, e favorece a criação de muitas EI utilizadas no meio social. Situando-se na linguagem, segundo Abella (1996:66), as expressões taurinas foram incorporadas à língua espanhola entre o século XV e XVI:

> "[...] que a maioria das expressões surgem por volta dos séculos XV e XVI, quando as festas das touradas tinham uma grande presença na vida diária das cidades e aldeias espanholas."[10]

[10] No original: "[...] que la mayoría de los refranes surgen en torno a los siglos XV y XVI, cuando los festejos taurinos tenían una gran presencia en la vida cotidiana de los pueblos y villas españolas." (tradução da autora)

Algumas expressões, foram encontradas, citadas no *Diccionario de la Real Academia* editado, em 1780 como por exemplo, *Para torear y para casarse hay que arrimarse*, ou *Ver los toros desde la barrera*. Hoje, elas são comumente encontradas em muitos titulares de jornais escritos, como forma de chamar a atenção, persuadir, mostrar domínio sobre determinado assunto, além de também estabelecer uma proximidade com o leitor/interlocutor.

Seguindo essa lógica, pode-se incluir também a ideia proposta por Xatara (1998:156), que diz que:

> "As EI, por sua vez, encontram-se, em sua grande maioria, no nível coloquial: linguagem informal, que usa palavras novas, imagens pitorescas, sentidas como "anormalidades", sem que a frequência de seus desvios constitua uma deformação que torne "inaceitáveis" as mensagens dadas - ter muita cera no ouvido, arriscar a pele, cheirar a defunto, vender seu peixe etc. O uso das EI nesse nível coloquial denota, na verdade, intimidade entre os interlocutores em uma situação de comunicação descontraída (Peytard & Génouvrier, 1970)".

Fazer a busca de EI em notícias de jornais é um desafio, e ao mesmo tempo uma oportunidade de entender, dentro de uma comunidade linguística, que compartilham da mesma língua, quando e como a utilizam. A linguagem jornalística, usada em sites populares de notícias consegue aproximar o leitor com o uso de EI como forma de interação, e foge de termos ou palavras rebuscadas, inclusive em seus intertítulos, utiliza essas expressões, muitas vezes, com um toque de humor.

4. Métodos para a criação do corpus

Em concordância com Berber Sardinha (2004: 97), para se criar um corpus computadorizado, os textos escritos ou falados têm que ser autênticos, ou seja, de falantes nativos; devem ser escolhidos criteriosamente conforme as necessidades da pesquisa e, por último, o corpus deve conter elementos significativos que representem as características do idioma dos falantes. Com base nessas premissas, foram definidos os critérios para a construção do corpus da pesquisa, como será visto a seguir.

4.1. Descrição dos parâmetros

De acordo com Viana (2010:30), é possível identificar diferentes tipos de *corpora*: monolíngue, bilingue ou multilíngue. Classificá-lo como sendo um corpus paralelo, composto de textos originais e um ou mais *corpora* e suas traduções; ou como sendo um corpus comparável constituído de *corpora* originais em duas ou mais línguas. Na sua estrutura, os *corpora* podem ser formados por textos escritos ou orais. Para criar um corpus é importante também definir: qual gênero textual ele pertence (por exemplo, se são cartas, notícias de jornais, verbetes, manuais de instruções etc.); de quando é o texto; qual a área específica; a fonte dos textos (*web* ou livros escaneados); se são apenas trechos ou textos completos; a proporção do número de palavras em cada textos para que tenham um balanceamento no tamanho do corpus; e por fim, se o texto é aberto (continua sendo alimentado por novos textos) ou fechado (finalizado e inalterado).

Conforme determinações pensadas para este estudo, ficou definido que será elaborado um corpus monolíngue, em espanhol, formado por textos completos e escritos, extraídos de *web* de notícias, entre o período de 2010 e 2020. Os sites serão da região da Espanha, de domínio público, e não serão considerados textos relacionados à atividade da tauromaquia. Não será delimitado o número de palavras por texto compilado, visto que o corpus criado não será comparado com outro paralelo. Impreterivelmente, nos textos coletados também deve conter o termo *toro*.

