DOI: 10.14393/DL52-v16n4a2022-5

As contribuições da Linguística de *Corpus* e do Processamento de Linguagem Natural na elaboração do protótipo do Dicionário Ideológico de Locuções

The contributions of Corpus Linguistics and Natural Language Processing in the elaboration of the Ideological Dictionary of Locutions prototype

Thyago José DA CRUZ[*]

RESUMO: Neste trabalho, buscamos demonstrar como recursos e ferramentas da Linguística de *Corpus* e do Processamento da Linguagem Natural puderam ser empregados na elaboração do protótipo do Dicionário Ideológico de Locuções, de caráter monolíngue e, ao mesmo tempo, onomasiológico e semasiológico. Esse tipo de repertório fraseográfico compõe-se de três grandes seções no corpo do dicionário: a parte sinóptico-analógica, a analógica (correspondendo ambas à parte onomasiológica da obra) e a alfabética (de característica semasiológica). No desenvolver desse projeto, utilizamos como *corpora* o *Corpus* Brasileiro e a *Web*. Como ferramenta para a elaboração do corpo do dicionário, empregamos o software *FieldWorks Language Explore*, o FLEx. Ao final, foi possível verificar que esses instrumentos computacionais foram de fundamental relevância para a realização do propósito da pesquisa.

PALAVRAS-CHAVE: Fraseografia. Linguística Computacional. Linguística de *Corpus*. Dicionário Ideológico.

ABSTRACT: In this work, we seek to demonstrate how resources and tools of *Corpus* Linguistics and Natural Language Processing could be used in the elaboration of the Ideological Dictionary of Locutions prototype, monolingual and, at the same time, onomasiological and semasiological. This type of phraseographic repertoire is composed by three large sections in the dictionary body: the synoptic-analogical part, the analogical (both corresponding to the onomasiological part of the work) and the alphabetical (with a semasiological characteristic). In developing this project, we used the *Corpus* Brasileiro and the Web as *corpora*. As tool for the elaboration of the dictionary body, we used the FieldWorks Language Explore software, FLEx. As a result, it was possible to verify that these computational instruments were of fundamental relevance for the accomplishment of the research purpose.

KEYWORDS: Phraseography. Computational Linguistics. Corpus Linguistics. Natural Language Processing. Ideological Dictionary.

[*] Doutor em Letras pelo Programa de Pós-graduação em Letras (CPTL/UFMS), professor da Faculdade de Educação (FAED/ UFMS). ORCID: https://orcid.org/0000-0001-5562-8485. thyago.cruz@ufms.br

1 Introdução

Há tempos que se reconhece a relevância nas pesquisas lexicográficas do uso de dados informatizados em *corpora* para a elaboração de dicionários. Em meados da década de 60 do século XX, a recolha para a descrição e a análise das unidades do léxico eram realizadas por meio de coleta e registro em fichas – tarefa que demandava tempo e, muitas vezes, gastos elevados. Foi a partir do advento da computação e do acesso mais facilitado a suas máquinas que se possibilitou uma revolução nas pesquisas de caráter lexicográfico (BIDERMAN, 1984, p. 17), pois os dados poderiam ser armazenados, selecionados, analisados, corrigidos, recuperados e com um custo muito menor.

Neste trabalho, fruto de uma pesquisa de doutoramento, objetivamos demonstrar a aplicabilidade dos recursos informáticos nos estudos linguísticos, em especial os advindos da Linguística de *Corpus* (doravante LC) e do Processamento de Linguagem Natural (PLN), para a elaboração do protótipo de um dicionário ideológico de locuções.

Com relação à finalidade do dicionário ideológico de locuções, esse modelo de obra fraseográfica tem como propósito armazenar e descrever as locuções de um modo onomasiológico (na parte sinóptico-analógica[1] e na analógica[2]) e também de um modo

[1] O dicionário ideológico de locuções se divide em três seções principais do corpo do dicionário: a parte sinóptica, a analógica e a alfabética. Os quadros sinóptico-analógicos estão antecedidos pelo plano de classificação de mundo, elaborado por meio da análise semântica do material fraseológico selecionado e se divide em duas macrocategorias com suas ramificações: mundo externo (água, animal, ar, botânica, coisa material, espaço, fogo, quantidade e qualidade, tempo, terra) e ser humano (ação, aspecto, ciclos da vida, ciência, corpo humano, faculdades cognitivas, profissão, relações de influência e/ou posse, relações de preferência, religião e crenças, sentido). Sua elaboração foi facilitada graças aos recursos do software FieldWorks Language Explore, o FLEx, que, após comandos pré-definidos, organizaram os verbetes sob o campo lexical ao qual pertencem, como dissertamos mais adiante neste artigo.

[2] Já a parte analógica, também facilitada sua organização graças aos recursos do FLEx, possui a mesma rede de analogia dos quadros sinóptico-analógico, porém com um visual menos poluído, isto é, os verbetes não estão inseridos em quadros e se dispõem em ordem alfabética, partindo do conceito (ou do lexema que constitui o fraseologismo) para as locuções que lhes são análogas.

semasiológico (na parte alfabética[3]), tendo como público-alvo os potenciais interessados em estudos linguísticos e também tradutores.

Reconhecidas as características do dicionário ideológico de locuções, passamos a um segundo momento em que discorremos sobre o Processamento de Linguagem Natural) e Linguística de *Corpus* para, na sequência, indicar como essas metodologias científicas foram empregadas na nossa pesquisa.

2 Algumas considerações sobre o Processamento de Linguagem Natural e sobre a Linguística de *Corpus*

O uso da tecnologia nos estudos linguísticos vem, a cada década, incrementando-se e aperfeiçoando-se. Conforme Impacta (2019) e Souza (2019), o campo de Linguística Computacional[4] teve início em meados da década de 50, no contexto da Guerra Fria, quando se empregaram computadores a fim de traduzir para o inglês, de forma célere e automática, documentos redigidos em outras línguas. Embora essas máquinas nem se comparem com as atuais, no que diz respeito à qualidade e eficiência, já se era possível traduzir, por exemplo, textos do russo para o inglês de um modo bem satisfatório.

Vieira (2004), por sua vez, coaduna com Impacta (2019) e Souza (2019) sobre a origem da Linguística Computacional, pois, para a pesquisadora:

> a área possui aproximadamente meio século de existência, começou juntamente com a área de Inteligência Artificial (IA), que tem o

[3] Com relação à parte alfabética, também conhecida como índice remissivo nessa tipologia de dicionário, trata-se da apresentação dos verbetes com entradas dispostas em ordem alfabética e que, na sequência, se oferecem a definição e exemplo de uso, além de outras marcas, como função gramatical e indicação direta aos quadros sinóptico-analógicos (obrigatórias) ou contorno, marcas de uso, outras informações (de cunho gramatical, ortográfico, pragmático ou histórico-cultural) e relações semânticas de sinonímia ou antonímia (estas últimas não obrigatórias).

[4] Embora não seja um posicionamento unânime e haja discussão na área, consideramos, neste artigo, a Linguística Computacional e o Processamento de Linguagem Natural como termos sinônimos.

> objetivo de reproduzir comportamento inteligente em sistemas computacionais, como a solução de problemas e automatização do raciocínio (VIEIRA, 2004, p. 1).

