

## Utilização de corpora extraídos da web em um dicionário enciclopédico do novo coronavírus

Using *corpora* taken from the web in an  
encyclopedia dictionary of the new coronavirus

**Ana Flávia Souto de Oliveira**

Universidade Federal de Santa Maria

**Guilherme Hatwig Piper**

Universidade Federal de Santa Maria

**Chrystian Revelles Gatti**

Universidade Federal de Santa Maria

**Resumo:** Neste trabalho, discutimos a compilação e o emprego de *corpora* no dicionário *LEXICOVID-19*. Com o *software* BootCaT, foram confeccionados *corpora* para três estágios do projeto. O AntConc e o Sketch Engine foram utilizados para análises. Inicialmente, o objetivo central da utilização dos *corpora* consistia na extração de exemplos e na complementação da lista de entradas. Após as análises, perceberam-se particularidades com relação às informações que poderiam ser extraídas, o que ampliou o uso dos *corpora*. Nossos resultados demonstram que, mesmo que o *LEXICOVID-19* utilize *corpus* de forma complementar, a utilização de *corpora* enriqueceu as informações apresentadas pelo dicionário.

**Palavras-chave:** Lexicografia; COVID-19; LEXICOVID-19; Corpus; BootCat

**Abstract:** In this paper we discuss the compilation and use of corpora in the LEXICOVID-19 dictionary. With BootCaT software, corpora were built for three stages of the project. AntConc and Sketch Engine were used for analysis. Initially the main purpose of using the corpora was to extract examples and to supply the list of entries. After the analyses, we realized particularities regarding the information that could be extracted, which broadened the use of the corpora. Our results show that, even though LEXICOVID-19 uses corpus in a complementary way, the use of corpora enriched the information presented by the dictionary.

**Keywords:** Lexicography; COVID-19; LEXICOVID-19; Corpus; BootCat

## Introdução

A pesquisa em *corpora*<sup>1</sup> fornece ferramentas para a análise linguística descritiva e quantitativa, facilitando a compilação, visualização e manipulação de grandes quantidades de dados linguísticos. Amplamente utilizada em projetos lexicográficos desde os anos 1980, a pesquisa linguística baseada em *corpus* é considerada um dos principais avanços na lexicografia moderna, pois “qualquer uso que ocorra frequentemente num *corpus*, e seja também encontrado numa variedade de tipos textuais, pode ser tomado de modo confiável como pertencente ao ‘núcleo’ estável da língua”<sup>2</sup> (ATKINS; RUNDELL, 2008, p.48, tradução nossa), e assim ser descrito adequadamente no dicionário.

Na lexicografia, dados provenientes de *corpora* podem subsidiar diversas etapas do processo de compilação de dicionários, como a seleção macroestrutural<sup>3</sup>, a extração de citações e exemplos de uso, a busca por e a análise de padrões colocacionais, bem como orientar a organização de acepções em dicionários que utilizam como critério a frequência. Hoje, é inviável conceber um projeto de dicionário sem recurso a dados provenientes de *corpora*, e o uso da *web* como *corpus* foi amplamente difundido nos últimos anos, pois, “assim como o advento de *corpora* transformou a maneira como os lexicógrafos trabalham, o advento da *web*, e seu rápido crescimento e difusão, mudou o panorama uma vez mais, muitas vezes de maneiras abrangentes”<sup>4</sup> (ATKINS; RUNDELL, 2008, p.53).

Neste trabalho, apresentamos e discutimos o método empregado na extração e utilização de *corpora web* para a elaboração do *LEXICOVID-19: dicionário enciclopédico do novo coronavírus* (OLIVEIRA, 2020). Trata-se de um dicionário *online* que repertoria palavras, termos e expressões relacionados à pandemia de COVID-19 de forma temática, com base nos contextos que dão suporte ao entendimento desses itens lexicais. A arquitetura geral da obra e sua relação com conceitos semântico-cognitivos são apresentadas na próxima seção deste trabalho.

No intuito de mapear a constante inclusão e evolução do léxico suscitado pela pandemia, optou-se, no projeto, pela compilação de *corpora* extraídos de forma automática da *web*, através do software BootCaT<sup>5</sup>. Foram gerados *corpora* para três etapas distintas do projeto: um *corpus* geral da COVID-19, um *corpus* para cada domínio temático e um *corpus* específicos para cenários complexos. Na seção de método, apresentamos as etapas de compilação dos *corpora*.

Após sua compilação, os *corpora* foram analisados com ferramentas do AntConc (versão 3.4.4w) e do Sketch Engine (versão *trial*). Na seção de discussão, apresentamos e problematizamos a utilização dos *corpora* nas diferentes fases de elaboração do dicionário, focando o debate nos

1 O presente trabalho não pretende discutir o estatuto da Linguística de *Corpus* (cf. Shepherd, 2009). Apresentamos a aplicação de *corpora* em diferentes etapas de um projeto lexicográfico, em que trabalhamos com um entendimento genérico de *corpus* como “uma coletânea de textos em linguagem natural, escritos ou falados, geralmente armazenados de forma organizada e informada, além de serem digitalizados a fim de que possam ser lidos por computador” (SHEPHERD, 2009, p.150).

2 [Any usage which occurs frequently in a corpus, and is also found in a variety of text-types, can confidently be regarded as belonging to the stable ‘core’ of the language] (ATKINS; RUNDELL, 2008, p.48).

3 A macroestrutura diz respeito ao conjunto de entradas repertoriadas pelo dicionário e, segundo Hartmann e James (2002, sv. *macrostructure*), é a “estrutura que permite ao compilador e ao usuário localizar a informação em uma obra de referência”.

4 [just as the advent of *corpora* transformed the way lexicographers work – the arrival of the web, and its rapid growth and penetration, changes the landscape once more, often in quite far-reaching ways] (ATKINS; RUNDELL, 2008, p.53).

5 Software que permite a criação de *corpus* a partir de *seeds* (sementes, palavras-chave), que são combinadas em *tuples* (tuplas) de, no mínimo, três itens lexicais, e que servem de base para a extração de páginas *web* em que tais combinatórias ocorrem. O funcionamento do *BootCaT* será apresentado de forma mais detalhada na seção de método de compilação dos *corpora*.

dados provenientes do *corpus* geral da COVID-19, no *corpus* do domínio ECONOMIA<sup>6</sup> e no *corpus* do cenário Vacina<sup>7</sup>.