A escolha pelo culturema *toro*, se deve ao fato de que ele representa uma figura alegórica de um país, e está presente na linguagem oral e escrita, formando diversas EI, e desse modo, constitui um símbolo importante neste estudo de corpus. Seguindo o conceito de Pamies Bertrán (2007), define-se culturemas, símbolos extralinguísticos culturalmente motivados, servindo de modelo para a criação de EI, ou seja, um elo entre a cultura e as unidades linguísticas, cujos símbolos são reconhecidos por um grupo de pessoas. Essas unidades fraseológicas têm influência em diversos eventos do cotidiano como festas, filmes, culinária, religião, entre outros, e acabam criando uma consciência social entre os falantes, neste caso, das touradas.

4.2. Planejamento do corpus

Para uma melhor visualização, o processo de criação do corpus está dividido em etapas, de acordo com os parâmetros estipulados para esta pesquisa.

Etapa 1: Busca de textos na *web*.

Em um primeiro momento, foi utilizada aferramenta de busca do Google no modo "Pesquisa avançada" e configurada para o idioma espanhol, inserindo o termo *toro* no campo de busca por palavras, e delimitada somente a região Espanha. Porém, como o Google indicou apenas seis páginas de busca, e estas, bem poucas seguiam os critérios estipulados, ou seja, não eram páginas de notícias, foi necessário utilizar outra estratégia de pesquisa, como acessar os sites dos jornais mais conhecidos do país e usar suas próprias ferramentas de busca, a escolha foi aleatória. Alguns dos *sites* selecionados da Espanha foram: *El Confidencia, El Mundo, El País, ABC, La Razón, El Imparcial, El Periódico, Cope, El Diario, El Correo* e *Última Hora*.

Contudo, também fazendo uso desse método foi detectado outros contratempos. Por exemplo, quando se buscava o termo *toro*, também foram encontrados termos sobre corridas de touro, *plaza de toros* ou nomes próprios como *Toro Rosso, Benicio del Toro* e da cidade chamada *Toro*(na província de Zamora). Outros problemas apontados foram: ausência do campo de busca (*La Razón*), número limitado de buscas (*El Correo*) ou buscas somente no título e não no corpo dos textos (*Última Hora*). Mas apesar desses percalços, foi possível coletar um número significativo de textos para a formação do corpus.

Etapa 2: Compilação dos textos com os devidos critérios.

Para os parâmetros, foram seguidas as seguintes regras: textos no idioma espanhol; notícias publicadas entre os anos de 2010 e 2020; *sites* da região da Espanha; textos com o termo *toro* no seu conteúdo, excluindo matérias sobre tauromaquia. Ao todo, foram coletados 265 textos.

Figura 1: Exemplo de fragmento de texto selecionado da internet[11]

Etapa 3: Etiquetagem (*markup*).

Na fase de etiquetagem, os textos completos foram copiados e colados no bloco de notas em formato .txt (Figura2), recebendo uma etiqueta no cabeçalho com as seguintes informações:

-Fonte do *site (URL)*;

- Assunto relacionado (por exemplo: esporte, política, economia etc.).

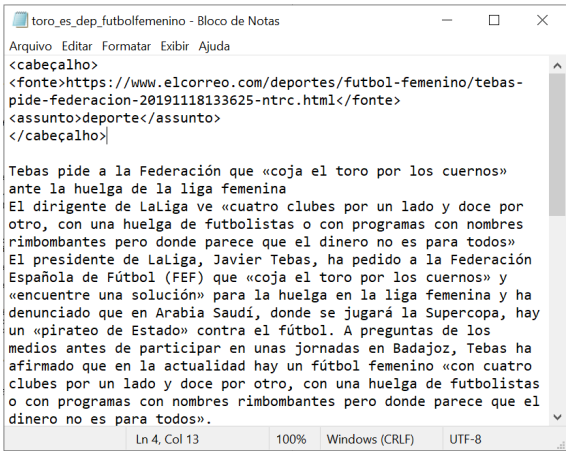

Figura 2: Exemplo de um texto colhido e etiquetado

[11]Extraído do *site*https://www.elcorreo.com/deportes/futbol-femenino/tebas-pide-federacion-20191118133625-ntrc.html - Acesso em: 08/02/2020

Etapa 4: Nomeação dos arquivos.

