# Dicionário da COVID-19: compilação do *corpus* e levantamento dos termos

DOI: http://dx.doi.org/10.21165/el.v51i1.3231

**Lucimara Alves Costa**[1]
**Beatriz Curti-Contessoto**[2]
**Ieda Maria Alves**[3]

## Resumo

Neste artigo, apresentamos um estudo preliminar sobre a temática da COVID-19 com base em dois *corpora* de estudo: o *Corpus Oficial*, composto de 993 textos sobre o coronavírus publicados em veículos oficiais, e o *Corpus Jornalístico*, constituído de 460 textos a respeito dessa temática. Neste trabalho, são apresentadas as fases iniciais do Dicionário em elaboração, que se referem à constituição do *corpus* de estudo. São também esboçadas algumas observações preliminares que o *corpus* constituído nos revela a respeito da terminologia da COVID-19. Nesse sentido, discutimos características gerais e iniciais desses textos que estão relacionadas não só a dados estatísticos, mas também às informações semânticas neles construídas. De modo particular, notamos, por meio dos padrões lexicais recorrentes em nossos *corpora*, que a pandemia trouxe transformações em nível vocabular que são o reflexo de novas formas de "viver" e "conviver" que nos foram impostas com o intuito de conter o contágio e a expansão do novo coronavírus.

**Palavras-chave:** COVID-19; terminologia; *corpus* oficial; *corpus* jornalístico.

1 Universidade Federal de Rondônia (UNIR), Porto Velho, Rondônia, Brasil; lucimaralves@unir.br; https://orcid.org/0000-0002-8481-6829

2 Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil; bfcurti@gmail.com; https://orcid.org/0000-0002-5497-5589

3 Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil; iemalves@usp.br; https://orcid.org/0000-0002-1803-3615

## COVID-19 Dictionary: compilation of the *corpus* and terminological identification process


## Abstract

This paper presents a preliminary study on the topic of COVID-19 based on two study *corpora*: the *Official Corpus*, which is composed of 993 texts on the coronavirus published in official vehicles, and the *Journalistic Corpus*, which has 460 texts on this theme. In this work, the initial phases of the Dictionary in preparation are presented and they refer to the constitution of the study *corpus*. Some preliminary observations that the constituted *corpus* reveals regarding the COVID-19 terminology are also outlined. In this sense, this paper discusses the general characteristics of these texts that are related not only to statistical data, but also to the semantic information built into them. In a particular way, it is noticed that, through the recurrent lexical patterns found in these two *corpora*, the pandemic brought changes at the vocabulary level. These linguistic transformations reflect new ways of "living" and "living together" that were imposed on us in order to contain the contagion and the expansion of the new coronavirus.



**Keywords:** COVID-19; terminology; official *corpus*; journalistic *corpus*.


## Introdução

A pandemia da COVID-19 constitui, sem dúvida, um dos maiores impactos ocorridos mundialmente, nas últimas décadas, com reflexos em vários âmbitos da sociedade, em especial na saúde, economia, educação, dentre outros setores.

Buscando amenizar os inúmeros problemas causados pela pandemia, a elaboração de um dicionário da COVID-19 insere-se como um instrumento que, por meio da linguagem e do léxico da língua portuguesa, procura minimizar os efeitos causados pela não compreensão da linguagem médica e, assim, contribuir para a acessibilidade e a difusão do conhecimento científico com a divulgação da terminologia da pandemia a um público não especializado na área médica, em uma plataforma *on-line*. Esse público, de maneira geral, tem dificuldades para compreender essa linguagem especializada e, não raro, manifesta problemas para interpretar as orientações e informações dos documentos oficiais de saúde a respeito da prevenção, da transmissão, do diagnóstico e do tratamento da doença.

Em função dessas dificuldades, foi proposto o projeto **Estudo e divulgação da terminologia da COVID-19**, que, sob a coordenação da Profa. Ieda Maria Alves e com apoio do Instituto de Estudos Avançados (IEA) da Universidade de São Paulo, está sendo desenvolvido nessa universidade, junto à área de Filologia e Língua Portuguesa (DLCV-FFLCH). Integra-se às atividades do Projeto TermNeo (Observatório de neologismos do

português brasileiro contemporâneo), que cumpre a finalidade de estudar aspectos da neologia geral e da neologia técnico-científica do português brasileiro contemporâneo, assim como a de elaborar glossários e dicionários terminológicos em algumas áreas de especialidade.

Desse modo, o estudo proposto visa a detectar, estudar e divulgar a terminologia da pandemia da COVID-19, doença causada pelo novo coronavírus, com o intuito de facilitar a compreensão dessas informações por falantes brasileiros não especializados na área médica. De maneira análoga a todo trabalho de caráter terminológico, este projeto é, necessariamente, interdisciplinar. Além das autoras deste texto, conta com a participação de pesquisadoras da área da Terminologia (Ana Maria Ribeiro de Jesus, Elenice Alves da Costa, Marcia de Souza Luz-Freitas), da Linguística de *Corpus* (Stella Tagnin, Malila Prado, Sandra Navarro), da Documentação (Cibele Araujo Marques dos Santos, Vania Mara Alves Lima) e terá suas definições avaliadas por pesquisadores da área médica que atuam em grupos de pesquisa junto ao IEA.

O estudo apresentado neste artigo está dividido em cinco partes. A primeira parte, chamada **Introdução**, apresenta as características e os objetivos do projeto, assim como a equipe constituída para a sua elaboração. As seções seguintes são dedicadas à descrição da constituição do *corpus* de estudo e da seleção dos termos. A primeira, denominada **Linguística de *Corpus* aliada ao trabalho lexicográfico e terminológico**, explicita como a Linguística de *Corpus* contribui para otimizar a coleta e a utilização de *corpora* nos trabalhos de cunho lexicográfico e terminológico. A segunda seção, dedicada à **Constituição do *corpus* de estudo***,* apresenta as características dos dois *corpora* constituídos para o estudo, assim como a metodologia adotada para a sua organização. Na seção seguinte, denominada **Observações preliminares sobre os dados levantados,** são apresentadas algumas características, ainda iniciais, que estão sendo observadas na constituição da terminologia da COVID-19.