Inicialmente, foi estabelecido que os *corpora* seriam utilizados somente para a complementação da lista de unidades lexicais arroladas no dicionário e para a seleção de exemplos de uso. Contudo, com o desenvolvimento do projeto e a partir dos dados provenientes das análises, o uso dos *corpora* foi ampliado, por exemplo, para o detalhamento de relações sintático-semânticas que auxiliaram na caracterização dos domínios e cenários apresentados.

Apesar de o projeto utilizar *corpora* de forma complementar às análises, sua utilização facilitou a extração de dados confiáveis, representativos e atualizados. Nas considerações finais, sistematizamos as diversas maneiras que utilizamos os *corpora* ao longo do projeto para identificação e validação de palavras, expressões e domínios temáticos do dicionário; para seleção de exemplos de uso; para o refinamento dos cenários para cada domínio e, finalmente, para detalhar relações sintático-semânticas que deram base para a construção das definições e infográficos desses cenários.

## 1 Domínios, cenários e palavras: o LEXICOVID-19 e o uso de corpora

*COVID-19 in Brazil: “So What?”* (COVID-19 no Brasil: “E daí?”). Esse é o título do editorial da revista *The Lancet* (2020, v.395, p.1461), uma das revistas de medicina mais prestigiadas do mundo. Para compreender inteiramente o subtítulo, o leitor deve ter um conhecimento mínimo sobre uma doença primariamente respiratória que assola o mundo, sobre o maior país da América do Sul (em termos econômicos e por suas dimensões continentais) e sobre o pronunciamento feito por seu presidente no dia 28 de abril de 2020, quando perguntado sobre o que achava do fato de o Brasil ter ultrapassado o número de mortes da China.

O enunciado ilustra um dos principais postulados da Semântica Cognitiva sobre a relação entre cognição e linguagem: as palavras dão acesso a um vasto repertório de conhecimento enciclopédico, de forma que são apenas a ponta do *iceberg* para uma vasta gama de processos conceituais envolvidos na produção e compreensão da linguagem (EVANS; GREEN, 2006).

Segundo Fillmore (1982), os discursos são sustentados por porções de conhecimento que se refletem, de forma geral, nas escolhas linguísticas realizadas pelos falantes e, de forma particular, no léxico. Como exemplo, no contexto da COVID-19, a expressão *isolamento vertical* pode ser entendida como uma “medida de distanciamento social em que apenas alguns grupos de pessoas ficam isolados, principalmente os que apresentam maior risco de desenvolver casos mais graves da doença”. Contudo, a defesa desta medida de prevenção está inserida em um modelo que apoia a continuidade das atividades econômicas, visto que o isolamento apenas de pessoas pertencentes ao grupo de risco permitiria que pessoas mais jovens continuassem a trabalhar e a consumir com relativa normalidade<sup>8</sup>. Fica claro, assim, que a defesa de um isolamento vertical parte de um entendimento da situação que prioriza a mitigação dos impactos econômicos, e não a questão

6 Utilizamos o seguinte padrão ao longo do texto para diferenciar os níveis de análise compreendidos pelo LEXICOVID-19: CAIXA ALTA para os domínios, Fonte Distinta para os cenários e *itálico* para palavras e expressões.

7 Pode ser acessado por diversos domínios, como POLÍTICA, CORONAVÍRUS e SAÚDE.

8 <https://brasil.elpais.com/ciencia/2020-04-11/os-youtubers-de-ciencia-que-o-brasil-escuta-em-tempos-decoronavirus.html>

sanitária imposta pela pandemia.

Com a propagação do vírus Sars-CoV-2 em âmbito global, houve uma rápida disseminação de termos e expressões relacionados à pandemia, principalmente em textos de caráter informativo – como notícias, reportagens e textos de divulgação científica – e de caráter técnico – boletins epidemiológicos e artigos acadêmicos. Percebe-se, assim, que os reflexos sociais, econômicos e políticos da pandemia trouxeram consigo questões linguísticas: falamos sobre *distanciamento social*, *cloroquina*, *teste RT-PCR*, *orçamento de guerra*, *auxílio emergencial*, *variante brasileira*, *vacina de vírus inativado*.

Assim como o entendimento da doença, o léxico relacionado à COVID-19 cresceu exponencialmente nos discursos cotidianos, seus principais reflexos no léxico do português brasileiro associados aos domínios da epidemiologia, saúde, economia e política (cf. Silva e Maia, 2021). Tendo em vista a necessidade de elaboração de uma obra de referência acessível ao público leigo, iniciamos a compilação do *LEXICOVID-19: dicionário enciclopédico do novo coronavírus* (OLIVEIRA, 2020). A organização da obra segue a linha de trabalhos que buscam aproximar a produção de obras de referência a teorias semântico-cognitivas (OLIVEIRA, 2010, 2015; CHISHMAN, 2016, 2018).

O *LEXICOVID-19* é uma obra lexicográfica temática de caráter enciclopédico, que busca integrar a lexicografia eletrônica e preceitos teóricos da Linguística Cognitiva (OLIVEIRA, 2010, 2015; CHISHMAN, 2018). No dicionário, as palavras e expressões estão vinculadas a um cenário (unidade de conhecimento que serve como base para o entendimento da expressão), que pode ser acessado através de um ou mais domínios (que funcionam como grandes áreas temáticas)<sup>9</sup>.

Nem todas as palavras de um mesmo cenário necessariamente se relacionam exclusivamente com cada domínio no qual o cenário está inserido. Por isso, a depender do domínio de onde se parte, o dicionário apresenta unidades lexicais distintas. Por exemplo – o cenário *Vacina* pode ser entendido segundo temas políticos (*negacionismo*, *plano nacional de imunização*, *ministério da saúde*), econômicos (*retomada da economia*, *passaporte de vacina*) ou relacionados à própria área científica (*adenovírus*, *vacina de RNA mensageiro*, *vírus inativado*). Entradas que atravessam todos os domínios, como o *plano nacional de imunização*, podem ser acessadas a partir de qualquer um deles.

No dicionário, existem quatro *layouts* de páginas padrão<sup>10</sup>: na página inicial, o consulente pode escolher por acessar as informações a partir dos domínios (LANGACKER, 2008): NOVO CORONAVÍRUS, SAÚDE, ECONOMIA e POLÍTICA, áreas que concentram a maior parte das discussões suscitadas pela pandemia no Brasil. O consulente pode ainda acessar as informações de forma mais tradicional, através da forma linguística, pela caixa de busca por palavra ou expressão.