No momento de nomeá-los e salvá-los no formato.txt, os arquivos receberam um código de etiqueta, de acordo com seu assunto, tal como descrito no quadro 1.

Quadro1. Códigos e assuntos

ASSUNTOS	CÓDIGO
Cultura	cul
Acontecimentos	suc
Esporte	dep
Economia	eco
Natureza	nat
Política	pol
Saúde	sal
Tecnologia	tec

Nesta etapa, o arquivo foi nomeado com o tema da pesquisa, as iniciais do país (ES para Espanha), o código do assunto e, por último, uma palavra-chave referente ao texto, como no exemplo abaixo:

toro_es_dep_futbolfemenino

É importante comentar que os textos copiados para o arquivo .txt, sofreu uma limpeza, onde foram retirados todos os elementos que não eram importantes para a pesquisa, deixando apenas os textos, a URL e o assunto, classificado pelo próprio jornal.

Desse modo, todos os arquivos foram especificados por temas e etiquetados com as informações necessárias para ajudar na pesquisa. Concluído este processo, os textos foram guardados em uma pasta separadamente, para que posteriormente, pudesse ser transferida para a ferramenta de software *AntConc* (Anthony, 2012)[12],que ajudará na extração de dados do corpus.

[12] *AntConc*: http://www.antlab.sci.waseda.ac.jp/antconc.index.html

5. Uso do *software* e análise do corpus

Após coleta e nomeação dos textos, foi feito o *upload* do arquivo no programa *AntConc*.

Através dessa ferramenta foi possível gerar listas de frequência e linhas de concordância, utilizando comandos como o *Word List*, que apresenta a lista das palavras mais recorrentes, o *Cluster/N-Gram* que detalha melhor as principais expressões e o *Concordance* que faz a concordância dos termos selecionados.

De acordo com a recolha efetuada, ao todo, foram 449 ocorrências com o termo *toro* no corpus compilado.(Figura 3).

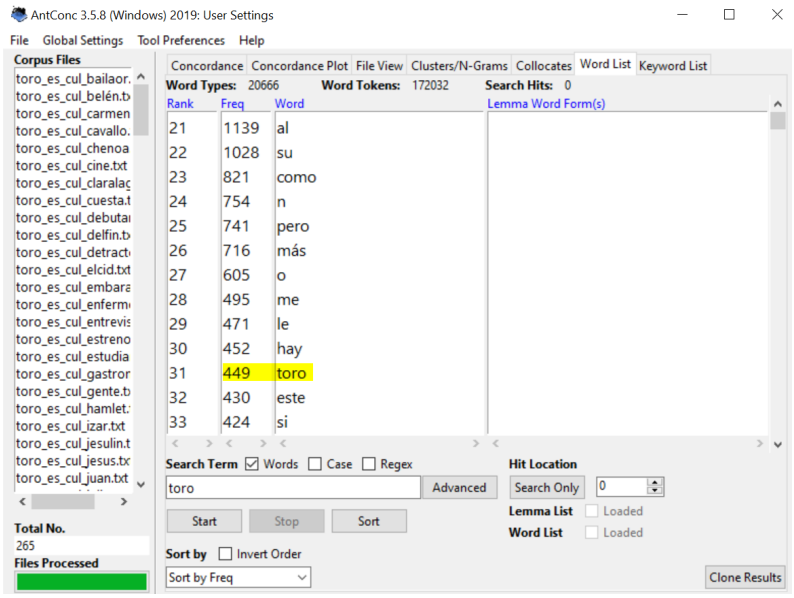

Figura 3: *AntConc - Word List*

Todavia, para refinar essa frequência, foi criado manualmente, uma lista de palavras em espanhol denominada *Stopwords*, formada por palavras que não aportam conteúdo relevante para esta pesquisa (artigos, preposições e advérbios, entre outros).