Poersch (1987, p. 97) reconhece a Linguística Computacional como "sendo uma ciência interdisciplinar na qual o linguista se serve de, fornece subsídio a, e interage com a ciência da computação". Em outras palavras, esse domínio de estudo considera as máquinas computacionais como uma ferramenta de trabalho que possui a finalidade de processar editar, controlar e/ou analisar dados linguísticos. A Linguística Computacional possibilita o desenvolvimento de *softwares* básicos para pesquisas. Desse modo colabora, mutuamente com a Informática, com o desenvolvimento cada vez mais apurado da Inteligência Artificial.

Domínguez Burgos (2002, p. 104) elenca algumas funcionalidades que programas advindos das pesquisas em Linguística Computacional podem exercer: construir modelos de teorias linguísticas; auxiliar no ensino de língua estrangeira; apontar as correções ortográficas e normativas em textos de um dado idioma; reconhecer a voz humana e processar a mensagem contidas nas frases enunciadas por qualquer indivíduo; elaborar sistemas que facilitem o trabalho do pesquisador, que antes era realizado manualmente, como a construção de verbetes de dicionários; criar jogos virtuais que utilizem, de alguma forma, os comandos da linguagem natural; realizar traduções automáticas ou auxiliar os tradutores nesse processo; e inclusive produzir voz artificial cada vez mais próxima da humana, com transmissão de informação em alto grau de inteligibilidade.

A linha divisória entre esses dois âmbitos de estudos nem sempre está muito clara, haja vista que, no decorrer de várias pesquisas, a LC e o PLN se entrecruzam. Contudo, concordamos com os argumentos de Finatto, Lopes e Ciulla que consideram que o PLN: "denota especificamente o objeto da pesquisa de desenvolvimento de sistemas computacionais capazes de processar objetos de natureza linguística" (2015,

p. 43) e não se trata de um sinônimo de Linguística de *Corpus*, uma vez que esta é mais conhecida pela comunidade científica em geral da área da Linguagem. Tentaremos, brevemente, delimitar estes dois âmbitos de estudos nos parágrafos que se seguem.

A Linguística de *Corpus* busca armazenar amostras de linguagem natural, advindas de uma ou de variadas fontes, tanto na modalidade escrita como na oral. Direciona-se, portanto, a explorar a linguagem por meio de evidências empíricas, efetivadas com recursos da informática. Esse armazenamento denominado de *corpus* linguístico é tratado computacionalmente e se configura como:

> Um conjunto de dados linguísticos (pertencentes 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 a análise (BERBER SARDINHA, 2004, p. 18).

Ainda segundo Berber Sardinha (2004), há alguns pré-requisitos a serem cumpridos para a formação de um *corpus* linguístico, tais como a origem (deve ser de textos de linguagem natural), a autenticidade (escritos ou falados por nativos), o conteúdo (deve passar por critérios pré-estabelecidos pelo seu criador para que aquilo que foi coletado responda às características almejadas na investigação) e a representatividade (ser de uma extensão considerável e representativa).

Exemplos de *corpus* linguísticos podem ser encontrados no: *Corpus* Brasileiro[5], *Corpus* do Português[6]; *Corpus de Referencia del Español Actual* (CREA)[7]; *Corpus Diacrónico del Español* (CORDE)[8] e *Corpus of Contemporary American English (COCA)*[9].

Já o PLN direciona-se aos estudos da linguagem relacionados com a criação e desenvolvimento de *softwares*, aplicativos e sistemas computacionais. Para Othero (2006, p. 343), "cabe à área de PLN justamente a construção de programas capazes de interpretar e/ou gerar informações em linguagem natural".

Segundo Finatto, Lopes e Ciulla (2015), ainda que possa trabalhar com algum *corpus*, a finalidade do PLN não se reduz à descrição de elementos linguísticos, mas se propõe a oferecer soluções a problemas pontuais que se relacionam com o reconhecimento e a reprodução da linguagem natural em uma dada escala, priorizando a relação baixo custo e alto benefício.

Dias-da-Silva (1996) considerava o PLN como um "laboratório em ebulição", pelo fato de que, a cada ano, a indústria tecnológica crescia (e ainda cresce!) vertiginosamente. Com isso, os sistemas vinculados ao PLN se sofisticam em seus mais variados programas informáticos e buscam oferecer produtos tecnológicos de forma acessível e útil. Para o autor:

> a pesquisa [em PLN] reveste-se de um caráter tecnológico e transforma-se em um objeto cobiçado pela voraz indústria da informática que, cada vez mais, precisa tornar seus produtos menos 'enigmáticos' e mais adaptados às necessidades dos seus clientes" (DIAS-DA-SILVA, 1996, p. 66).

[5] Disponível em: https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS Acesso em 23 ago. 21.
[6] Disponível em: https://www.corpusdoportugues.org/ Acesso em 23 ago. 21.
[7] Disponível em: https://corpus.rae.es/creanet.html Acesso em 23 ago. 21.
[8] Disponível em: https://corpus.rae.es/cordenet.html Acesso em 23 ago. 21.
[9] Disponível em: https://www.english-corpora.org/coca/ Acesso em 23 ago. 21.

O pesquisador em PLN, portanto, empenha-se em tornar os sofisticados *softwares* em produtos de manuseio mais simples para o usuário, para que se diminua o risco de que este os abandone; ou prefira utilizar métodos mais exaustivos por não compreender o funcionamento do sistema informático criado ou por considerar que o tempo dispensado para entender seu funcionamento acabe sendo muito grande, equiparando-se ao período utilizado se adotassem os tradicionais procedimentos manuais.

Atualmente, temos como exemplos de programas de PLN: o Extrator Automático de Termos para Ontologias em Língua Portuguesa (ExATOlp), que emprega técnicas modernas de PLN em textos escritos em português, o *Neural Machine Translation* (NMT), um sistema de tradução automática sofisticado capaz de lidar com estruturas complexas da língua de modo satisfatório (como as unidades fraseológicas) e o *FieldWorks Language Explorer* (FLEx), a ser mais bem explorado neste artigo em tópico posterior.

Após essa contextualização sobre a Linguística de *Corpus* e o PLN, direcionamo-nos aos *corpora* empregados na elaboração do protótipo do dicionário ideológico de locuções e sobre o uso e a viabilidade da ferramenta computacional FLEx para a construção da parte sinóptico-analógica, da parte analógica e da parte alfabética do referido modelo de obra fraseográfica.

3 *Corpus* Brasileiro e *Web* como *corpus:* A importância do uso de *corpora* para a elaboração de um dicionário ideológico de locuções

Conforme Penadés Martínez (2015, p. 73), as locuções que compõem um dicionário de locuções não aparecem *ex nihilo* e nem devem partir exclusivamente da competência linguística do fraseógrafo – é aconselhável que sejam extraídas de fontes primárias e de secundárias. Neste contexto, é que surge a importância de *corpora* linguísticos que, ademais de possibilitarem a extração das unidades, permitem a

verificação da frequência de uso, a extração de exemplos, a identificação de significados ou de combinatórias que, muitas vezes, o redator jamais intuiria existir. Para a elaboração de um dicionário, portanto, é necessário que se parta de pelo menos um *corpus*.