Ao acessar um domínio, como o NOVO CORONAVÍRUS, cada página de domínio apresenta como categorias de informação: definição, Cenários do domínio, vídeo de canal do YouTube, Infográfico do domínio e Saiba Mais. Como ilustração, para o domínio NOVO CORONAVÍRUS, são apresentados os seguintes cenários: Difusão e Propagação, Doença, Grupo de Risco, Me-

<sup>9</sup> Por exemplo, as expressões *grupo prioritário*, *insumo farmacêutico ativo – IFA*, *vacinação heteróloga* e *registro de uso emergencial* fazem parte do cenário *Vacina*, que, por sua vez, pode ser acessado através dos domínios NOVO CORONAVÍRUS, SAÚDE e POLÍTICA.

<sup>10</sup> Ou seja, a obra segue um programa constante de informações (PCI); são apresentadas informações de mesmo tipo para verbetes que apresentam a mesma natureza.

didadas de Prevenção, Sintomas, Testes Diagnósticos, Tratamento, Vacina e Vírus.

Os cenários dos domínios são sub-domínios que organizam elementos, portanto, dão sustentação de forma mais direta ao entendimento de palavras e expressões que evocam tais cenários. A Semântica de *Frames* (FILLMORE, 1982) nos auxilia, como teoria, a conceber as definições das unidades lexicais, já que a compreensão delas pode demandar a evocação de um cenário maior que lhes serve de fundo.

Em nosso *corpus* de POLÍTICA, a unidade lexical *repatriação* evoca, para sua definição o contexto maior de um lugar estrangeiro, um local de origem (pátria), o cidadão desta pátria que está no estrangeiro (que será repatriado), a atuação do governo de seu país para ajudá-lo a retornar etc. Todos estes são participantes distintos do mesmo fato, e, a partir do ângulo de cada participante envolvido, a situação pode ressaltar peculiaridades.

O cenário *Medidas de Prevenção*, por exemplo, agrupa, entre outros, itens como barreira sanitária, distanciamento social, higienização das mãos, isolamento domiciliar, lockdown, medidas não farmacológicas, quarentena e uso de álcool gel. Ele pode ser acessado a partir de todos os domínios temáticos do dicionário, visto que as medidas de prevenção contra o novo coronavírus são medidas sanitárias (como higienização das mãos, isolamento domiciliar e uso de álcool gel), que dependem em larga medida de decisões políticas (como barreira sanitária, distanciamento social e quarentena) e que impactam fortemente no funcionamento da economia (implementação precoce da quarentena e lockdown).

Para cada cenário, são apresentadas as seguintes categorias informacionais: definição, palavras do cenário, imagem ou infográfico, vídeo de canal do YouTube e Saiba Mais. Alguns dos cenários, como *Vacina* e *Medidas de Prevenção*, foram considerados complexos, visto que apresentam uma grande quantidade de unidades lexicais vinculadas, dependem de mais de um domínio para sua caracterização e apresentam diferentes atores e ações complexas<sup>11</sup>.

Por fim, as palavras e expressões recebem um verbete que apresenta variantes, definição (no contexto da COVID-19), exemplo de uso e equivalente em língua inglesa. Está ainda prevista a possibilidade de inserção de imagem, em caso de itens para os quais definições ostensivas sejam consideradas facilitadoras para o entendimento.

Ao longo do projeto, percebeu-se que, conforme a realidade se rearranjava, com as contínuas descobertas científicas, novidades tecnológicas e tomadas de decisão por autoridades, o dicionário precisava ser constantemente atualizado<sup>12</sup>. Devido a essa evolução gradual do léxico da COVID-19 e do estabelecimento de novos conceitos lexicalizados, surgiu, no projeto do dicionário, a necessidade de delimitar um método de compilação de *corpora* que abarcasse textos desses domínios temáticos e que permitisse a atualização constante de seus conteúdos.

Exemplos dessa necessidade constante de atualização dos dados repertoriados pelo *LEXICO-*

11 Na seção de análise, abordaremos a forma como essa complexidade pode ser incorporada ao verbete de *Vacina*, através do uso do *corpus* para o detalhamento de relações sintático-semânticas e da criação de um infográfico.

12 Cabe notar, ainda, a problemática de uma definição fixa e sólida dos termos, tendo em vista a instabilidade dos próprios referentes, por exemplo, no caso do entendimento de *Grupo de risco*: numa redação do portal G1 no dia 6 abril de 2020, gestantes ainda não eram consideradas parte do grupo de risco. Ainda no mesmo mês, no dia 15, o Ministério da Saúde declarou que gestantes e puérperas eram mais suscetíveis às infecções e estavam entre as populações mais vulneráveis aos efeitos do vírus. Além disso, a própria ideia de que existe um grupo de risco já é questionada e está sendo entendida, atualmente, num modelo de comportamento de risco. Isto impõe uma necessidade de revisão e renovação constante dos dados apresentados, uma vez que o dicionário deverá estar sempre alinhado com as mais recentes descobertas e especificações científicas.

VID-19 são os cenários *Vacina*, *Variante* e *Irregularidades*, que precisaram ser criados para dar conta do léxico que passou a ser difundido com relação a essas porções de conhecimento, como *vacinação heteróloga*, *variante Delta* e *CPI da COVID*, respectivamente. A palavra *vacina*, por exemplo, que originalmente era apenas uma palavra de um dos cenários do domínio NOVO CORONAVÍRUS (e, no início de nosso projeto, apenas uma expectativa de figurar como cenário), pode hoje constituir um cenário inteiro, devido às tantas implicações em novos contextos. O *Plano Nacional de Imunização* e o processo de *registro* (com a participação dos *institutos de pesquisa* e das *agências reguladoras*), as diversas vacinas aprovadas para uso no Brasil e no mundo, como a *vacina da Janssen (Ad26.COV2.S)*, a *vacina da Moderna (mRNA-1273)*, a *vacina da Oxford/Astrazeneca (Covishield)* e a *vacina da Pfizer (Comirnaty)*.

Iniciamos a compilação de *corpora* no projeto com a intenção de complementar a lista das unidades lexicais e extrair exemplos de uso para palavras e expressões. Contudo, com o desenrolar do trabalho de análise já do primeiro *corpus* (e próprio desenvolvimento do dicionário) percebemos que a utilização dos *corpora* extraídos da *web* poderia ser ampliada. Na próxima seção, detalhamos o método utilizado na compilação dos *corpora* nas diferentes etapas do projeto do dicionário.