Ela foi anexada em *Tool Preferences*, no software *AntConc*, para priorizar a aparição de mais substantivos. Depois de atualizada a página, o termo *toro* lematizado *(toro* e *toros)* apareceu nas primeiras posições, seguido por *partido* na terceira posição; *España* na quinta e *gobierno* y *equipo* logo em seguida, apontando as palavras de conteúdo mais recorrentes no corpus (Figura 4).

Através da função *Word List* tentou-se relacionar os assuntos, com base nas palavras de maior frequência: *partido, gobierno y equipo*. Contudo, utilizar essa estratégia foi inviável pelo fato de algumas palavras transitarem entre diferentes temas, por exemplo, *partido* e *presidente* podem estar tanto em política, quanto em esporte.

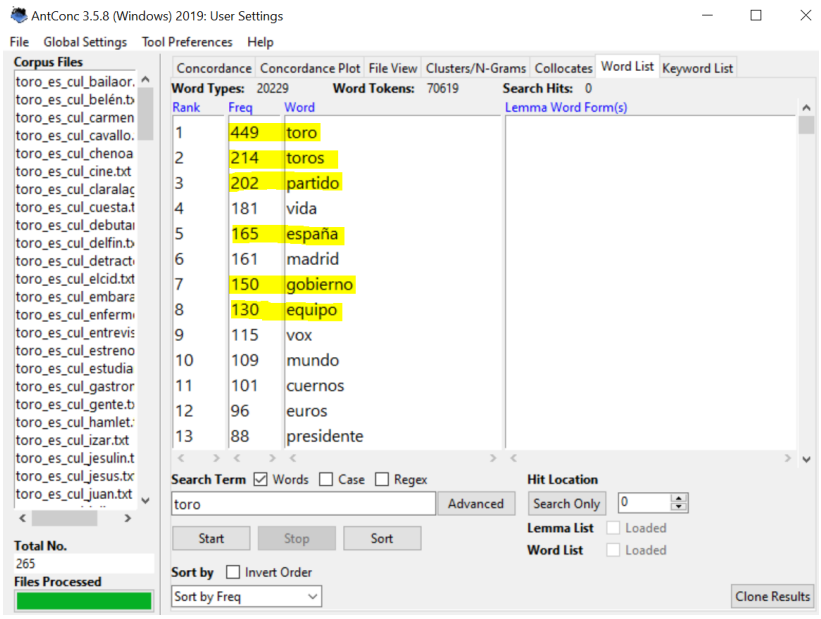

Figura 4.*Word lis*te liminando as *Stopword*

Assim, foi necessário utilizar outra função do *AntConc*, o *Cluster/N-Grams*, que serve para mostrar uma sequência de "n" elementos que se repetem no texto —foi estabelecido o número de cinco palavras — junto com a palavra de pesquisa *toro*, o que ajudou a visualizar as expressões com mais ocorrências.

Ao clicar em cada linha, ou seja, em cada N-Gram, a página é automaticamente direcionada para a função *Concordance*, onde esse grupo de palavras (nódulo) está contextualizado.

Lematizou-se as expressões para a busca, isto é, os verbos que acompanhavam as expressões foram lematizados com o uso do asterisco na parte fixa, o que facilitou a busca dos termos relacionado à expressão inteira.

Então, ao clicar na primeira linha, no rank 1 (Figura 5), *el toro por los cuernos*, a página é direcionada para *Concordance*, onde aparece a expressão destacada dentro de um contexto, como apresentado na figura 6, seguinte.