Tendo em consideração essas informações sobre o labor lexicográfico aliado ao emprego de *corpus* (*corpora*) na redação de dicionários, escolhemos o *Corpus Brasileiro* e a *Web* como fontes primárias para a extração de algumas unidades fraseológicas para a elaboração do protótipo do dicionário ideológico de locuções, além da coleta de exemplos, de identificação dos significados e das estruturas argumentais, actanciais[10] ou informações pragmáticas (distintas das apresentadas nos dicionários analisados). Cabe-nos ressaltar que a grande maioria das locuções foi retirada de fontes secundárias – nas obras Ferreira (2010), Tesouro do Léxico Patrimonial Galego-Português[11] e o Dicionário de Expressões Idiomáticas[12] – uma vez que começamos a extração e a seleção das locuções por meio delas.

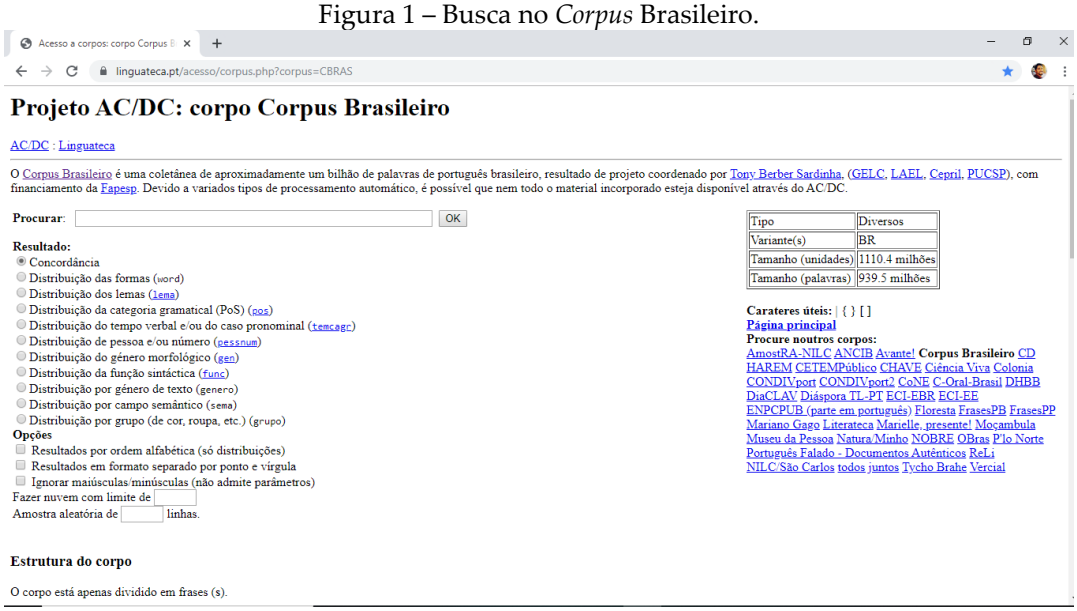
A escolha de *Corpus* Brasileiro deveu-se ao fato de que, atualmente, trata-se de um dos maiores *corpora* construído de palavras do português brasileiro, ao reunir mais de um bilhão delas. Idealizado e elaborado pelo Grupo de Estudo de Linguística de *Corpus* (GELC), coordenado pelo professor e pesquisador Tony Berber Sardinha, a base de dados online possibilita a busca de unidades lexicais (simples e compostas), terminológicas e fraseológicas; possui registros coletados dos mais variados gêneros textuais tanto em sua modalidade escrita como oral, além de apresentar a frequência de ocorrência da unidade pesquisada, no *corpus*. O *Corpus Brasileiro* pode ser acessado por meio do SketchEngine (plataforma paga) ou na Linguateca (acesso gratuito).

---

[10] As estruturas argumentais e actanciais referem-se à valência das locuções verbais e também de algumas nominais, adjetivas e adverbiais (PENÁDES MARTÍNEZ, 2015, p. 188)
[11] Disponível em: http://ilg.usc.es/Tesouro/pt. Acesso em: 20 jan. 2022.
[12] Disponível em: http://www.deipf.ibilce.unesp.br/pt/index.php Acesso em: 20 jan. 2022.

Para se localizar uma unidade fraseológica, deve-se desmembrar e separar por meio de aspas todos os seus elementos constituintes. Por exemplo, para a unidade "a cavalo" digita-se no campo *Procurar*: "a" e, em seguida, "cavalo". Deve estar selecionada a opção concordância para que ele faça a busca exata destas palavras. Vejamos o *layout* dessa página na *Web*:

Figura 1 – Busca no *Corpus* Brasileiro.



Fonte: elaborado pelo autor.

Com relação às locuções verbais, uma vez que as retiradas de fontes secundárias estão registradas em sua forma infinitiva, realizamos várias tentativas de buscas no *Corpus*, para cada unidade, conjugando-se o elemento verbal de sua estrutura em diferentes tempos e modos. Como, a princípio, para a elaboração deste protótipo de dicionário de locuções, preocupamo-nos somente se o fraseologismo em questão ainda está em uso e não registramos se seria uma forma pouco frequente, frequente ou muito frequente, este critério apresentado se tornou viável.

Ao todo eliminamos, até a elaboração da versão final do protótipo de dicionário, 76 locuções, extraídas das fontes secundárias, mas que não tinham nenhuma ocorrência no *Corpus* Brasileiro, nem na *Web*. Em contrapartida, incluímos mais 67

novas locuções, que correspondem a aproximadamente quinze por cento das locuções frequentes nos *corpora* analisados. A inclusão desses novos fraseologismos foi realizada por meio dos seguintes passos:

1- ou partiram de anotações de locuções ouvidas, por nós, em diálogos, em programas radiofônicos ou televisivos, que possuíssem o sema de /+ rural/ [13]e que houvesse ocorrência em um dos dois *corpora*;
2- ou por meio de digitação de partes dos fraseologismos já coletados nas fontes secundárias, em especial, os de maiores extensões, no *Corpus* Brasileiro (como em "botar o carro na frente dos bois", realizaram-se as seguintes buscas: "carro" "na" "frente" "do" "boi" > "frente" "do" "boi" > "do" "boi");
3- ou por meio da busca por unidade lexical que remetesse ao campo lexical do rural (por exemplo, digitávamos no campo de busca a unidade lexical "pato" e verificávamos, uma por uma, qual constituía um fraseologismo). Este método é mais moroso por apresentar, na maioria das vezes, muitas ocorrências de aparição do lexema buscado.

Verificada a importância do *Corpus* Brasileiro e o modo empregado de busca, seleção e extração de unidades, discutimos sobre a consideração e o uso da *Web* como *corpus*.