## 2 Método de compilação dos *corpora*

Foram compilados *corpora* em distintas etapas do projeto através do *software* BootCaT<sup>13</sup> (1.21). O BootCaT é um *software* que permite a criação de *corpus* randômico a partir da combinação de pelo menos cinco palavras-chave [*seeds*] fornecidas pelo usuário, que são combinadas em *tuples* (tuplas) de, no mínimo, três itens lexicais, e que servem de base para a extração de páginas *web* em que tais combinatórias ocorrem. O *software* então realiza a coleta de *links* com resultados de pesquisa *web* [*queries*] que são salvos e convertidos pelo programa em arquivos de texto (.txt).

O BootCaT permite que o usuário especifique, previamente à compilação do *corpus*, a plataforma de busca (*Google* ou *Bing*), o número de palavras-chave usado nas combinatórias [*tuple length*], a quantidade de tuplas randômicas incluídas na busca e o número de páginas *web* retornadas para cada tupla. Além disso, é possível excluir combinatórias irrelevantes para a busca (aquelas que não são específicas o bastante para gerar buscas relevantes para o *corpus* em construção).

Ao longo do projeto, compilamos *corpora* para diferentes fases: *corpus* geral COVID-19, *corpora* para os domínios SAÚDE, ECONOMIA e POLÍTICA e *corpora* para cenários complexos, aqui ilustrado por *Vacina*. Inicialmente, buscou-se um método padronizado para a compilação dos *corpora*, compreendendo os passos que serão descritos na seção 2.1. Porém, na análise individualizada, percebeu-se que cada domínio temático ou cenário apresentava particularidades, o que exigiu uma reformulação dos passos para a compilação dos *corpora*.

Dessa forma, nas subseções a seguir, detalhamos individualmente o método de compilação dos três *corpora*: *corpora* geral COVID-19, *corpus* do domínio ECONOMIA e *corpus* do cenário *Vacina*.

13 Download disponível em: <<https://bootcat.dipintra.it/?section=download>>.

## 2.1 *Corpus* geral COVID-19

Para a compilação do primeiro *corpus* (realizada em 4 de maio de 2020), partimos da recolha de palavras e expressões relacionadas à COVID-19 e ao novo coronavírus em glossários publicados até então sobre a COVID-19, em pronunciamentos oficiais do ministério da saúde e em textos midiáticos (notícias e reportagens) que tivessem uma ligação direta com os domínios a serem estudados, com prioridade para itens com conotação neutra<sup>14</sup>. Esse primeiro *corpus* tinha como objetivo permitir a análise do léxico ligado à COVID-19 de uma maneira geral, validar os domínios já estabelecidos para o dicionário e a identificar os primeiros cenários desses domínios.

Após seleção qualitativa das principais palavras e expressões elencadas, foram selecionados 18 itens para serem utilizados como sementes<sup>15</sup> no BootCat. Os itens selecionados foram: *achatamento da curva, cloroquina, coronavírus, covid-19, crise, distanciamento social, EPI, epidemia, grupo de risco, hidroxiclороquina, isolamento social, lockdown, OMS, pandemia, profissionais da saúde, quarentena, respirador, surto*.

Para a criação do *corpus* com o *software*, foram selecionadas 30 tuplas de três sementes e realizada a seleção qualitativa das tuplas, com a exclusão daquelas sem relação direta com a COVID-19. Foram solicitadas 30 páginas por combinatória que, após processo de compilação, resultaram em um *corpus* de 768 textos (com 39591 *types* e 1131181 *tokens*) composto, em sua maioria, de reportagens, artigos, documentos e proclamações governamentais.

Foi realizada análise qualitativa da lista de palavras com o AntConc. Como não dispúnhamos de um *corpus* de referência para realizar a extração de *Keyword List*, a busca por termos e expressões foi realizada através das ferramentas de *Word List* e *Clusters/N-grams* para a recolha de unidades lexicais, tendo como critério a frequência de uso e a conexão direta com os domínios do dicionário.

Finalmente, foi utilizado o concordanciador do AntConc para a recolha de exemplos de uso de cada uma das unidades lexicais selecionadas, providenciando dados adicionais para a inclusão destas no dicionário. É importante destacar que, paralelamente à geração do *corpus* randômico através do BootCat e sua análise, adicionamos às listas de palavras-entrada, ao longo de todo o projeto, palavras e expressões retiradas de materiais complementares, como notícias, vídeos, mídias sociais e pronunciamentos políticos.

## 2.2 *Corpus* do domínio ECONOMIA

Após a validação/identificação dos domínios lexicais, foram criados *corpora* específicos para cada domínio, utilizando unidades lexicais representativas do *corpus* genérico como sementes, assim como outras sementes selecionadas de artigos, notícias e boletins on-line representativos de cada domínio. Para criação desses *corpora*, foi dada atenção especial à formulação de tuplas,

<sup>14</sup> A ideia por trás da priorização de uma conotação neutra é a possibilidade de maior abrangência à busca do BootCat, por exemplo, entre palavras *emprego* e *desemprego*, selecionamos *emprego*, sendo que, logicamente, a palavra estaria presente em textos lidando com ambos *emprego* e *desemprego*.

<sup>15</sup> É interessante notar que, caso precisássemos realizar a mesma seleção de palavras e expressões relacionadas à pandemia no atual momento, os itens lexicais elencados seriam outros, como *auxílio emergencial* e *vacina*.

para garantir a representatividade do domínio para qual foram criados.

Conforme descrito na presente seção, o trabalho de compilação do *corpus* da ECONOMIA serviu como teste para a elaboração dos *corpora* dos domínios, principalmente no que diz respeito ao método de recolha de sementes e a combinação das sementes em tuplas. Sendo que o primeiro *corpus* compilado para o domínio não se mostrou tão útil quanto o *corpus* geral da COVID-19, pois, devido ao caráter genérico de algumas palavras-chave, resultou em textos da área anteriores à pandemia.