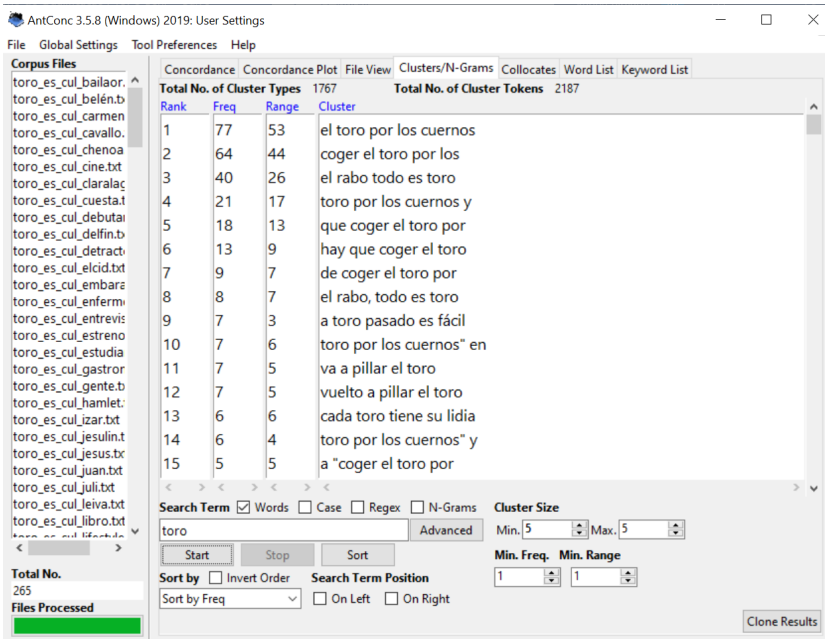

Figura 5.*AntConc - Clusters/N-Grams*

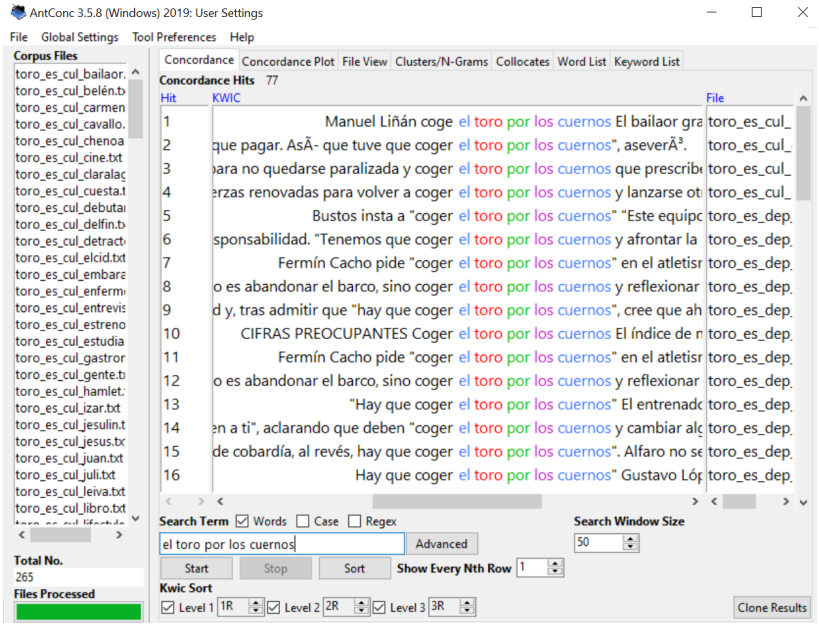

Figura 6. *AntConc – Concordance (rank 1*)

Na coluna *File,* à direita do quadro da figura 6, quando somados todos os assuntos, conforme os códigos das etiquetas, verifica-se que *el toro por los cuernos* possui mais resultados em política (31 ocorrências) e esporte (23 ocorrências) e contempla um total de 77 ocorrências.