Para muitos pesquisadores, como Kilgarriff e Grefenstette (2003), a *Web* pode ser considerada como o maior *corpus* linguístico existente na atualidade. Valença e Sabino (2016) fazem uma relação entre a quantidade de páginas indexadas pelo

[13] O critério de selecionar somente as unidades fraseológicas que possuíssem o sema de /+rural/ se deve ao fato de que, por se tratar de uma pesquisa de doutorado e de que seria necessário, portanto, realizarmos recortes, elegemos, como ponto de partida, as locuções que continham um significado que as ligasse a esse conteúdo semântico (como, por exemplo, "aberto dos peitos", que se diz da cavalgadura ou do animal de sela que vive caindo devido a esforços extremos realizados) ou que um dos lexemas que compunha o fraseologismo se vinculasse ao referido sema (como na locução "amolar o boi", de significado "aborrecer", em que há um elemento em sua estrutura que se remete ao mundo rural, isto é, "boi").

buscador *Google* (mais de 60 bilhões) com o possível número de palavras existentes nelas, o que levaria, segundo as autoras, a uma quantidade quase imensurável desses lexemas.

Concordamos com Lüdeling, Evert e Baroni (2007) que assinalam que muitos *corpora* linguísticos tradicionais (armazenados em um banco de dados e sistematizados mediante uma programação) são adequados para determinados propósitos de pesquisa, como ao se desejar verificar quais são as unidades lexicais mais frequentes em um dado gênero textual. Todavia, percebemos, no decorrer desta pesquisa, que muitas unidades que não estavam presentes no *Corpus* Brasileiro, apareciam na *Web*, além de que o número de ocorrência dos fraseologismos naquele era, na grande maioria dos casos, inferior ao demonstrado neste. Sobre essa perspectiva, Lüdeling *et alii*, (2007, p. 7) acrescentam:

> [...] há casos em que os dados necessários para responder ou explorar uma pergunta não podem ser encontrados em um corpus padrão, porque o fenômeno em consideração é raro (dados escassos), pertence a um gênero ou registro não representado no corpus ou decorre de uma época em que os dados do corpus não cobrem (por exemplo, são novos demais) (LÜDELING *et alii*, 2007, p. 7) [14].

Essa afirmação vem ao encontro do assinalado por Valença e Sabino (2016, p. 482), isto é, "a frequência de fraseologismos presentes nesses *corpora* [tradicionais] é relativamente baixa, o que torna esses ambientes de pesquisa menos relevantes que a *web* para pesquisas fraseológicas". Concordamos com as autoras de que se torne menos relevante a busca em *corpora* tradicionais se partirmos somente do prisma da análise

[14] […] there are cases in which the data needed to answer or explore a question cannot be found in a standard corpus because the phenomenon under consideration is rare (sparse data) belongs to a genre or register not represented in the corpus, or stems from a time that the corpus, or stems from a time that the corpus data do not cover (for example, it is too new) (tradução nossa).

de frequência de usos dos elementos investigados. Já no que diz respeito ao controle e sistematização rigorosos do *corpus* pelo linguista, à dificuldade de um motor de busca eficiente, ao risco de encontrar uma linguagem de caráter artificial e à possível presença de erros ortográficos ou desvios gramaticais no material linguístico coletado, trazem, segundo Colson (2007), objeções a esse modo de pesquisa.

Ao levar esses fatores em consideração, tanto as vantagens como as desvantagens de empregar a *Web* como *corpus*, decidimos incluí-la, pela possibilidade de que ela pode trazer à vista do pesquisador unidades fraseológicas, significados novos e exemplos que não foram encontrados em um *corpus* tradicional, mas sempre com olhos atentos também para as dificuldades e objeções elencadas por Colson (2007).

Os passos assumidos para a busca das locuções, seus significados, seus contornos[15] e exemplos, foram:

1. acessamos o site *Google*, digitamos no campo de busca a locução entre aspas (para que fosse exatamente esta expressão). Em seguida, na opção de "Configurações", escolhemos "Pesquisas avançadas". Nela, selecionamos o idioma (Português) e o país (Brasil) – o site fez as buscas somente nesse idioma e espaço virtual;
2. para as locuções verbais, atuamos como fizemos nos passos metodológicos para a busca no *Corpus* Brasileiro, ou seja, conjugamos o lexema verbal que forma parte da estrutura do fraseologismo, em vários tempos e modos (por exemplo: cair do cavalo, caindo do cavalo, caído do cavalo, caio do cavalo, caiu do cavalo, caímos do cavalo, cairá do cavalo etc.);
3. para as locuções de extensões maiores, aos poucos, eliminávamos um de seus elementos constituintes para verificar a possibilidade de alguma variação

[15] Os contornos lexicográficos correspondem aos traços subcategoriais ou contextuais que acompanham a "definição propriamente dita" do verbete.

lexical dentro do fraseologismo. Há também a possibilidade de inserir, apenas uma vez por busca, um asterisco no lugar da palavra excluída – isso faz com que o buscador mostre outros elementos que podem compor as unidades pesquisadas (por exemplo, para a locução "caminho das cabras" é recomendável colocar como "caminho * cabras", pois, se não, há a possibilidade de não se encontrar nenhuma variação do fraseologismo);

4. os erros e desvios gramaticais que apareceram, em especial nos exemplos extraídos, foram corrigidos;

Discutidos os processos de utilização do *Corpus* Brasileiro e da *Web* como *Corpus*, passamos à seção que descreve como o software *FieldWorks Language Explorer* (FLEx) foi empregado na pesquisa, bem como seus benefícios e limitações.

3 Processamento de Linguagem Natural: O FLEx como ferramenta útil para a construção de verbetes

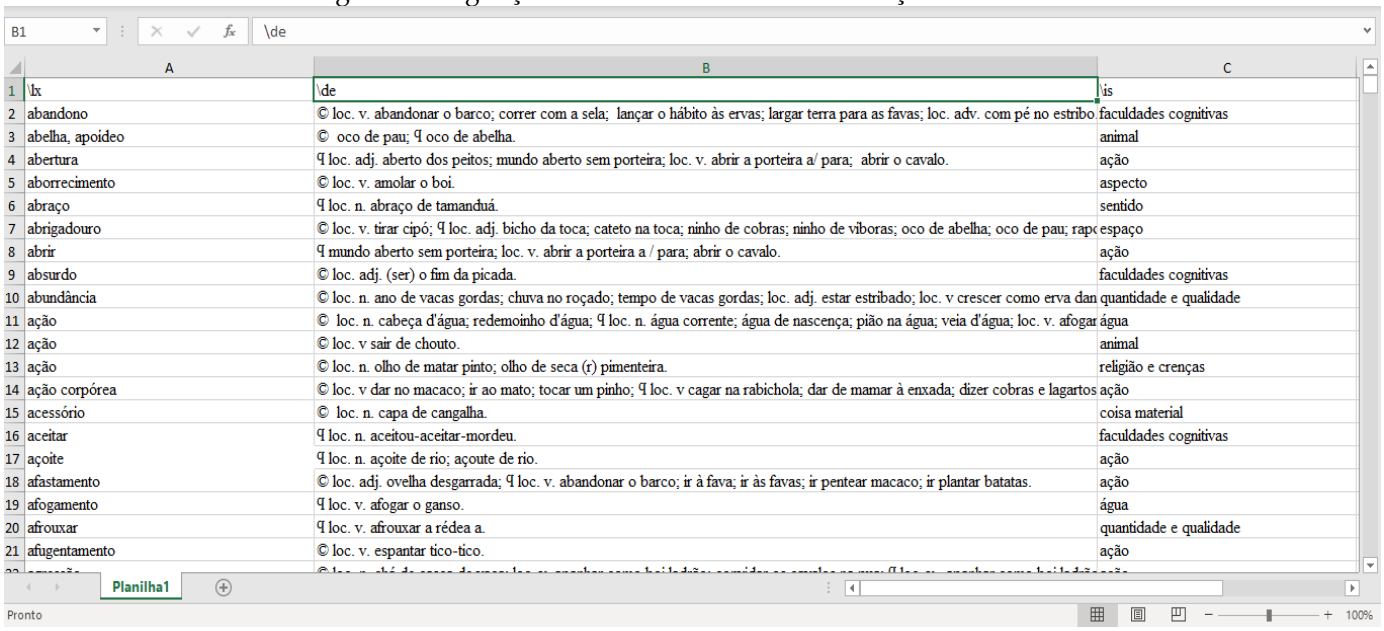
Para a elaboração dos verbetes (dos quadros sinóptico-analógicos, da parte analógica e da parte alfabética) empregamos o *software FieldWorks Language Explorer*, o FLEx. Trata-se de um programa criado e disponível para download[16], gratuitamente, pelo *Summer Institute of Linguistic* (*SIL International*). Essa ferramenta, que também pode ser usada nos estudos e pesquisas da Linguística de *Corpus*, organiza os dados e gera, de um modo automático, os verbetes para o usuário, mediante as configurações pré-estabelecidas.