Neste trabalho, então, são apresentadas as fases iniciais da elaboração do Dicionário, que se referem à constituição do *corpus* de estudo, seguindo as orientações da Profa. Stella Tagnin e de suas colaboradoras. São também esboçadas algumas observações iniciais que o *corpus* constituído nos revela a respeito da terminologia da COVID-19.

## Linguística de *Corpus* aliada ao trabalho lexicográfico e terminológico

A Linguística de *Corpus* (LC), *grosso modo*, ocupa-se da coleta e da exploração de *corpora*, entendidos, neste trabalho, como um conjunto de dados linguísticos textuais coletados com a função de solidificarem e orientarem as pesquisas linguísticas. Por essa razão, ela "vem mudando a maneira como se investiga a linguagem, nos seus mais diversos níveis, colocando à disposição do analista quantidades de dados antes inacessíveis" (KADER RICHETER, 2013, p. 13, *apud* BERBER SARDINHA, 2009).

Segundo Berber Sardinha (2000), mesmo antes da invenção do computador já existiam *corpora*. Nesse sentido, já na Antiguidade e na Idade Média, produziam-se *corpora* de citações da Bíblia. Outro argumento que comprova essa afirmação é o fato de que "na Grécia Antiga, Alexandre, o Grande, definiu o *Corpus* Helenístico" (BERBER SARDINHA, 2000, p. 525).

Entretanto, é indiscutível o fato de que a inovação tecnológica e a invenção dos computadores, nos anos 60, ocasionou um *boom* no trabalho com *corpus* e a solidificação da LC que, atualmente, possui uma grande influência e muito contribui para o desenvolvimento de pesquisas em linguagem.

Direcionado ao trabalho lexicográfico e terminográfico, o avanço tecnológico e a criação de grandes *corpora* eletrônicos proporcionaram um enorme aumento na produção e uma considerável melhoria na qualidade das obras lexicográficas e terminológicas. Hwang (2010, p. 43) aponta que o advento da informática propiciou o surgimento de um grande momento da história lexicográfica, a Lexicografia informatizada, que trouxe consigo "uma verdadeira transformação nas condições de trabalho dos lexicógrafos (facilidade, rapidez e novas possibilidades de produção material dos dicionários), mas também uma evolução nas formas de pensamento e nas práticas lexicográficas atuais".

Conforme este autor, as novas possibilidades de coleta, armazenamento e análise de dados disponibilizados pela informática ocasionaram o surgimento de uma Lexicografia que não se voltasse apenas para a produção de dicionários, opondo-se, portanto, à dicionarística, e voltando-se mais para o trabalho de levantamento, descrição, análise e armazenagem de informações sob a forma de bancos de dados lexicográficos, podendo resultar ou não na confecção dos dicionários (QUEMADA, 1987, *apud* LEHMANN, 1995). O mesmo ocorreu com a Terminologia, uma vez que essas novas possibilidades propiciadas pela informática contribuíram para o aumento na quantidade e também na qualidade dos trabalhos e produtos terminológicos e terminográficos (cf. PARADIS; AUGER, 1987).

Nesse sentido, conforme destaca Orenha (2004, p. 3), a Linguística de *Corpus* veio revolucionar a prática léxico-terminográfica, uma vez que:

> Graças à *Linguística de Corpus,* podemos mais facilmente, e de maneira mais rápida e eficiente, levantar e selecionar não apenas palavras, mas também combinações de palavras. Por meio de um *corpus*, podemos analisar essas combinações em seus contextos naturais e, principalmente, com um contexto muito maior à disposição do lexicógrafo/terminógrafo, que antes dependia de uma árdua busca manual.

Ainda de acordo com Orenha (2004), outra vantagem propiciada pela LC às pesquisas lexicográficas e terminológicas diz respeito à possibilidade de serem realizadas análises estatísticas permitidas pelo uso de ferramentas de pesquisas específicas, como as disponibilizadas nos programas computacionais *WordSmith Tools* (SCOTT, 2015) e *AntConc* (ANTHONY, 2012), sendo este último utilizado nesta pesquisa.

A Terminologia se serve da Linguística Computacional, da Engenharia Linguística e, por extensão, da LC para constituir seu próprio objeto de trabalho (*corpora* de documentos/textos especializados) e utiliza suas ferramentas e aplicações para gerir e organizar melhor o trabalho e o processo terminográfico.

Como se destaca em IULA (2015), convém ressaltar que, mesmo sendo inúmeros os aportes e os recursos propiciados pela Linguística Computacional e pela LC para as pesquisas terminológicas e lexicográficas, a maioria dos sistemas e das ferramentas que funcionam em Terminologia e Linguística, todavia, é semiautomática e requer a intervenção e a contribuição do usuário/pesquisador. Essa união homem-máquina é o que garante a eficácia dos resultados das pesquisas linguísticas que utilizam, como aporte, a LC[4].

Dentre as distintas "facilidades" que podem ser propiciadas por essas ferramentas computacionais e que otimizam o tempo e trabalho do terminólogo e das pesquisas terminológicas/terminográficas, podemos citar:

> - a seleção automática do *corpus* e do conjunto de informações relacionadas a um determinado tema;
>
> - a seleção e extração de termos de uma área de trabalho a partir de um *corpus*;
>
> - a análise morfológica dos termos, com o objetivo de oferecer propostas neológicas tanto de caráter formal quanto semântico;
>
> - a atribuição de termos a um ou mais campos de especialidade, de acordo com sua frequência de ocorrência nos textos;
>
> - a elaboração automática de definições a partir da análise do conteúdo do texto de onde procede o termo;
>
> - a elaboração automática da estrutura conceitual de um campo de especialidade. (IULA, 2016, p. 5, tradução nossa).