Analisando agora o rank 3,na figura 5, o N-Gram [hasta] *el rabo todo es toro*, é possível observar as etiquetas do código dos assuntos, na coluna *File*, que há mais ocorrências em esporte (27 casos) e política (13 casos), com um total de 49. (Figura 7)

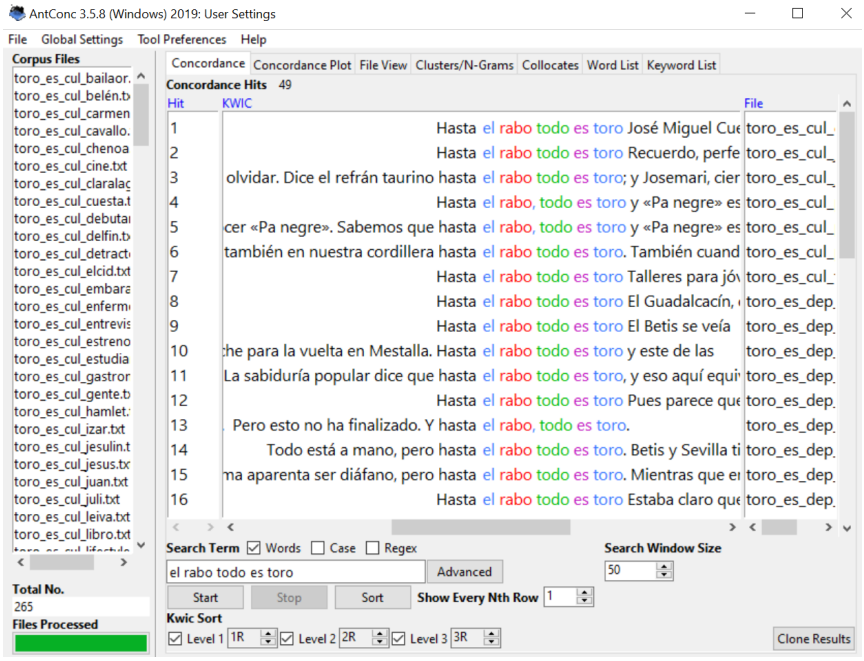

Figura 7:*AntConc – Concordance* (rank 3)

Ao fazer-se uso dos variados instrumentos de análise disponíveis dentro da ferramenta *AntConc*, é possível, quantitativamente, observar a frequência dos dados do corpus, ficando visível o uso do termo *toro* dentro das EI. Também através do método de amostragem foi possível identificar com mais objetividade em quais seções/assuntos elas são empregadas no contexto jornalístico.

Considerou-se analisar até o rank 15, dado que a ferramenta *AntConc* revela uma lista extensa de termos, e o propósito é identificar somente as mais frequentes. A seguir, estão apresentadas as cinco EI encontradas com mais frequência no corpus criado, contextualizadas, com fragmentos retirados das notícias e com o respectivo número de ocorrências, em ordem decrescente.

EXPRESSÕES IDIOMÁTICAS	EXEMPLO DE USO NA WEB DE NOTÍCIAS	OCORRÊNCIAS
Coger el toro por los cuernos	El presidente del PP, Mariano Rajoy, ha criticado la gestión de los socialistas en el Gobierno, tanto en el central como en el gallego, y ha dicho de ellos que "son incapaces de **coger el toro por los cuernos**". *La Voz Digital*	80
Hasta el rabo todo es toro	"Hay que hacer muchas cosas muy bien en ataque y muchas cosas muy bien en defensa. **Hasta el rabo todo es toro**". *Hoy.es*	48
A toro pasado	"**A toro pasado**, todos sabemos lo que habría que haber hecho», aseveró Duque, que insistió: «No creo que se pueda decir que los ministros no han hecho nada." *El Correo*	42
Pillar el toro	"A las empresas les va a **pillar el toro** para adaptarse a las nuevas exigencias europeas de protección de datos." *ABC.es*	37
Cada toro tiene su lidia	"Pensará el ministro que **cada toro tiene su lidia**, que es justo lo que busca cada ministril, su momento de bilateralidad, su gloria, por sus pobres y su nación." *En Castilla la Mancha*	6

Tabela 2: Expressões idiomáticas que incluem o termo *toro*.

Em suma, ao analisar o corpus utilizando os vários comandos disponíveis no *AntConc*, fica evidente a existência das EI, nos textos compilados, possibilitando identificarem quais seções essas expressões são mais empregadas nas notícias da web.