Para a construção dos verbetes dos quadros sinóptico-analógicos e da parte analógica do dicionário ideológico de locuções, primeiramente, digitamos todos os dados em uma planilha de formato XLS. Haveria a possibilidade de colocá-los diretamente no programa, mas como ainda não tínhamos a segurança se este poderia

[16] Disponível em: https://software.sil.org/fieldworks/ Acesso em: 20 jan. 2022.

ser útil para a elaboração da microestrutura das diferentes seções, preferimos deixar gravado em um arquivo XLS, caso fosse necessário utilizar outro programa ou fazer manualmente em formato DOCS. Para tanto, a digitação de algumas codificações, no formato XLS, foi necessária para que o FLEx pudesse ler os dados e organizar os verbetes, como podemos ver na figura 2:

Figura 2 – Digitação dos comandos e das informações no Excel.



	A	B	C
1	\lx	\de	\is
2	abandono	© loc. v. abandonar o barco; correr com a sela; lançar o hábito às ervas; largar terra para as favas; loc. adv. com pé no estribo	faculdades cognitivas
3	abelha, apoideo	© oco de pau; ¶ oco de abelha.	animal
4	abertura	¶ loc. adj. aberto dos peitos; mundo aberto sem porteira; loc. v. abrir a porteira a/ para; abrir o cavalo.	ação
5	aborrecimento	© loc. v. amolar o boi.	aspecto
6	abraço	¶ loc. n. abraço de tamanduá.	sentido
7	abrigadouro	© loc. v. tirar cipó; ¶ loc. adj. bicho da toca; cateto na toca; ninho de cobras; ninho de viboras; oco de abelha; oco de pau; rapc	espaço
8	abrir	¶ mundo aberto sem porteira; loc. v. abrir a porteira a / para; abrir o cavalo.	ação
9	absurdo	© loc. adj. (ser) o fim da picada.	faculdades cognitivas
10	abundância	© loc. n. ano de vacas gordas; chuva no roçado; tempo de vacas gordas; loc. adj. estar estribado; loc. v crescer como erva dan	quantidade e qualidade
11	ação	© loc. n. cabeça d'água; redemoinho d'água; ¶ loc. n. água corrente; água de nascença; pião na água; veia d'água; loc. v. afogar	água
12	ação	© loc. v sair de chouto.	animal
13	ação	© loc. n. olho de matar pinto; olho de seca (r) pimenteira.	religião e crenças
14	ação corpórea	© loc. v dar no macaco; ir ao mato; tocar um pinho; ¶ loc. v cagar na rabichola; dar de mamar à enxada; dizer cobras e lagartos	ação
15	acessório	© loc. n. capa de cangalha.	coisa material
16	aceitar	¶ loc. n. aceitou-aceitar-mordeu.	faculdades cognitivas
17	açoite	¶ loc. n. açoite de rio; açoute de rio.	ação
18	afastamento	© loc. adj. ovelha desgarrada; ¶ loc. v. abandonar o barco; ir à fava; ir às favas; ir pentear macaco; ir plantar batatas.	ação
19	afogamento	¶ loc. v. afogar o ganso.	água
20	afrouxar	¶ loc. v. afrouxar a rédea a.	quantidade e qualidade
21	afugentamento	© loc. v. espantar tico-tico.	ação

Fonte: elaborado pelo autor.

O código "\lx" indica que os elementos daquela coluna devem ser lidos pelo programa como a entrada do verbete. Já o código "\de" corresponde à definição e "\is", ao domínio semântico (no caso, desta pesquisa, inserimos as macrocategorias).

Devido às características dos verbetes dos quadros e da parte analógica, tomamos alguns procedimentos para que o programa cumprisse com o objetivo que esperávamos dele. Como na definição necessitávamos da diferenciação das classes gramaticais das locuções e elas se encontravam dentro da planilha codificada por "\de", foi necessário que puséssemos manualmente, dentro de cada célula, as

referidas classificações, além também dos símbolos "©"[17] e "ɋ"[18]. Não nos pareceu viável abrir mais colunas com a codificação "\de", o que possibilitaria a inclusão do comando "\ng" (informação gramatical), pois, como decidimos que os verbetes destas duas partes se regeriam primordialmente pelos sentidos ou pela presença do lexema que constituiu o fraseologismo, o FLEx não conseguiria decodificar os dados para dispô-los como esta pesquisa almejava. Logo, elaboramos dois arquivos XLS, um com as locuções acompanhadas das classes gramaticais e outros sem essas informações.

Após todo esse processo, os dois arquivos XLS foram convertidos em formato *Simple File Manager* (SFM), por meio do programa SheetSwiper[19], pois se trata de uma extensão que pode ser lida pelo FLEx. Ao abrir o programa, escolhemos o arquivo que se desejava a conversão e em alguns segundos o novo formato já estava disponível. Salientamos que os dois arquivos SFM, no que diz respeito à inserção dos dados no FLEx, passaram pelo mesmo processo descrito a seguir.

O passo seguinte foi a inserção desses dados no programa FLEx. Para tanto, escolhemos a opção "Create a new Project", selecionamos o arquivo que se deseja abrir, nomeamos o projeto e seguimos alguns pequenos comandos que o programa solicita e o arquivo em seguida é aberto. A tela indicada para a elaboração dos quadros sinóptico-analógicos é a denominada "Dicionário Classificado", que organizará as etiquetas (isto é, as entradas dessa seção do dicionário ideológico) e as indicações lemáticas (isto é, as definições dessa seção do dicionário ideológico) a partir do

---

[17] O símbolo "©" presente na definição dos verbetes indica que, na sequência, apresentamos as locuções que se remetem conceitualmente ao lexema etiquetado na entrada. Por exemplo, na entrada "abundância", temos as unidades fraseológicas "ano das vacas gordas" e "chuva no roçado", dentre outras, que se relacionam conceitualmente ao elemento lematizado.