4 Sabemos que existem diferentes concepções em torno do que se entende por LC em termos de reconhecimento teórico e/ou metodológico. Contudo, não entramos no mérito dessa questão por não ser esse o nosso foco. Para fins de desenvolvimento deste trabalho, entendemos que a LC nos auxilia a estudar a linguagem por meio de *corpora* (AIJMER; ALTENBERG, 1991), e, em consonância com Matuda e Tagnin (2014, p. 221), "acreditamos que a LC pode ser utilizada para outros fins, que não se restrinjam a pesquisas linguísticas".

Para a realização do nosso trabalho, parte do qual discorremos neste artigo, dando um enfoque à compilação de um *corpus* de estudo direcionado à temática da COVID-19, bem como sobre a seleção e a extração das unidades mais frequentes e significativas desse *corpus*, utilizamos, primeiramente, o recurso computacional *BootCat* (para a compilação dos textos) e, posteriormente, o programa computacional *AntConc* (para a seleção e a extração dos candidatos a termo), conforme explicitamos na próxima seção.

Sendo assim, orientamo-nos por uma abordagem direcionada pelo *corpus* (*corpus-driven*, cf. Tognini-Bonelli, 2001), já que é o próprio *corpus* que nos indica os caminhos de análise linguística possíveis de serem percorridos (MATUDA; TAGNIN, 2014). Para seguir essa abordagem, observamos os dados que, por sua vez, levaram-nos a levantar hipóteses sobre as características da linguagem veiculada pelas fontes consultadas que tratam da COVID-19. Nesse processo, os programas computacionais aos quais recorremos contribuíram, principalmente, com o levantamento dos candidatos a termo, bem como com um vislumbre analítico dos dados por eles organizados tanto de um ponto de vista quantitativo quanto qualitativo. Esses candidatos devem ser verificados com base nos pressupostos da Terminologia. Na seção seguinte, explicamos a metodologia empregada com o intuito de constituirmos nosso *corpus* de estudo.

## Constituição do *corpus* de estudo

Nosso *corpus* de estudo se constitui de dois *corpora*, a saber: o *Corpus Oficial* (CO), que é composto de 993 textos sobre a COVID-19 publicados em veículos oficiais e disponibilizados nos *sites* da Organização Mundial da Saúde (OMS), do Ministério da Saúde do Brasil, das Secretarias de Saúde dos estados brasileiros, da Organização Pan-Americana da Saúde (OPAS), da Fundação Oswaldo Cruz (Fiocruz) e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); e o *Corpus Jornalístico* (CJ), de caráter complementar, constituído de 460 textos jornalísticos compilados a partir dos *sites* da *Folha de S. Paulo*, *O Estado de S. Paulo* e *O Globo*[5].

A metodologia que adotamos a fim de constituirmos o *corpus* de estudo foi baseada na *web* como *corpus* (cf. KILGARRIFF; RIGAU, 2013). Assim, entendemos, apoiadas em Renouf, Kehoe e Banerjee (2007), que esse tipo de recurso tem um forte potencial no sentido de nos permitir observar a dinamicidade das línguas, identificando os padrões lexicais e, mais especificamente, os termos referentes à temática da COVID-19 que vêm sendo veiculados em *sites* brasileiros, verificando se há diferenças com relação à escolha terminológica a depender do veículo que divulga informações a respeito da pandemia.

5 A escolha dos jornais se pautou em dados estatísticos do Instituto Verificador de Comunicação (IVC) que representam a circulação de jornais no Brasil. *Folha de S. Paulo*, *O Globo* e *O Estado de S. Paulo* são os três primeiros colocados no *ranking* dos cinco jornais de maior circulação no país em 2018 e 2019 (MEIOEMENSAGEM, 2020).

Recorremos, então, a essa metodologia por meio da utilização das ferramentas *BootCat Bootstrap Corpora and Terms from the web*, versão 1.21 (ZANCHETTA; BARONI, 2011) e *AntConc* (ANTHONY, 2012). Primeiramente, utilizamos o *BootCat* com o intuito de encontrar os textos disponibilizados na *web* que abordassem a temática em pauta e que foram publicados até março de 2021. Essa etapa foi realizada duas vezes, de modo que os *corpora* supracitados fossem constituídos separadamente, tal como explicamos na sequência.

A compilação do CO se deu a partir de uma lista de sementes (*seeds*) que contém unidades lexicais frequentemente associadas ao tema da pandemia de COVID-19, tais como, *coronavírus*, *covid*, *pandemia*, *cepa* e *variante*, dentre outras. Além disso, restringimos a varredura que o programa faz no *Google* de modo que essas sementes fossem encontradas apenas nos *sites* das organizações, das fundações e das secretarias mencionadas anteriormente. Como retorno, obtivemos 249 *Urls*, das quais foram baixados os documentos em *txt* que correspondem aos *subcorpora* do CO. Em seguida, checamos cada arquivo a fim de eliminar possíveis textos repetidos ou não relativos ao tema em pauta, por exemplo, e de etiquetar os documentos selecionados para facilitar sua identificação. Desse modo, os arquivos foram identificados com o código <Fonte oficial, Dia-Mês-Ano, Sobrenome, Abreviação do nome do(a) autor(a) do texto, se for o caso>, tal como em <Fiocruz, 09-03-20, Teixeira, S.>, por exemplo. Ao final desse processo, esse *corpus* passou a conter 4.203.126 milhões de palavras.