5.1. Resultado: Inventário por assunto

Identificadas as cinco expressões mais empregadas nos textos de jornais de notícias online, agora é possível analisar detalhadamente, em quais assuntos essas EI foram empregadas. (Tabela 3)

Expressões Idiomáticas	Cul	Dep	Eco	Nat	Pol	Sal	Suc	Tec	TOTAL
Coger el toro por los cuernos	6	23	6	3	**32**	0	10	0	80
Hasta el rabo toro es toro	7	**27**	2	0	12	0	0	0	48
A toro pasado	6	10	2	0	18	2	3	1	42
Pillar el toro	9	5	6	0	11	0	3	3	37
Cada toro tiene su lidia	2	0	1	0	2	1	0	0	6

Tabela 3: Resultados das EI por assuntos

Nessa perspectiva, observando a tabela 3, em termos de valores absolutos, é constatado uma predominância das EI nos assuntos relativos à política e ao esporte. Por exemplo, nota-se uma alta ocorrência, da expressão *Coger el toro por los cuernos*, com 32 casos em política e, a expressão *Hasta el rabo todo es toro* com 27 casos em esporte. Esses resultados obtidos, confirmam que as EI são comumente empregada sem textos jornalísticos na *web*, e estão presentes em várias seções.

É interessante notar que, duas categorias tão distintos como esporte e política, possuem em comum as EI referentes ao touro e, que podem transmitir a mesma mensagem ao leitor. Observa-se também que, mesmo socialmente, a tourada não tendo muitos adeptos, na atualidade, as EI relativas a ela, são muito empregadas em notícias, dado que o público leitor, no geral, é capaz de reconhecer seu sentido metafórico, como já mencionado por Pamies Bertrán (2008) acerca do entrelaçamento do cultural com o lexical, que gera a compreensão dessas expressões dentro de uma comunidade.

Seguindo a teoria de Lakoff & Johnson (2002:45), "a metáfora está infiltrada na vida cotidiana, não somente na linguagem, mas também no pensamento e na ação.", sendo construídas nos ambientes socioculturais específicos de cada comunidade de fala, e pode-se notar explicitamente, através das notícias compiladas presentes no corpus.Com efeito, a elevada ocorrência das expressões com o termo *toro* indica que, no processo de criação de um corpus*,* a seleção de critérios é fundamental para garantir que os dados obtidos nas pesquisas sejam confiáveis.

6. Considerações finais

Neste artigo, foi enfatizada a necessidade de se pensar primeiramente, nos parâmetros para construir um corpus específico, ademais de mostrar como aplicar a ferramenta de *software AntConc*. Após definido o objetivo do corpus, os critérios adotados para a sua construção, são de extrema importância, pois é parte fundamental da metodologia de trabalho. Seguindo o conceito de Tagnin (2015: 20), foi assegurado a necessidade de garantir que os textos selecionados fossem representativos do campo da pesquisa, além de possuírem fontes confiáveis, sendo de jornais locais conhecidos.

O percurso descrito, confere que a Linguística de Corpus, utilizada como metodologia, pode colaborar com os estudos linguísticos temáticos, pois disponibiliza recursos à pesquisa, oferecendo ferramentas computacionais facilitadoras para a análise dos dados. Essas ferramentas oferecem dados quantitativos, além de possibilitar a análise qualitativa desses dados compilados, permitindo que se trabalhe com dados reais para averiguar aspectos culturais específicos de um grupo, neste caso, utilizando expressões provindas do meio taurino, na língua espanhola.

Como afirma Xatara (1995: 195), "muitas vezes o léxico de uma língua não dispõe em seu acervo de unidades lexicais apropriadas para expressar certas nuances de sentimento, emoção, ou sutilezas de pensamentos do falante", logo, faz-se uso das expressões com combinatórias inusitadas que buscam um efeito de sentido, e isso independe do perfil do público leitor e de sua ideologia.