A compilação do primeiro *corpus* para o domínio ECONOMIA foi realizada por meio da seleção das seguintes palavras-chave, presentes em reportagens referentes aos efeitos econômicos da pandemia de COVID-19, disponíveis em jornais online, assim como itens relacionados à ECONOMIA recolhidos do *corpus* geral: *abertura, assistência econômica, auxílio emergencial, cadastro único, contágio, coronavírus, covid-19, mercado de trabalho, crédito, economia, emprego, empréstimo, farmacêutica, mercado, pandemia, pib, produto interno bruto, projeção, projeções, reabertura, vacina.*

Estas palavras-chave foram, em seguida, organizadas em 23 tuplas randômicas de 3 unidades lexicais, a partir das quais foram extraídos 20 *links* para análise. Os links então foram utilizados pelo software BootCaT para a geração de um *corpus* do domínio ECONOMIA, resultando num *corpus* contendo 535 textos (56.647 types e 1.575.740 tokens), dentre os quais se encontravam declarações governamentais, relatórios, notícias e matérias jornalísticas.

Durante a análise deste primeiro *corpus* da ECONOMIA com o *software* AntConc, percebeu-se que uma grande quantidade de unidades lexicais não apresentava nenhuma conexão óbvia com a pandemia de COVID-19 ou com o domínio em questão. Após uma inspeção mais aprofundada (principalmente durante a recolha de exemplos de uso de unidades lexicais, nas quais o arquivo de texto em si é acessado) percebeu-se que muitos dos textos recolhidos pelo BootCaT eram datados de meses ou até anos anteriores ao surgimento inicial da COVID-19. Após a inspeção das palavras-chave e de suas combinações em tuplas, foi determinado que esta falta de representatividade surgiu devido a combinações que não representavam suficientemente a relação entre o domínio ECONOMIA e a COVID-19, assim como a coleta excessiva de *links* por tupla, abrindo, dessa maneira, a possibilidade para a extração de textos antiquados e não relevantes.

Por estes motivos, foi realizada a compilação de um novo *corpus*, com a inclusão de novas sementes, extraídas de reportagens e artigos disponíveis na web: *abertura, assistência econômica, auxílio emergencial, cadastro único, contágio, coronavírus, covid-19, crédito, demanda, economia, emprego, empréstimo, farmacêutica, indústria, investimentos, mercado de trabalho, PIB, pandemia, Produto Interno Bruto, projeção, reabertura, restrições, vacina, bolsa de valores.*

Para evitar o problema de qualidade dos textos encontrado na formulação do primeiro *corpus*, foi tomada a decisão de formular tuplas de 4 unidades lexicais, com a extração de apenas 10 *links* por tupla, aumentando, assim, a relevância dos textos encontrados; também foi inserida em cada tupla ao menos uma unidade lexical de fizesse referência direta à pandemia de COVID-19. Portanto, pelo menos uma das unidades lexicais *coronavírus, COVID-19, pandemia, vacina, e auxílio emergencial* foi de inclusão obrigatória em todas as tuplas, que continham uma maioria de termos relacionados à economia. Como resultado, o novo *corpus* resultou em um total de 201 textos, con-

tendo 16.077 types e 218.790 tokens e foi mais representativo do campo de análise do dicionário.

Após a extração dos *corpora*, e por não dispormos de um *corpus* de referência para a utilização da ferramenta do AntConc, foram geradas listas de *keywords* no *software* Sketch Engine versão trial<sup>16</sup> e realizada seleção qualitativa de unidades lexicais. Posteriormente, esses itens foram associados a cenários que representassem amplamente os *frames* necessários para compreensão das unidades lexicais.

### 2.3 *Corpus* do cenário Vacina

Como reflexo da pesquisa e do desenvolvimento de vacinas contra a COVID-19, foi necessário adicionar ao dicionário um cenário específico que apresentasse as expressões e trouxesse informações acerca das vacinas. Como o *corpus* genérico da COVID-19 foi compilado em momento anterior ao surgimento das primeiras vacinas<sup>17</sup> e o *corpus* do domínio SAÚDE não apresentava a especificidade necessária à descrição do cenário *Vacina*, decidiu-se por compilar um *corpus* específico para esse cenário.

A partir da identificação de itens lexicais que já compunham a nominata do dicionário e da leitura de notícias, reportagens e textos de divulgação científica, foram recolhidos itens lexicais representativos do cenário *Vacina* no contexto da COVID-19. Com base em uma análise qualitativa dessas expressões, foi realizada uma nova seleção, considerando sua representatividade, pertinência e relação com o desenvolvimento de vacinas específicas contra a COVID-19, excluindo-se aquelas que não se enquadravam nesses critérios. As expressões resultantes foram: *acordo de transferência de tecnologia, aprovação, cobertura vacinal, CoronaVac, COVAX, distribuição de doses, dose, efetividade, eficácia, estudo clínico / ensaio clínico, fase de teste, grupo prioritário/grupos prioritários, imunização, imunizantes, Insumo Farmacêutico Ativo (IFA), Moderna, Oxford, pesquisa e desenvolvimento (P&D), registro, reserva de segunda dose, segurança, Pfizer, plano nacional de imunização / programa nacional de imunização (PNI), Sputnik-V, uso emergencial, vacina, vacinação.*

Essas expressões foram utilizadas como sementes no BootCaT para a compilação do *corpus* do cenário. Selecionamos 40 tuplas, com retorno de 30 páginas por tupla. Após a exclusão das tuplas que não tinham relação direta com a COVID-19, o *corpus* resultou em 1156 textos, com um total de 80362 types e 3823125 tokens.

Novamente, utilizamos o Sketch Engine para extração da lista de *keywords* e de expressões sintagmáticas, que foram analisadas qualitativamente. Além disso, utilizamos a ferramenta de concordanciador para a seleção de exemplos de uso e para o detalhamento das relações sintático-semânticas, conforme apresentado na próxima seção.

16 Na versão trial do Sketch Engine, há um limite para o tamanho do *corpus* que pode ser carregado. No caso da segunda versão do *corpus* da ECONOMIA, foi possível inserir 190 textos, com um total de 257.700 tokens. Contudo, como nosso trabalho com *corpus* não apresenta um caráter exaustivo e quantitativo, julgamos apropriado utilizar essa ferramenta do *software*, visto que, com o maior cuidado com o método utilizado para a compilação do segundo *corpus* da ECONOMIA, ele mostrou-se mais representativo do domínio em questão.

17 Cabe destacar que há menção a vacinas no primeiro *corpus*, mas todas as ocorrências discutem a necessidade de desenvolvimento de vacinas para o controle da pandemia ou relatam pesquisas em andamento sobre as vacinas contra a COVID-19.