O mesmo passo a passo foi realizado com o intuito de compilarmos o CJ. A partir da mesma lista de *seeds*, buscamos os textos divulgados na *web* pelas fontes já citadas, trocando apenas os endereços dos *sites* oficiais pelos dos jornais. Após a busca feita pelo programa, chegamos a 122 *Urls* a partir das quais baixamos arquivos em *txt* que constituem os *subcorpora* do CJ. Também realizamos um processo de limpeza dos arquivos baixados pelo programa e de etiquetagem, atribuindo-lhes, assim, o código <Fonte do jornal, Dia-Mês-Ano>, tal como em <ESP, 10-03-21>. Depois de fazermos esse procedimento, chegamos a um *corpus* com 340.583 mil palavras.

Além do cuidado com a seleção dos *sites* e dos documentos específicos para a constituição dos *corpora*, preocupamo-nos com o critério da delimitação e da representatividade ou da relevância do mesmo. De acordo com Cabré (2007), para se chegar à delimitação exata sobre a constituição de um *corpus* e, acima de tudo, sobre sua dimensão, podemos nos nortear com perguntas como: para que se constitui o *corpus* que vamos elaborar? Que finalidade esse *corpus* deve cumprir? A que estudos linguísticos ou contribuições queremos que dê lugar?

Para autores como Leech (1991, p. 27, *apud* BERBER SARDINHA, 2000), por exemplo, o critério da representatividade está relacionado ao tamanho ou à extensão do *corpus*, ou seja, quanto maior for sua extensão, mais representativo será. Entretanto, consideramos

que, como ressalta Berber Sardinha (2004, p. 45), o *corpus* pode ser entendido como a "amostra de uma população cuja dimensão se desconhece" e, sendo assim, é impossível determinar ou quantificar a língua como um todo. Logo, não se pode estabelecer com precisão qual seria a extensão de uma amostra representativa.

Seguindo, então, o que propõe Berber Sardinha (2000), partimos do princípio de que não existem critérios objetivos para determinar a representatividade de um *corpus*. Mesmo porque, antes de se determinar se um *corpus* é representativo ou não, há que se considerar duas questões fundamentais: do que e para quem esse *corpus* é representativo e relevante?

Podemos concluir, portanto, que, para a função que se propõe, que é criar um dicionário da terminologia da COVID-19 voltado para não especialistas, nosso *corpus* é uma amostra válida, representativa e relevante dessa temática, tanto pelos documentos selecionados, quanto pela quantidade total dos arquivos que compõem os *corpora* inicialmente compilados. Tais *corpora* estão sendo complementados, seguindo a mesma metodologia, uma vez que a terminologia estudada tem se mostrado bastante dinâmica.

Depois de termos realizado a etapa de constituição de nossos *corpora* de estudo, passamos ao programa *AntConc* com o intuito de darmos um tratamento aos textos compilados, analisando as ocorrências dos itens lexicais presentes nos *corpora* e identificando os termos mais pertinentes para o nosso estudo. Essa análise foi realizada separadamente para que os dados dos *corpora* não se misturassem. Assim, poderíamos verificar que tipos de ocorrência, em termos qualitativos, são encontrados no CO e no CJ e compará-los. Para realizarmos essa etapa de nossa investigação, foi necessário utilizarmos um *corpus* de referência que foi disponibilizado pela Profa. Stella Tagnin, que também integra a equipe do presente projeto. Esse *corpus*, chamado de *Generalidades*, traz um compilado de textos em Português Brasileiro sobre temas diversos, com um total de 2.125.210 milhões de palavras[6].

Partimos, então, para a observação e levantamento das unidades lexicais presentes nos *corpora*, organizadas em diferentes tipos de listas, dentre as quais estão as de palavras-chave (*keyword list*), de palavras (*wordlist*), de *clusters* e de concordâncias (*concordance*). A *keyword list* permite criar uma lista de palavras-chave ao se comparar a lista de palavras mais frequentes do *corpus* com a frequência de um *corpus* de referência. Já a *word list* gera uma lista de todas as palavras que constam no *corpus*, podendo ser organizada alfabeticamente ou por ordem de frequência. O recurso *clusters*, por sua vez, compila

6 Na literatura, existe a indicação de que o *corpus* de referência, preferencialmente, seja de três a cinco vezes maior do que o *corpus* de estudo (cf. BERBER SARDINHA, 2004). Contudo, no nosso caso, cumpre dizer que o *corpus Generalidades* foi suficiente para nos auxiliar no levantamento das *keywords*, ainda que não atenda a esse critério.

uma lista ordenada das palavras próximas ao termo pesquisado que são determinadas a partir da quantidade de itens combinados que desejamos buscar. Assim, se a unidade lexical *coronavírus* for buscada com *clusters* com 2 a 4 elementos, por exemplo, aparecerá uma lista com combinações contendo duas, três e quatro palavras que ocorrem com esse termo no *corpus*. Por fim, a ferramenta *concordance* gera as concordâncias ou contextos de ocorrências de determinado termo, a ser pesquisado dentro dos textos de nosso *corpus*.

Munidas, então, desses recursos, utilizamo-nos da seguinte forma: primeiramente, observamos a frequência das unidades lexicais encontradas no *corpus* com o auxílio da ferramenta *wordlist* com o intuito de verificar quais unidades têm maior ocorrência em cada um de nossos *corpora*; em seguida, contrapomos o CO e o CJ ao *corpus* de referência *Generalidades* com o intuito de selecionar as palavras-chave de cada um dos *corpora* com o auxílio da ferramenta *keywords*; feita essa seleção, que contém 525 unidades extraídas do CO e 461 unidades do CJ, comparamos as duas listas a fim de obtermos uma lista mesclada final e, a partir dela, buscamos os *clusters* de cada uma delas, utilizando a extensão de 2 a 4 elementos, a fim de se verificar suas co-ocorrências e encontrar, assim, candidatos a termos sintagmáticos; por fim, buscamos as concordâncias de cada uma das unidades selecionadas com o intuito de verificar seus contextos de uso, que nos dão uma ideia de sua pertinência ao tema da COVID-19 e que, em etapas posteriores, servirão de base para elaborarmos os verbetes de nosso Dicionário.