Pactuando com Halliday (1991, *apud* Berber Sardinha, 2004, p. 30), a linguagem é vista como probabilidade, pois as EI, não necessariamente apareceram com a mesma frequência em todas as categorias. Por isso, o corpus é um importante recurso linguístico para analisar a frequência e estimar a probabilidade do uso das EI nos jornais de notícias da *web*, e a partir daí, gerar variados dados linguísticos.

Efetivamente, observa-se na Espanha, que as EI aportam um forte valor cultural. Cada língua e cada região possui um conjunto de expressões criadas conforme a visão de mundo de seus falantes, sendo possível utilizá-las em vários contextos.

Por fim, esta análise da frequência das EI, aqui demonstrada, almeja contribuir para os estudos de expressões populares, que seguem presentes

na língua espanhola, e que trazem valores culturais e históricos embutidos significativamente, na linguagem.

Referências Bibliográficas

Abella, C. (1996):*!Derecho al toro! El lenguaje taurino y su influencia en lo cotidiano*. Madrid: Anaya & Mario Muchnik.

Alcantud, J.A.G. (1999): Toros y Moros. El discurso de los orígenes como metáfora cultural. In: *Revista de Estudios Taurinos*: 67-90.

Anthony, L.(2012): Advancing AntConc: Design and performance improvements for multi-language. In: Japan Association For English Corpus Studies (JAECS) *Annual Conference, 2012*, Okasa: Osaka Universtity.

Berber Sardinha, T. (2004): *Linguística de Corpus*. Barueri: Manole.

Davies, M. (Org.). *Corpus del Español*. Acesso em: 30/09/2019. Disponível em: <https://www.corpusdelespanol.org/>.

Fuentes, Carlos. (2001): *O espelho enterrado: reflexo sobre a Espanha e o Novo Mundo*. Rio de Janeiro: Rocco.

Lakoff, G; Johnson, M. ([1980] 2002):*Metáforas da vida cotidiana*. (Coordenação da tradução de Mara Sofia Zanotto). Campinas/São Paulo: Mercado de Letras/Educ,.

Luque Duran, J.d.D.; Manjón Pozas, F.J. (1998a): Fraseología, Metáfora y lenguaje taurino. In: Luque Durán, J.d.D.; Pamies Bertrán, A. (eds.), *Léxico y fraseología*, Granada: Método: 40-70.

Pamies Bertrán, A. (2007): El lenguaje de la lechuza. Apuntes para un diccionario intercultural. In: Luque Durán, J.d.D.; Pamies Bertrán, A. (eds.), *Interculturalidad y lenguaje: el significado como corolario cultural*. Granada: Granada Linguistica / Método, vol. 1: 375-404.

Pamies Bertrán, A. (2008): El simbolismo cultural en el lenguaje. Ponencia presentada a la *III Conferencia Internacional de Hispanistas de Rusia*. Moscú,19-21 de mayo.

Pamies Bertrán, A. 2020 "El componente cultural en la variación diatópica: la fraseología taurina española". *Estudios de Lingüística. Universidad de Alicante (ELUA)*. Anexo7: 59-72.

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

Tagnin, S.E.O. (2013):*O jeito que a gente diz*. Barueri: Editora Disal.

Tagnin,S.E.O. (2015): (Org.).Corpora na tradução. In: *A Linguística de Corpus na e para a Tradução*. São Paulo: Hub Editorial: 19-56.

Viana, V. (2010): Linguística de Corpus: conceitos, técnicas & análises. In: Viana, V.; Tagnin, S.E.O. (Org.). *Corpora no Ensino de Línguas Estrangeiras*. São Paulo: Hub Editorial: 25-96.

Xatara, C.M. (1995): O resgate das expressões idiomáticas. In: *ALFA: Revista de Linguística*, v.39: 195-210.

Xatara, C.M. (1998): O campo minado das expressões idiomáticas. In: *ALFA: Revista de Linguística*, v. 42 –Número especial - *O estado da arte nas ciências do léxico: lexicologia, lexicografia e terminologia*. Disponível em: <http://hdl.handle.net/11449/107755>.