### 3 Discussão

#### 3.1 *Corpus* geral COVID-19

O *corpus*, inicialmente desenvolvido para representar o léxico da COVID-19 de forma genérica, foi compilado com o objetivo de ampliar a nominata do dicionário, confirmar os domínios anteriormente propostos, assim como identificar os primeiros cenários de cada um dos domínios. Após o processo de extração do *corpus*, este foi então analisado com o *software* AntConc, principalmente com uso das ferramentas *wordlist*<sup>18</sup> e *clusters/n-grams*. Percebeu-se que, apesar de ser possível uma validação dos domínios anteriormente propostos, a representação de alguns, mais notavelmente POLÍTICA e ECONOMIA, era rasa, ou quase imperceptível, o que demandou uma criação de *corpora* específicos para cada domínio para um estudo apropriado.

Foi possível ampliar de forma consistente a lista de palavras e expressões dos domínios SAÚDE e NOVO CORONAVÍRUS. Uma das ferramentas utilizadas para a análise do *corpus* geral da COVID-19 que se mostrou mais produtiva foi a de *clusters/n-grams*. A partir da busca no *corpus*, foi possível ampliar consideravelmente a lista de palavras candidatas a inserção como entrada. O mesmo recurso foi utilizado amplamente em itens como *distanciamento* e *isolamento*, que ampliaram ainda mais a lista de candidatos a entradas. O busca por expressões com *medidas* e *social*, por exemplo, resultou nas seguintes expressões: (i) *clusters* com *medidas*: *medidas de distanciamento (social) (ampliado)*, *medidas de isolamento (social)*, *medidas de prevenção*, *medidas de controle*, *medidas restritivas/ de restrição*, *medidas de contenção*, *medidas de prevenção e controle*, *medidas de higiene*, *medidas não farmacológicas*, *medidas sanitárias*, *medidas de biossegurança*, *medidas de flexibilização*. (ii) *clusters* com *social* e *isolamento*: *distanciamento social*, *distanciamento social ampliado*, *distanciamento social seletivo*, *distanciamento social intermitente*, *isolamento social*, *isolamento domiciliar*, *isolamento vertical*, *isolamento (social) horizontal*, *isolamento respiratório*, *isolamento total = lockdown*, *isolamento de contato*, *isolamento seletivo*.

Em outros cenários replicamos tal busca, como em Irregularidades, com itens como *operação*, que resultou, entre outros, em *operação Favorito*, *operação Placebo*, *operação Tris in Idem*, *operação Mercadores do Caos*. Tal passo mostrou-se proveitoso para diferentes domínios e cenários, mesmo que com a necessidade de identificar qualitativamente de forma prévia termos possíveis de serem avaliados dessa maneira. Entradas como *operação panaceia*, *operação placebo* e *operação fiel da balança* constataram a reincidência de uma “cena-operação” que envolve participantes recorrentes como: Polícia Federal, autoridades políticas, desvios, licitações etc., com o pano de fundo para a fraude ou ato ilícito consistindo geralmente em algum produto, aparelho ou ocasião diretamente relacionada à pandemia (oxigênio, máscaras, contratação de leitos etc.).

#### 3.2 *Corpus* do domínio Economia

Apesar do fato de que nem todas as unidades lexicais utilizadas no domínio foram retiradas

<sup>18</sup> Listagem de todas as palavras utilizadas dentro do *corpus* assim como o contexto textual na qual se encontram.

dos *corpora*, estes ainda serviram como uma importante ferramenta na validação de unidades lexicais recolhidas de outras fontes (jornais, revistas, editais etc.). Um exemplo de *corpora* se encontra no cenário *Setores Econômicos*, boa parte do qual pôde ser recolhido diretamente do *corpus* inicial, simplesmente pela pesquisa da presença da palavra *setor* com a ferramenta *cluster/n-gram*. Dentro de um grande número de termos que se encaixam na definição utilizada para *Setores Econômicos* no dicionário<sup>19</sup>, foram inseridos *setor privado*, *setor público*, *setor produtivo*, *setor de saúde* e *setor de transporte*.

A partir da análise deste *corpus inicial*, pudemos compilar diversas palavras e expressões que permitiram identificar cenários necessários para apropriada descrição do domínio. Os cenários *Impactos Econômicos*, *Setores Econômicos*, *Escolas de Pensamento*, *Atores e instituições Econômicas*, *Medidas de Prevenção*, *Atividades Econômicas*, *Agentes regulamentadores*, *Políticas Econômicas*, *Índices Econômicos*, *Debates Econômicos* e *Relações Econômicas* serviram como cenários finais para a apropriada classificação das unidades lexicais selecionadas. Porém, devido aos problemas envolvendo a inclusão de informações não relevantes na compilação do *corpus*<sup>20</sup>, foi necessária, para contínua análise do domínio, a compilação de um *corpus* novo com uma seleção de palavras-chave mais cuidadosa, assim como tuplas maiores e uma coleta menor de links por tupla, assegurando um *corpus* mais representativo, com textos atualizados e relevantes.

Outro exemplo do uso do *corpus* para a exploração do domínio ECONOMIA foi sua utilização como uma ferramenta de refinamento do domínio, possibilitando, por exemplo, a verificação da frequência e dos contextos de uso de palavras quase sinônimas ou variantes. Isso auxiliou na tomada de decisão sobre o caráter de sinonímia (se estavam sendo utilizadas com o mesmo significado referencial), sobre qual delas seria escolhida para figurar como entrada principal no dicionário e qual seria considerada uma variante. Um exemplo disso foram as unidades lexicais *lockdown*, *bloqueio total* e *tranca-rua*, as três fazendo referência ao mesmo fenômeno, o completo isolamento social para evitar novos contágios. Porém, como *lockdown* era o item de uso mais frequente no *corpus*, *bloqueio total* e *tranca-rua* foram incluídos como variantes.