Até o momento, finalizamos o levantamento dos *clusters* do CO e do CJ que consideramos como candidatos a termos sintagmáticos. Estamos em fase de limpeza e organização desses dados no sentido de selecionarmos as unidades sintagmáticas mais significativas para a temática em pauta, de encontrarmos possíveis variantes terminológicas e de verificarmos quais candidatos são, de fato, termos da área à luz dos pressupostos da Terminologia. Para tanto, adotamos o critério da relevância semântica, que considera a importância (ou não) desse termo para a temática da COVID-19, independentemente da frequência atingida pelo termo em nossos *corpora*. Somados a ele, estão os critérios apresentados por Barros (2007), utilizados, em Terminologia, para se verificar o grau de lexicalização dos sintagmas terminológicos e para determinar os limites das unidades terminológicas sintagmáticas. Assim, observamos, principalmente: se esses candidatos a termos denominam conceitos específicos do domínio em pauta; se há dependência semântica entre os elementos de um possível termo sintagmático; se essas unidades se encontram definidas nos textos do *corpus* e se são bastante frequentes, dentre outros. Na sequência, elaboraremos os verbetes do dicionário proposto. Nesse processo, também serão consultados especialistas da área médica, que avaliarão se as definições e as notas explicativas do Dicionário estão adequadamente redigidas, considerando que todo trabalho terminológico, por ser interdisciplinar, necessita de especialistas da área enfocada, indispensáveis para o adequado acompanhamento do trabalho do terminólogo.

Na próxima seção, apresentamos uma breve análise, em termos quantitativos e qualitativos, sobre os dados levantados até o momento em nossa investigação.

## Observações preliminares sobre os dados levantados

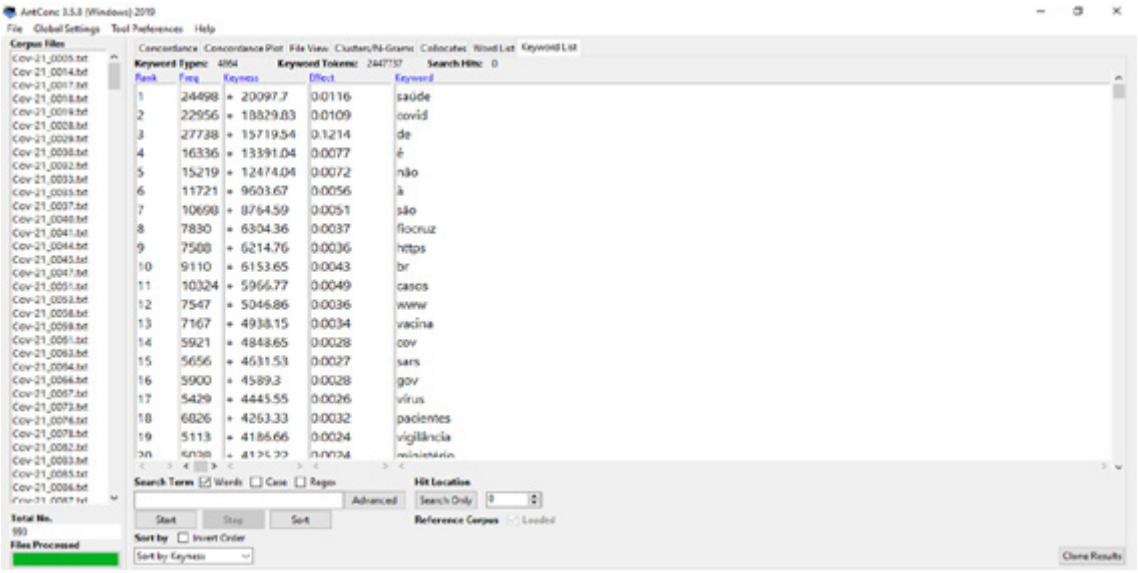
Graças aos recursos tecnológicos de tratamento textual aos quais temos acesso hoje em dia, é possível não só trabalharmos com um grande volume de textos de uma forma mais rápida e eficaz, como também levantarmos alguns dados estatísticos sobre eles. Além dessas observações quantitativas, ao adicionarmos a esse trabalho o olhar humano do pesquisador, diversos aspectos qualitativos podem ser analisados. É sobre esses dois aspectos que tratamos nesta seção.

O primeiro dado levantado com o *AntConc* diz respeito, como dissemos, à lista de palavras (*wordlist*) mais recorrentes em nossos *corpora*. Neles, observamos que as palavras gramaticais representam os itens mais recorrentes. Outras unidades como *saúde* e *covid* figuram como as primeiras palavras substantivais com maior frequência em ambos os *corpora*, o que pode ser justificado com base no recorte temático feito quando de sua compilação. Essas unidades ocupam, respectivamente, as posições 15 e 18 no CO, com 24.498 e 22.956 ocorrências, e 24 e 22 no CJ, com 1.362 e 1.693 atestações.

Além disso, do ponto de vista quantitativo, temos, no CO, 89.336 *word types* e 4.203.126 *word tokens*, e, no CJ, há 19.999 *word types* e 340.583 *word tokens*. Esses dados significam que há cerca de 90 mil tipos de palavras em um total de quase 4 milhões no CO enquanto, no CJ, essa quantidade cai para quase 20 mil tipos em meio a aproximadamente 340 mil palavras.