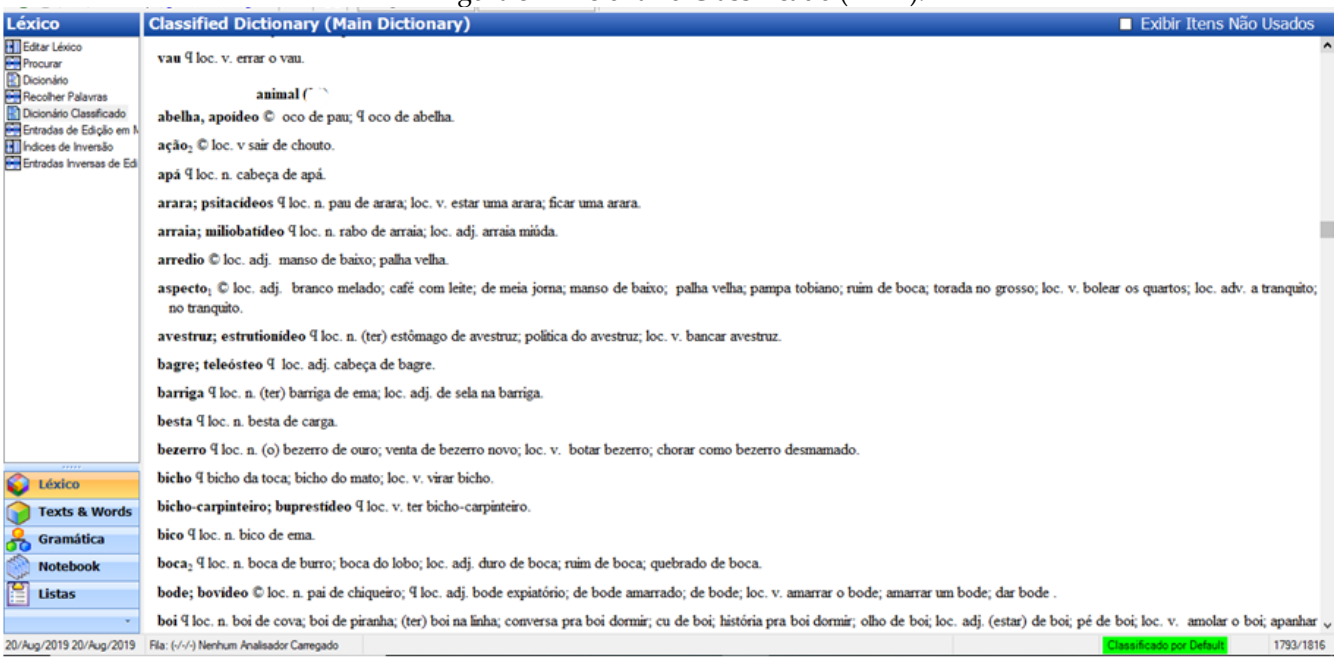
[18] O símbolo "ɋ" presente na definição dos verbetes sinaliza que o consulente encontrará, na sequência, locuções que, em sua estrutura, possuam o lexema (ou um radical em comum) do elemento lematizado. Por exemplo, na entrada "afrouxar", há na definição a locução "**afrouxar** a rédea"; ou em "assar", há "mão de mucura **assada**".

[19] Disponível gratuitamente para download em https://software.sil.org/sheetswiper/ Acesso em: 20 jan. 2022.

domínio semântico. No entanto, previamente, o usuário do programa deve direcionar-se à tela "Recolher Palavras" para configurar os domínios semânticos, pois, nesta seção, caso não se apague o que está identificado como "vernáculo", no momento em que os dados estiverem no "Dicionário Classificado," as macrocategorias aparecerão escritas repetidamente. Caso se julgue melhor, essa duplicação pode ser corrigida quando estiver em posse dos dados em formato DOC.

Após o processo de compilação de palavras, selecionamos a tela "Dicionário Classificado" e já aparecem os verbetes disponibilizados do modo almejado, como vemos na figura a seguir:

Figura 3 – Dicionário Classificado (FLEx).



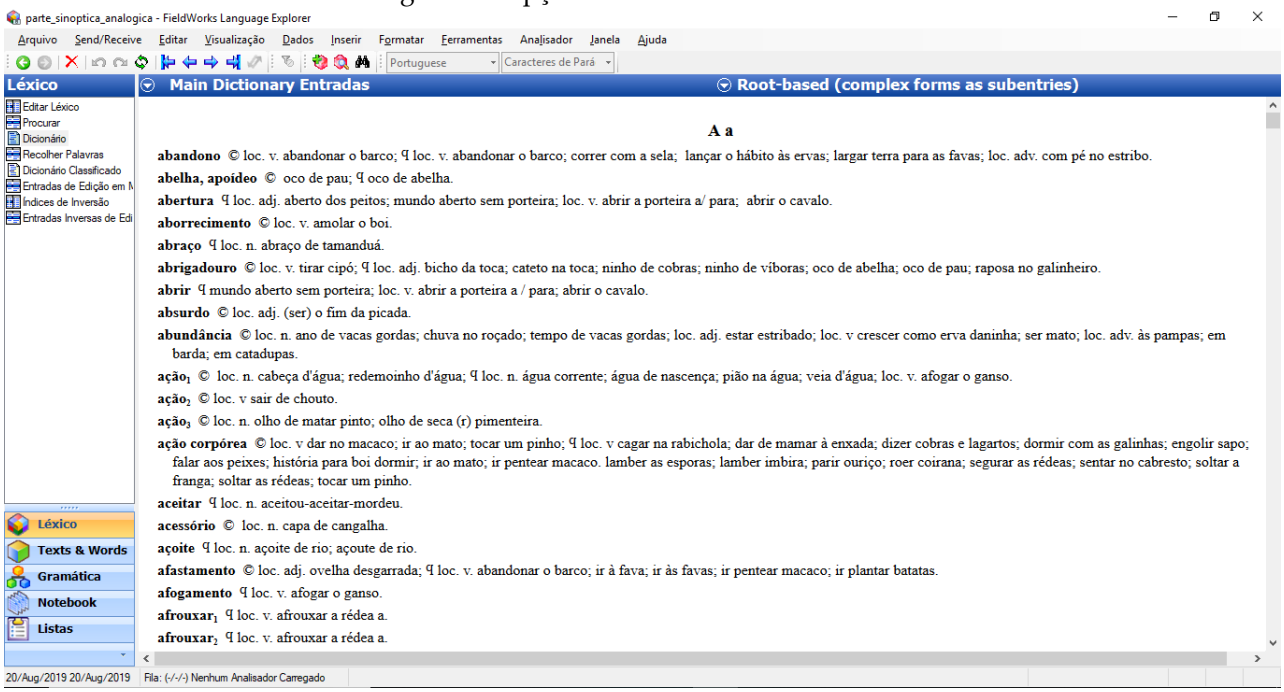
Fonte: elaborado pelo autor.