Uma peculiaridade notada entre as palavras-chave que foram selecionadas através da análise dos *corpora* da ECONOMIA foi que muitas delas também figuraram no domínio POLÍTICA, um fenômeno de certa forma já esperado, devido à inerente proximidade entre ambos os domínios. Devido a esta intersecção dos dois domínios, foi necessário diferenciar entre unidades lexicais que poderiam receber apenas uma definição, principalmente no que diz respeito aos itens lexicais do cenário *Políticas Econômicas*, e aquelas que necessitavam de definições distintas, ou seja, que foram consideradas homônimas, assim possibilitando a compreensão do impacto dessa unidade dentro de cada esfera de conhecimento. Esse foi o caso para unidades lexicais do cenário *Escolas de Pensamento*, dentre as quais estão *capitalismo* e *comunismo*, que também são encontradas no domínio POLÍTICA, mas com definições diferentes. No âmbito dos Debates

<sup>19</sup> Diferentes setores da economia afetados pela pandemia, como indústria farmacêutica, turismo e o setor de serviços. Em suma, as áreas específicas que sofreram algum impacto direta ou indiretamente durante a pandemia de Covid-19. Em geral os efeitos causados pela pandemia foram adversos, com setores de turismo, transporte, e eventos sofrendo graves quedas na capacidade de geração de renda durante os períodos de quarentena, porém, setores como o *delivery* e serviços on-line em geral passaram por um período marcado por grande crescimento.

<sup>20</sup> Como já explicado na seção de metodologia, a inclusão de informações não relevantes ao domínio provavelmente ocorreu devido a uma combinação de tuplas com palavras-chave não suficientemente específicas à economia e a coleção de um número desnecessário de *links* por tupla.

Políticos, seus usos incluem conotações e nuances pejorativas, como o uso de comunismo = “totalitarismo” ou de capitalismo = “opressão”. Enquanto que, na esfera de Debates Econômicos, as definições de ambos os itens são mais sucintas e apresentam uma versão neutra do significado de cada palavra.

Também foi notado que muitas unidades lexicais retiradas dos corpora perderam sua conotação econômica entre o período de tempo de criação dos corpora e as análises mais recentes, refletindo mudanças de hábitos sociais trazidos pela pandemia e pelo isolamento social. Unidades lexicais como *lives* e *delivery*, que, inicialmente, eram tratados por textos quase que exclusivamente como meios de manutenção do funcionamento de estabelecimentos comerciais e atividades econômicas, passaram a refletir uma modificação no comportamento social em geral. Tal mudança de conotação em unidades lexicais pode ser facilmente atribuída à natureza rapidamente mutante de um vocabulário associado a um evento que ainda está em desenvolvimento, o que levou à elaboração de um novo domínio para a análise, o SOCIAL.

Podemos sugerir que o entrelaçamento e sobreposição entre os domínios ECONOMIA e POLÍTICA se dá devido ao fato que eles são marcados por cenários muitas vezes genéricos ou que são relacionados não apenas à COVID-19, mas também às crises sociais, políticas e econômicas exacerbadas pela pandemia. Como resultado, as discussões acerca de tais crises são discutidas pela mídia e pelo público de uma maneira generalizada, criando, portanto, a necessidade de conhecimentos de mundo que são pertinentes a mais de um domínio para sua compreensão.

### 3.3 Corpus do cenário Vacina

O *corpus* do cenário *Vacina* foi um dos principais exemplos da importância da constante atualização dos *corpora* utilizados no projeto para contemplar o léxico da COVID-19. É interessante destacar que, na análise do *corpus* geral da COVID-19, conseguimos identificar algumas expressões relacionadas à vacina, porém elas eram, em sua maioria, relativas à pesquisa e desenvolvimento de vacinas e aos estudos clínicos que estavam sendo desenvolvidos.

Após a análise do *corpus*, com as ferramentas *clusters/n-grams* e *keywords*, percebeu-se que o cenário *Vacina* mostrou-se complexo: apresentava uma grande quantidade de unidades lexicais (mais de 60); dependia de mais de um domínio para sua caracterização (NOVO CORONAVÍRUS, SAÚDE e POLÍTICA) e apresentava diferentes atores (agências reguladoras, institutos de pesquisa e empresas farmacêuticas) e ações complexas (estudos clínicos, registro, vacinação).

Assim, surgiu a necessidade de sistematizar as informações a serem apresentadas para o cenário. A partir da categorização das unidades lexicais em subcenários, percebeu-se que o *corpus* poderia auxiliar na identificação e na caracterização das principais ações envolvidas nesse cenário. Por exemplo, para o subcenário (processo de) Registro de vacina, buscou-se pelo item ANVISA (que, no Brasil, é a agência responsável pela análise técnica e autorização para uso) com as ferramentas do concordanciador e *clusters/n-grams*. Foram identificadas, manualmente, as principais ações pelas quais a Agência era responsável ao longo do processo de registro e os participantes que mais figuravam nas sentenças referentes a esse processo (ou seja, os elementos

que participam do subcenário *Registro*).

A partir da compilação desses dados, e baseando-se nos contextos de uso presentes no *corpus*, foi possível esquematizar o cenário da seguinte forma: [Agência Reguladora] – *regula (o uso)/ realiza análise técnica de/ aprova (o uso)/ autoriza (o uso emergencial)/ libera (o uso)/ avalia (registros e licenciamento)* [de Vacina] – [Desenvolvida por] – [Instituto de Pesquisa ou empresa farmacêutica].

O mesmo tipo de análise foi realizada para outros sub-cenários de Vacina, como Pesquisa e desenvolvimento, Estudos Clínicos e Plano Nacional de Imunização, com exemplificado a seguir: [Ministério da Saúde] – [Elabora] – [PNI] – [que] – *define diretrizes nacionais de imunização/ estabelece grupos prioritários/ organiza a logística de distribuição de vacinas/ orienta os estados/ elenca ações de vacinação*.

Esses dados foram utilizados tanto na construção da definição do cenário quanto na composição de um infográfico, que traz informações acerca desses subcenários através de relações esquemáticas (sv. *Vacina*, LEXICOVID-19).

## 4 Considerações finais

Pelo fato de os *corpora* terem sido utilizados como uma ferramenta de apoio na elaboração do LEXICOVID-19, eles estão sendo empregados em diferentes etapas de compilação de informações, dependendo da necessidade de análise de cada domínio ou cenário. O *corpus* randômico foi utilizado para extração inicial de uma lista de palavras para compor a nominata e confirmar os domínios previamente estabelecidos para a obra. A confirmação de unidades lexicais no domínio ECONOMIA foi realizada principalmente através da análise qualitativa dos *corpora* do domínio. A identificação de *clusters* (por exemplo, com os itens *licitação, operação, distanciamento, isolamento*) foi amplamente utilizada para expandir a lista de entradas do dicionário, principalmente dos domínios SAÚDE e NOVO CORONAVÍRUS. Os *corpora* serviram ainda para a extração de exemplos de uso de unidades lexicais de todos os domínios. Por fim, os *corpora* tiveram papel importante nas análises de cenários que apresentam maior complexidade, como *Vacina* e *Irregularidades*.