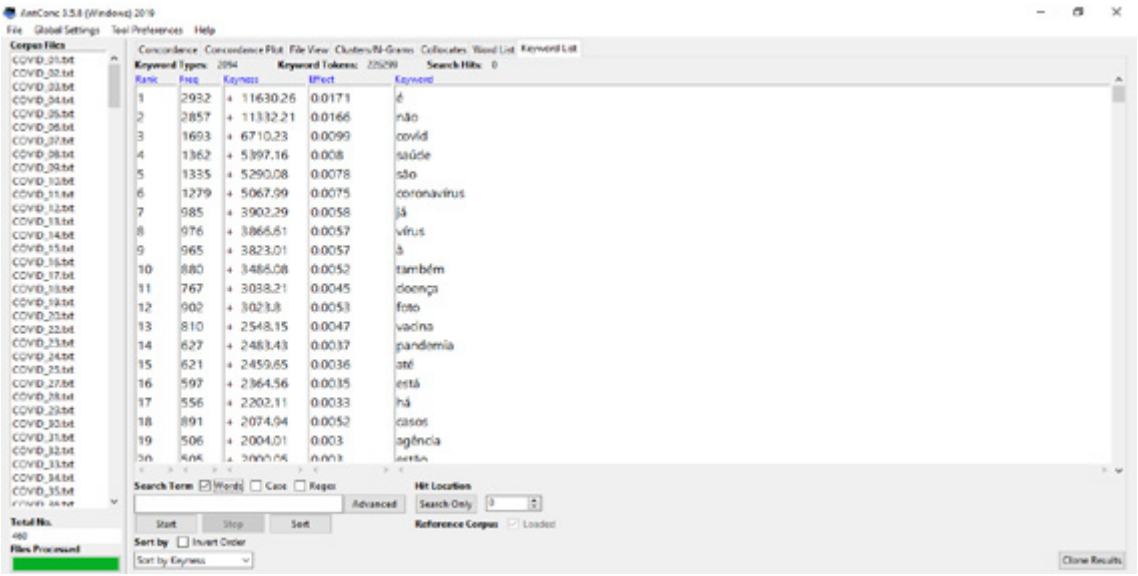
Após o levantamento da *wordlist,* verificamos as *keywords* a partir da comparação com o *corpus* de referência escolhido. A seguir, apresentamos duas imagens referentes às primeiras palavras-chave que ocorrem em nossos *corpora* de estudo:

**Figura 1.** Recorte da lista de palavras-chave mais frequentes do CO




**Fonte:** Elaboração própria utilizando o programa *AntConc*.

**Figura 2.** Recorte da lista de palavras-chave mais frequentes do CJ




**Fonte:** Elaboração própria utilizando o programa *AntConc*.

É interessante observar nas Figuras 1 e 2 que os dois *corpora* trazem, em alguns casos, as mesmas palavras com alta frequência. São exemplos desse dado: *saúde*, *covid*, *casos* e *vacina* — o que nos dá uma ideia do tipo de informação acerca da pandemia que tem sido mais veiculada por essas fontes. Em termos quantitativos, as informações diferem em virtude da extensão dos *corpora*. Nesse sentido, encontramos, no CO, 4.864 *keyword types* e 2.447.737 *keyword tokens*. Já no CJ, há 2.094 *keyword types* e 226.299 *keyword tokens*.

Das quase 3 milhões de *keywords* encontradas no total, selecionamos 672. Para chegarmos a esse número total, juntamos as 532 unidades retiradas do CO às 461 do CJ, excluindo as ocorrências repetidas cuja frequência foi somada. Nesse processo, foram escolhidas unidades substantivais, verbais e adjetivais que nos pareciam ter maior relevância em relação à temática da COVID-19, e foram descartados outros tipos, tais como os adverbiais.

Com relação às formas de uso dos substantivos levantados, observamos que diversas palavras-chave (como *medidas*, *ações*, *efeitos* e *casos*, dentre outras) são mais recorrentes, no CO e no CJ, em sua forma plural do que singular. *Medidas*, por exemplo, possui aproximadamente o triplo de ocorrências em relação à sua forma no singular (3.528 x 1.041 no CO e 505 x 169 no CJ). Esse dado pode indicar uma característica muito particular à terminologia relacionada à COVID-19: a de que a ideia de conjunto, de coletivo, seja essencial aos conceitos aos quais essas unidades se referem.

Na sequência, apresentamos um recorte da lista final em ordem alfabética das *keywords* mescladas, na qual a cor preta se refere às palavras do CJ, a cor vermelha diz respeito às palavras do CO e a cor laranja representa as palavras que ocorrem nesses dois *corpora*:

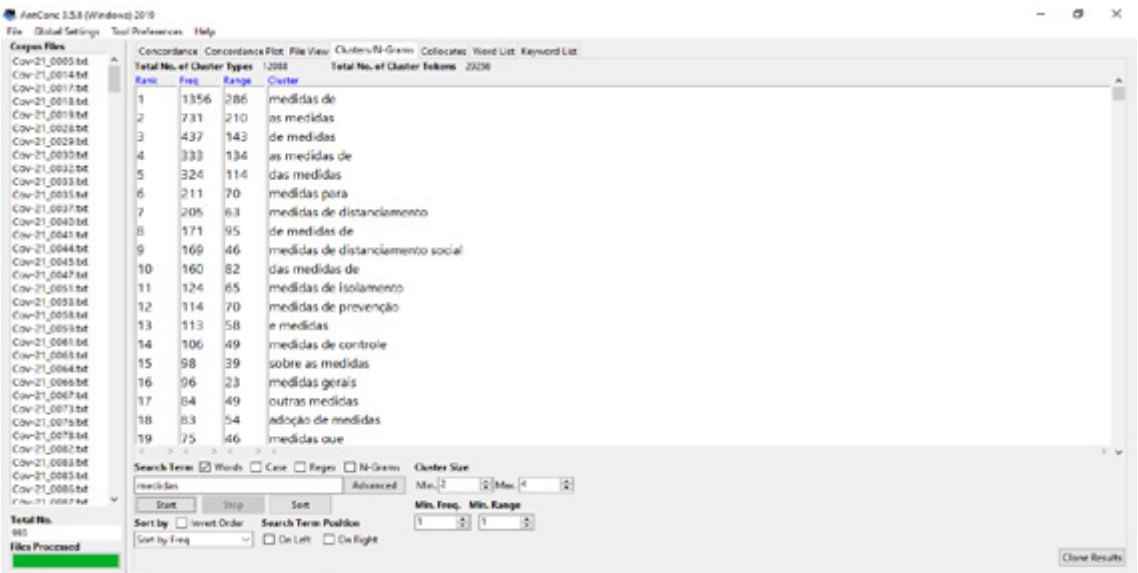
1. ação
2. acelerar
3. acesso
4. achatar
5. ações
6. adenovírus
7. adoção
8. adotadas
9. adotar
10. adverso/adversos
11. aerossóis/aerossol
12. aferição
13. afetados
14. afrouxamento
15. agência
16. agendamento
17. agente
18. aglomeração/ aglome-rações
19. agravamento
20. agravar
21. aguda/agudas
22. ajuda
23. álcool
24. alerta
25. alta
26. alta
27. alterações
28. alvo
29. ambientes
30. ambulâncias
31. ambulatorial
32. ambulatório
33. amostra/amostras
34. ampolas
35. anafilaxia
36. análise/análises
37. antibiótico/ antibióticos
38. anticoagulação
39. anticoagulante/ anticoagu-lantes
40. anticorpo/anticorpos
41. antígeno/antígenos
42. antirretrovirais
43. antivirais/antiviral
44. anvisa
45. aplicação
46. aplicação
47. aprovação
48. aquisição
49. áreas
50. armazenamento
51. arterial
52. artificial
53. asma
54. assintomática/ assintomáticas/ assintomático/ assintomáticos
55. assistência
56. associação
57. astrazeneca
58. atenção
59. atendimento/ atendimen-tos
60. ativo
61. aumento
62. autocuidado
63. autoimune/ autoimunes
64. automedicação

A comparação das listas mescladas de *keywords* nos revela algumas características textuais interessantes. Nesse sentido, é possível observar, no recorte apresentado anteriormente, que o CJ tende a trazer palavras menos opacas, tais como, *aplicação*, *automedicação* e *avanço*, enquanto o CO veicula uma terminologia "mais especializada" (*adenovírus*, *anafilaxia* e *antirretrovirais*, por exemplo). Palavras mais "populares" são encontradas nos dois *corpora* (embora mais frequentes no CJ), e, de modo geral, são elas que se destacam em laranja (álcool, *aplicação*, *aprovação*, dentre outras).

Em meio a essas unidades em comum, observamos também o tipo de unidade lexical que vem sendo mais divulgado por essas fontes no Brasil. Nesse sentido, encontramos *aglomeração*, *distanciamento*, *cloroquina*, *isolamento*, *ivermectina* e *lockdown*, por exemplo, que refletem o modo como a população e os governantes brasileiros reagiram à situação pandêmica, seja em relação a medidas de prevenção ao espalhamento da doença (*distanciamento*, *isolamento* e *lockdown*), seja sobre o desrespeito a essas determinações governamentais por parte da população em geral (*aglomeração*), ou ainda acerca da defesa e do incentivo ao chamado *tratamento precoce* (por meio do consumo de *cloroquina* e *ivermectina*).

Feita a seleção das *keywords*, essas unidades foram utilizadas para realizarmos a busca por combinatórias com 2 a 4 componentes. Para tanto, valemo-nos da ferramenta *cluster* do programa *AntConc*, conforme explicamos na seção anterior. No total, selecionamos 5.898 expressões. A título de ilustração, a Figura 3 traz um recorte da lista de *clusters* de *medidas*:

**Figura 3.** Recorte da lista de *clusters* com *medidas* do CO




**Fonte:** Elaboração própria utilizando o programa *AntConc*.

A Figura 3 nos mostra combinatórias interessantes com *medidas*. Temos, por exemplo, *medidas de distanciamento* (205), *medidas de isolamento* (124) e *medidas de prevenção* (114), que apresentam maior recorrência no CO. Essas expressões, que se referem às ações determinadas por meio de decretos provenientes de diferentes níveis da organização governamental brasileira (federal, estadual e municipal) com o objetivo de conter a disseminação da COVID-19, também figuram entre os primeiros *clusters* no CJ por ordem de frequência. A recorrência dessas unidades sintagmáticas nos dois *corpora* pode ser um indicativo do quanto essas medidas foram constantes no cotidiano dos brasileiros durante o período abarcado pelos *corpora*.

Observamos, no decorrer do levantamento dos *clusters*, uma alta produtividade com relação aos adjetivos encontrados nos *corpora*. Dentre eles, *novo* (e suas flexões em gênero e número) se mostrou bastante recorrente, com um total de 9.809 ocorrências nos dois *corpora* juntos, vindo frequentemente associado aos substantivos *coronavírus*, *casos*, *cepa*, *variantes* e óbitos, por exemplo. Outros adjetivos, tais como, *grave*, *agudo*, *sério*, *severo*, *moderado* e *leve*, são também frequentes, especialmente no CO, o que pode ser explicado pelo fato de a terminologia estudada se referir a uma temática que traz, além de *doença*, outros campos relacionados (*diagnóstico*, *sintomas*). Verificamos ainda que os adjetivos *sintomático(a)s*, *assintomático(a)s* e *precoce(s)* são particularmente produtivos e co-ocorrem com diferentes substantivos (*indivíduo*, *paciente*, *doente*, *infecção*, etc., no caso dos dois primeiros, e *tratamento*, *isolamento*, *consulta*, *diagnóstico*, *intervenção*, dentre outros, no caso do terceiro exemplo adjetival). Esse dado pode indicar uma característica dos discursos difundidos (oficial e extra-oficialmente) sobre a COVID-19 no Brasil, que destacaram a relevância dos sintomas (que podem ou não estar presentes nos doentes; daí a importância de termos um cuidado redobrado para evitar a disseminação e o contágio), bem como a ideia de que seria possível minimizar os efeitos da doença se atitudes precoces fossem tomadas (algo de que não há, até o momento, comprovação científica).