Para essa seção, o FLEx não possui a disponibilidade de conversão direta dos dados para o formato DOCS, somente para o XHTML. Por isso, foi necessário que os dados fossem convertidos e exportados para este último, posteriormente, aberto no navegador "Chrome", o qual possibilita salvá-lo em formato PDF para, em seguida,

ser convertido em DOCS. Para este trabalho, a conversão de PDF para DOCS foi realizada pelo site PDFCandy[20], uma vez que foi o que melhor realizou essa atividade, com perdas praticamente nulas do formato inicial. Realizado isso, cabe ao fraseógrafo definir a formatação do tamanho da fonte e outras que julgar necessário (como a inserção dos quadros e da numeração de cada verbete).

Para a parte analógica, com o arquivo que possui os verbetes sem as marcas de classificação gramatical, já convertido em formato SFM, seguimos os passos indicados até a abertura do arquivo no FLEx. Após isso, em vez de utilizar a tela "Dicionário Classificado", selecionamos a seção "Dicionário", que apresentará os verbetes dispostos em ordem alfabética sem as indicações das macrocategorias. É um processo semelhante a ser empregado na formação da parte alfabética, como perceberemos mais adiante. A seguir, a figura de alguns verbetes da parte analógica, no dicionário FLEx.

Figura 4 – Opção "Dicionário" no FLEx.



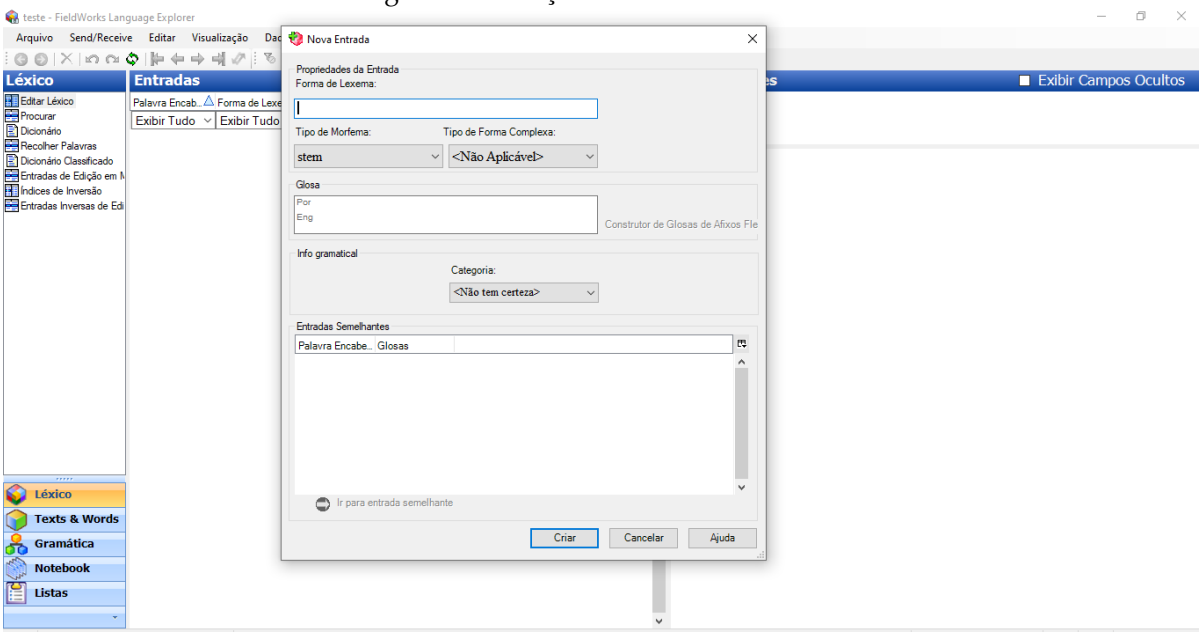
Fonte: elaborado pelo autor.

[20] Disponível em: https://pdfcandy.com/pt/result/ebde06fa.html Acesso em: 20 jan. 2022.

A exportação dos dados dessa tela para o formato DOCS é mais simples. Basta realizar o download, também gratuito, do programa Pathway[21]. Uma vez que esteja instalado no computador, devemos selecionar no FLEx a aba "Arquivo", logo após "Exportar" e escolher a opção "Dictionary, Reversal Index Pathway (various outputs)". Em seguida, elegemos a exportação para o programa Open Office/ Libre Office, que é compatível com o formato DOCS.

Com relação à parte alfabética, também foi realizada por meio do FLEx. No entanto, a modo de testar outras possibilidades, decidimos digitar os dados diretamente no programa. Para tanto, ao abri-lo, selecionamos a opção "Create a new Project". Escrevemos o nome do projeto e escolhemos o idioma. Selecionamos, na aba "Inserir" a opção "Entrada" e surgiu a seguinte janela:

Figura 5 – Inserção de entradas no FLEx.

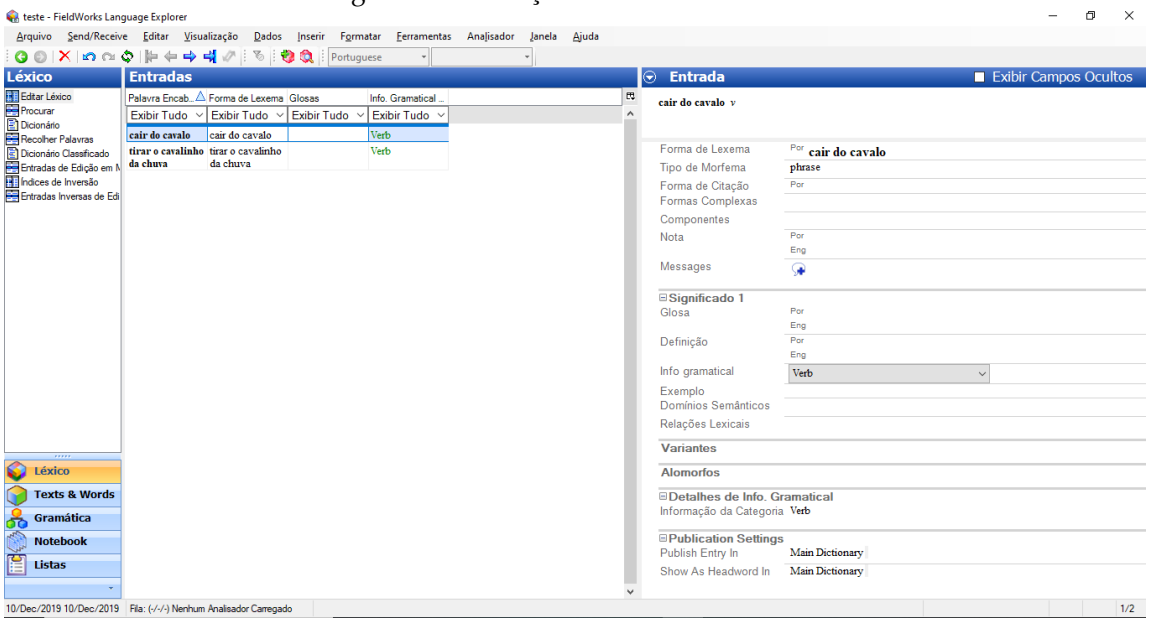


Fonte: elaborado pelo autor.

[21] Disponível em https://software.sil.org/products/ Acesso em: 20 jan. 2022.