Os resultados demonstraram a pertinência da utilização de um extrator de *corpus web* para o projeto, principalmente pela demanda de constante atualização dos *corpora*, mesmo que com a necessidade de adaptar algumas das etapas do método de recolha, facilmente alteradas pelas opções disponíveis no próprio *software*. Além de questões semânticas referentes à descrição de algumas unidades lexicais, o aparecimento de novos impactos, como a infecção por síndrome pós-covid e a Síndrome Inflamatória Multissistêmica, mostram que estamos diante de uma realidade em si mesma mutável, que demanda uma revisão contínua.

Nossos resultados mostram que o método estabelecido para a utilização do *corpus* no projeto possibilitou a análise de cenários mais específicos dentro dos domínios, como *Vacina*, e que os critérios utilizados na construção dos *corpora* e a análise de dados dependem não apenas do objeto de análise, mas também do tipo de dados que se deseja encontrar. Portanto, mesmo que a utilização de *corpus* tenha uma natureza em grande parte quantitativa, é fundamental destacar

a relevância de critérios qualitativos para a análise dos dados presentes no *corpus* e para sua inserção no dicionário, principalmente com relação à seleção de sementes e à recolha de unidades lexicais para compor a lista de entradas.

Constatou-se uma disparidade na proporção entre a aproveitabilidade da coleta randômica nos contextos cientificamente ligados ao novo coronavírus – como os da área médica –, e os relativamente periféricos, como é o caso dos temas políticos e econômicos, cujos *corpora* trouxeram revistas, blogs, portais de notícias e editoria política. Por outro lado, os endereços vinculados às unidades lexicais do domínio NOVO CORONAVÍRUS, por exemplo, remetiam para boletins epidemiológicos e artigos de divulgação científica.

Esta especificidade no teor dos dados político-econômicos, exigiu, igualmente, adequação na metodologia previamente planejada. Circundante ao vírus em si, o tema político-econômico pôde ser melhor aproveitado a partir de materiais complementares como pronunciamentos políticos, reportagens e diversas mídias. Isso se deve ao caráter pouco técnico dos cenários em questão, os quais são evocados segundo circunstâncias mais amplas, discrepando da objetividade das unidades diretamente ligadas a tópicos de Medicina e Biologia.

Além disso, há a consolidação de novos costumes, trazidos à tona pelas circunstâncias impostas no enfrentamento da pandemia. Termos como *delivery*, *homeoffice*, *lives no YouTube* e *reuniões no Meet*, ainda que já fossem conhecidos e fizessem parte de nossa realidade, não eram tão recorrentes em nosso dia-a-dia a ponto de caracterizarem, como hoje, as práticas sociais e de consumo inseparáveis do período de isolamento social. A recorrência destes hábitos nos levaram a cogitar a idealização de um domínio SOCIAL ou SOCIEDADE, onde pudéssemos agrupá-los junto às consequências exclusivamente sociais da crise, como por exemplo o impacto geral da pandemia na saúde mental da população<sup>21</sup>.

## Referências

ANTHONY, L. **AntConc** [v.3.59]. 2020. Página Inicial. <https://www.laurenceanthony.net/software/antconc/>. Acesso em: 15 de Outubro, 2021.

ATKINS, S. BT; RUNDELL, M. **The Oxford guide to practical lexicography**. Oxford University Press, 2008.

BARONI, M.; BERNARDINI, S. **BootCaT**, 2004. Página Inicial. <https://bootcat.dipintra.it/>. Acesso em: 15 de Outubro, 2021.

CHISHMAN, R. (Org.). **Dicionário Olímpico**, 2016. Página Inicial. <https://www.dicionarioolimpico.>

<sup>21</sup> Ver <https://www.saude.df.gov.br/wp-conteudo/uploads/2018/03/Nota-Informativa-A-Sa%C3%BAde-Mentale-a-Pandemia-de-COVID19-poss%C3%ADveis-impactos-e-dicas-de-gerenciamento-para-a-popula%C3%A7%C3%A3o-geral.pdf>

[com.br/](http://com.br/). Acesso em: 15 de Outubro 2021

CHISHMAN, R. et al. Dicionário Olímpico: a Semântica de Frames encontra a lexicografia eletrônica. **Linguística de Corpus: perspectivas**. Porto Alegre: Instituto de Letras-UFRGS, p. 265-298, 2018.

EVANS, V; GREEN; M. **Cognitive linguistics: an introduction**. Edimburgo: Edinburgh University Press, 2006.

FILLMORE, C. J. Frame Semantics. In: LINGUISTIC SOCIETY OF KOREA (ed.). **Linguistics in the Morning Calm: Selected Papers from SICOL-1981**. Seoul: Hanshin Publishing Company, 1982. p. 111-137.

ARTMANN, R. R. K.; JAMES, G. **Dictionary of lexicography**. London/New York: Routledge, 2002.

LANGACKER, R. W. **Cognitive grammar: a basic introduction**. Oxford/New York: OUP, 2008.

OLIVEIRA, A. F. S. (org.). **LEXICOVID-19: dicionário enciclopédico do novo coronavírus**. 2020. Página Inicial. <https://www.lexicovid19.com.br/>. Acesso em: 15 de Outubro, 2021.

OLIVEIRA, A. F. S. (org.). **A multiplicidade semântica em learners' dictionaries: por uma abordagem semântico-cognitiva para a organização das acepções**. Tese de doutorado. Instituto de Letras – UFRGS, 2015.

OLIVEIRA, A. F. S. (org.). **Subsídios da semântica cognitiva para a disposição das acepções nos Learner's Dictionaries**. Dissertação de mestrado. Instituto de Letras – UFRGS, 2010.

LANCET, The. COVID-19 in Brazil: “So what?”. **Lancet (London, England)**, v. 395, n. 10235, p. 1461, 2020.

SHEPHERD, T. M. G. O estatuto da Linguística de corpus: metodologia ou área da Linguística?. **Matraga-Revista do Programa de Pós-Graduação em Letras da UERJ**, v. 16, n. 24, 2009.

SILVA, F. M.; MAIA, J. S. Neologismos na mídia em meio à pandemia da Covid-19. **Fórum Linguístico**, v. 18, n. 2, p. 6079-6100, 2021.