## Considerações finais

Como explicamos anteriormente, encontramo-nos em processo de organização dos dados levantados a partir do CO e do CJ com o auxílio da ferramenta *clusters* do programa *AntConc*. Até o momento, identificamos variantes diversas, tais como, *evento adverso pós vacinação* x *EAPV*, *ficha de notificação* x *formulário de notificação*, óbito x *morte*, dentre outras, cuja classificação configura, juntamente com a elaboração das definições e a organização do Dicionário proposto, etapas futuras de nosso trabalho.

O estudo preliminar que trouxemos neste artigo revela que o tratamento de *corpora* textuais por meio de ferramentas computacionais pode nos dar indícios de características gerais dos textos que os compõem. Essas características estão relacionadas não só a dados estatísticos, mas também às informações semânticas neles construídas. De

modo particular, essas informações têm relação direta com aspectos socioculturais que, inevitavelmente, subjazem aos discursos especializados. No caso da temática da COVID-19, notamos, por meio dos padrões lexicais recorrentes em nossos *corpora*, que a pandemia trouxe transformações em nível vocabular que são o reflexo de novas formas de "viver" e "conviver" que nos foram impostas com o intuito de conter o contágio e a expansão do novo coronavírus.

## Agradecimentos

As autoras Lucimara Alves Costa e Beatriz Curti-Contessoto agradecem a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento das pesquisas.

## REFERÊNCIAS

AIJMER, K.; BENGT, A. *English Corpus Linguistics*. New York: Longman, 1991.

ANTHONY, L. *AntConc* (Version 3.5.8) [Windows]. Tokyo, Japan: Waseda University. Disponível em: http://www.laurenceanthony.net/software/antconc/. Acesso em: 14 jul. 2020.

BARROS, L. A. *Conhecimentos de Terminologia geral para a prática tradutória.* São José do Rio Preto: NovaGraf, 2007.

BERBER SARDINHA, T. Linguística de Corpus: histórico e problemática. *Delta,* v. 16, n. 2, p. 323-367, 2000.

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

BERBER SARDINHA, T. Lexicology and corpus linguistics – an introduction. *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada*, *[S. l.]*, v. 25, n. 2, 2009. Disponível em: https://revistas.pucsp.br/index.php/delta/article/view/28245. Acesso em: 16 maio. 2022.

CABRÉ, M. T. Constituir un corpus de textos de especialidad: condiciones y possibilidades. *In*: BALLARD, M.; PINEIRA - TRESMONTANT, C. (ed.). *Les corpus en linguistique et en traductologie*. Arras: Artois Presses Université, 2007.

HWANG, A. Lexicografia: dos primórdios à nova Lexicografia. *In*: HWANG. A. D; NADIN, O. L. (org.). *Linguagens em Interação III:* estudos do léxico. Maringá: Clichetec, 2010.

IULA. Terminología, ingeniería lingüística y lingüística computacional [en línea]. En Grupo IulaTerm. *Diploma de postgrado online: Terminología y necesidades profesionales, 9a ed*. Barcelona: IULA. Universidad Pompeu Fabra, 2015.

KADER, C. C. C.; RICHTER, M. G. Linguística de Corpus: possibilidades e avanços. *Instrumento: R. Est. Pesq. Educ*. Juiz de Fora, v. 15, n. 1, p. 13-23, jan./jun. 2013.

KILGARRIFF, A.; RIGAU, I. EsTenTen, a vast web corpus of Peninsular and American Spanish. *In*: VARGAS-SIERRA, C. (org.). *Corpus resources for descriptive and applied studies. Current Challenges and Future Directions*: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013). Alicante, Spain, 2013. p. 12-19. DOI: https://doi.org/10.1016/j.sbspro.2013.10.617.

LEHMANN, A. Présentation*. Langue Française*, Montrouge, n. 106, 1995.

MATUDA, S.; TAGNIN, S. A terminologia do futebol: um estudo direcionado pelo *corpus*. *Letras & Letras*, v. 30, n. 2, p. 214-243, 2014.

MEIOEMENSAGEM. *Circulação dos maiores jornais do País cresce em 2019*, 2020. Disponível em: https://www.meioemensagem.com.br/home/midia/2020/01/21/circulacao-dos-maiores-jornais-do-pais-cresce-em-2019.html. Acesso em: 02 mar. 2022.

ORENHA, A. Aplicações léxico-terminográficas da Línguística de Corpus: relato da elaboração de um glossário bilíngue de colocações na área de negócios. *Intercâmbio*, v. 13, p. 1-8, 2004.

PARADIS, C.; AUGER, P. La terminotique ou la terminologie à l'ère de l'informatique. *Meta*, v. 32, n. 2, p. 102-110, 1987.

RENOUF, A.; KEHOE, A.; BANERJEE, J. WebCorp: an integrated system for web text search. *In:* HUNDT, M.; NESSELHAUF, N.; BIEWER, C. (org.). *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 2007.

SCOTT, M. *WordSmith Tools version 6*. Oxford: Oxford University Press, 2015. Disponível em: https://lexically.net/wordsmith/version6/index.html. Acesso em: 31 maio 2019.

TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins, 2001. DOI: http://dx.doi.org/10.1075/scl.6

ZANCHETTA, E.; BARONI, M.; BERNARDINI, S. Corpora for the masses: the BootCaT front-end. *Corpus Linguistics 2011 Conference*, Birmingham. Abstracts. Birmingham: University of Birmingham, 2011.