No espaço "Forma de Lexema", dispomos a unidade fraseológica a ser lematizada. Automaticamente, a opção "Tipo de morfema" é preenchida como "phrase". Em "Tipo de forma complexa" recomendamos selecionar a opção "Não aplicável" para que no verbete não apareça a marca de elemento sem especificação ("unspecified complex form"). Pela característica assumida para o dicionário ideológico de locuções, não preenchemos a opção de glosa. Identificada a função gramatical que a locução pode assumir, selecionamos a opção que lhe é correspondente (nominal, adjetival, etc.) em "categoria". Caso houvesse mais de uma função, a depender do contexto de onde foram extraídas, decidimos dispô-las em acepções distintas. Por fim, selecionamos a opção "Criar". Realizado isso, o programa expõe a seguinte tela:

Figura 6 – Formação de verbetes no FLEx.



Fonte: elaborado pelo autor.

Com relação às marcas lexicográficas, devemos ir à aba "Configurar" e, em seguida, "Coluna" para adicionar a opção "Dialect Labels (Entry)"[22]. Já no espaço destinado ao significado, é possível preencher a definição e o exemplo, cada um em seu campo específico. Se há a necessidade de inserir mais exemplos ou mais acepções, basta clicar em "Inserir Exemplo" ou "Inserir Significado", respectivamente. No que diz respeito ao contorno lexicográfico, uma vez que decidimos colocá-lo entre parênteses antes das definições, realizamos isso manualmente, isto é, digitando-o diretamente no início de cada definição, no campo da "definição".

Para a inclusão das relações de sinonímia e/ ou antonímia entre as unidades, selecionamos, com o botão direito do *mouse*, a opção "Relações lexicais", e foi escolhida a remissão que almejávamos estabelecer. O *software* a faz, automaticamente, após o usuário determinar quais as unidades devem estar relacionadas.

Como a parte alfabética foi elaborada no FLEx em projetos diferentes das partes sinóptica-analógica e analógica e por não identificarmos um recurso que pudesse fazer a remissão dos verbetes aos quadros sinóptico-analógicos, realizamos essa inclusão de forma manual, após a construção de todos os verbetes.

5 Considerações finais

Discorremos, neste artigo, como a Informática e a Linguística podem estar unidas em outros ramos científicos, como a Linguística de *Corpus* e o Processamento de Linguagem Natural. Com o desenvolver das investigações nesses dois âmbitos,

[22] Mediante as considerações de Cardoso (2010, p. 25), quem atribui a existência de dialetos a um dado espaço geográfico aliado aos traços de formação etnolinguística subjacente a um ato de fala, cabe ressaltar que as marcas de uso "vulgar", presente no modelo de dicionário, não são consideradas como formas de dialeto. Contudo, assim como ocorre no preenchimento das indicações gramaticais e pragmáticas no espaço destinado no FLEx para informações enciclopédicas, realizou-se deste modo pela disposição que o programa colocaria essas estruturas, aproximando o verbete construído à configuração do verbete prototípico elaborado para esta tese.

pesquisadores, cada vez mais, têm a oportunidade de ter às mãos bancos de armazenamento de material linguístico e/ou ferramentas computacionais que facilitem seu trabalho e que possam reduzir a chance de equívocos na pesquisa e de corrigi-los caso haja.

Demonstramos como a LC e o PLN foram de fundamental importância para a elaboração do protótipo do dicionário ideológico de locuções. Apresentamos, ainda, alguns pequenos obstáculos (como o caráter limitado de um *corpus* fechado, ou os riscos de não se trabalhar com uma linguagem natural ou as incorreções gramaticais que podem estar nas unidades léxicas da *Web*, ou as impossibilidades ou dificuldades de comandos no FLEx, dada à peculiaridade da obra fraseográfica delineada). Contudo, esses impasses puderam ser facilmente contornados, o que não desvalida a utilidade e o proveito que essas ferramentas computacionais podem proporcionar ao pesquisador.

Referências

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

BIDERMAN, M. T. C. A ciência da Lexicografia. **Alfa Revista de Linguística**. São Paulo, n. 28 (supl.), p. 1-26, 1984.

CARDOSO, S. A. **Geolinguística**: tradição e modernidade. São Paulo: Parábola Editorial, 2010.

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

DIAS-DA-SILVA, B. C. **A fase tecnológica dos estudos da linguagem:** o processamento automático das línguas naturais. Tese (Doutorado em Linguística e Língua Portuguesa) – Faculdade de Ciências e Letras, UNESP, Araraquara, 1996.

DOMÍNGUEZ BURGOS, A. Lingüística computacional: un esbozo. **Boletín de lingüística**, v. 18, p. 104-119, 2002. Disponível em: http://saber.ucv.ve/ojs/index.php/rev_bl/article/view/1437 Acesso em: 20 jan. 2022.

FERREIRA, A. B. H. **Dicionário Aurélio da Língua Portuguesa**. 5. ed. Rio de Janeiro: Positivo, 2010.

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**. Uberlândia, MG. Vol. 9, n. 5 (dez. 2015), p. 41-59, 2015. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/28670/17075. Acesso em: 20 jan. 2022. DOI https://doi.org/10.14393/DLE-v9n5a2015-3

IMPACTA. **Conheça tudo sobre área de Linguística Computacional!** [S.l.], 2019. Disponível em: https://www.impacta.com.br/blog/conheca-tudo-area-linguistica-computacional/. Acesso em: 20 set. 2021.

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

LÜDELING, A; EVERT, S; BARONI, M. Using web data for linguistic purposes. *In*: HUNDT, M; NESSELHAUF, N; BIEWER, C (org.). **Corpus linguistics and the web**. Amsterdam: Rodopi, 2007.

OTHERO, G.A. Lingüística Computacional: uma breve introdução. **Letras de hoje**, v. 41, n. 2, 2006.

PENÁDEZ MARTÍNEZ, I. **Para un diccionario de locuciones:** de la lingüística teórica a la fraseografía práctica. Alcalá: Universidad de Alcalá, 2015. DOI https://doi.org/10.4067/S0718-93032016000100010

POERSCH, J. M. Linguística computacional: elaboração do diploma para a língua portuguesa. **Letras de Hoje**, v. 22, n. 1, 1987.

SOUZA, J. W. C. **Me vê um texto menor, por favor?** [S. l.], 2019. Disponível em: http://www.roseta.org.br/2019/07/31/me-ve-um-texto-menor-por-favor/ Acesso em: 20 set. 2021.

VALENÇA, E. M; SABINO, M. A. O uso da Web como corpus em pesquisas fraseológicas: Uma prática prejudicial ou um recurso valioso? **Calidoscopio**, v. 14, n. 3, p. 480-488, 2016. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/cld.2016.143.11 Acesso em: 20 jan. 2022. DOI https://doi.org/10.4013/cld.2016.143.11

VIEIRA, R. Lingüística Computacional: uma entrevista com Renata Vieira. **Revista Virtual de Estudos da Linguagem - ReVEL**. Vol. 2, n. 3, agosto de 2004. Disponível em: http://www.revel.inf.br/files/entrevistas/revel_3_entrevista_renata_vieira.pdf Acesso em: 20 jan. 2022.

Artigo recebido em: 27.08.2021 Artigo aprovado em: 21.01.2022