30º CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO
25 a 29 de novembro 2024
Bibliotecas Fortes: Sociedade Democrática
Recife, PE

Eixo 3 – Formação e identidade profissional

Modalidade: trabalho completo

Diretrizes metodológicas para compilação e manipulação de corpus para Sistematização de Terminologias: a compilação e a manipulação do corpus da COVID-19

Methodological guidelines for compiling and manipulating corpus for Systematization of Terminologies: compilation and manipulation of the COVID-19

Valdirene Pereira da Conceição – Universidade Federal do aranhão (UFMA)

Maria Leoquiane Oliveira Guimarães – Universidade Federal do aranhão (UFMA)

Roosewelt Lins Silva – Universidade Federal do aranhão (UFMA)

Resumo: Estudo sobre a Sistematização da Terminologia da Covid-19. É uma pesquisa exploratória de natureza descritiva, que tem como objetivo analisar o processo de compilação e manipulação do corpus da terminologia da COVID-19 para auxiliar no desenvolvimento de modelagem léxico-ontológica, os estudos, compartilhamento de informações e na recuperação da informação. Adota a Teoria Comunicativa da Terminologia, os estudos do léxico e do Processamento de Linguagem Natural (PLN)como subsídio para elaboração de vocabulários controlados, organização e recuperação de informação. Apresenta como resultado da pesquisa em andamento a etapa de copilação do corpus, assim como a descrição de etapas que vão da seleção de fontes, geração do corpus, extração e montagem do corpus, implementação e disponibilidade em ambientes computacionais.

Palavras-chave: COVID-19. Terminologia de domínio. Estudos léxicos. Onto-léxico. Processamento de Linguagem Natural (PLN).

Abstract: Study on the Systematization of Covid-19 Terminology. It is an exploratory research of a descriptive nature, which aims to analyze the process of compiling and manipulating the corpus of COVID-19 terminology to assist in the development of lexical-ontological modeling, studies, information sharing and information retrieval. It adopts the Communicative Theory of Terminology, lexical studies and Natural Language Processing (NLP) as a basis for creating controlled vocabularies, organizing and retrieving information. It presents the corpus compilation stage as a result of ongoing research, as well as a description of steps ranging from source selection, corpus generation, removal and assembly of the corpus, implementation and availability in computational

environments.

Keywords: COVID-19 Domain terminology. Lexical studies. Onto-lexicon. Natural Language Processing (PLN).

1 INTRODUÇÃO

O estudo, ora em tela, é parte do Projeto "Sistematização da Terminologia da COVID - 19: desenvolvimento de um modelo léxico-ontológico para o domínio da COVID-19", cujo objetivo geral consiste em sistematizar a terminologia da COVID-19, por meio de uma estrutura conceitual (ontologia) a partir da compilação de corpus em língua portuguesa - língua de chegada (LC) - e do uso da nomenclatura em inglês - língua de partida (LP), de vários recursos informacionais e fontes com características diversas, para auxiliar os estudos e pesquisas no enfrentamento da pandemia da COVID-19 e outras síndromes gripais, bem como a organização e recuperação da informação em diversos ambientes. É vinculado ao Departamento de Biblioteconomia, possui uma equipe multidisciplinar agregando bibliotecários, docentes e discentes das áreas de Biblioteconomia, Enfermagem e Ciências da Computação em seus diferentes níveis acadêmicos.

Os estudos sobre a COVID-19, ainda hoje, é uma das áreas centrais das atividades de pesquisa, desenvolvimento e inovação no mundo, especialmente nos países desenvolvidos (FGV, 2024). Em 2020, a Organização Mundial da Saúde (OMS) declarou o estado de pandemia pelo vírus SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), agente etiológico da doença COVID-19 (Zhu et al., 2020). Desde então, os países que detêm tecnologia avançada no campo da ciência têm se empenhado em disponibilizar de modo intensivo investimentos que são aplicados nesse domínio emergente de conhecimento, na perspectiva de conhecer sobre o vírus e suas variações, e ainda, buscar formas de combatê-lo para salvar a existência humana ameaçada pelo surgimento do vírus SARS-CoV-2, que se espalhou de forma pandêmica pelo mundo todo, naquele ano.

No contexto do avanço da ciência em relação às pesquisas no que tange ao desenvolvimento científico e tecnológico demandado da urgência de encontrar o meio de deter o estado de calamidade em saúde pública causado pelo momento pandêmico, além de investimentos tanto da iniciativa pública como privada é preponderante a

sistematização de repertórios vocabulares em língua portuguesa sobre a COVID-19, para que as áreas de interesse possam dialogar entre si, fazendo uso de uma linguagem unificada sobre a referida temática.

Logo, sistematizar terminologias significa criar termos confiáveis de forma a facilitar a comunicação especializada, além de demonstrar que a língua portuguesa está apta para nomear conceitos técnicos e científicos, em meio à especificidade dos estudos sobre a COVID 19 (Cabré, 2003; Almeida, 2006). Em outras palavras, ao mesmo tempo em que se promove a disseminação de conhecimentos e de tecnologias, fomentam-se os estudos sobre a organização do conhecimento e a elaboração de repertório em língua portuguesa, que traduza os artefatos e produtos resultantes das pesquisas, posto que são poucos os glossários e/ou dicionários sobre a COVID-19, em língua materna.

Os estudos iniciais para a concepção do desenvolvimento de uma estrutura conceitual (modelo léxico–ontológico) do domínio da SARS-COVID-19 apontaram uma grande variedade de possíveis tópicos e atividades, ligados à pesquisa acadêmica e aplicações no desenvolvimento de produtos e serviços comerciais/bases de dados, que podem ser adotados na estruturação conceitual do domínio. De fato, um levantamento de portais e páginas da Internet dedicados à COVID-19, em inglês, português e espanhol, mostrou que a estruturação do domínio varia enormemente, dependendo dos interesses específicos da Instituição que se dedica à pesquisa e mesmo de autores, que dela tratam. Por exemplo, há páginas em que a cobertura da literatura se restringe aos sintomas e tratamento, enquanto outras se concentram na fabricação de vacinas e sequelas da COVID-19. Mencione-se, também, a carência de glossários abrangentes, pelo menos online. Os glossários de COVID-19 encontrados são limitados, em abrangência e profundidade, sendo praticamente quase todos em inglês. Isso não é surpreendente, haja vista a natureza inter e multidisciplinar dessa área que ainda está se consolidando.

Na perspectiva conceitual da Organização do Conhecimento (OC)/Ciência da Informação (CI), Domínio de Emergência Subta (DES) – é o termo adotado para nomear áreas do conhecimento que levaram nos últimos anos à eclosão de situações de realidade não antecipadas por projetos e estudos (Barité, Fontans, 2021). A exemplo dos eventos tais como o (de) Ebola, demolição das torres gêmeas, e, atualmente, a pandemia da COVID-19 causada pelo vírus SARS-CoV-2 que têm modificado o modo de

saber e fazer, pelo menos, durante a pandemia, dos diversos atores (organizações profissionais, pesquisadores, gerentes de bibliotecas e designers de sistemas de organização do conhecimento) de uma comunidade de discurso e prática. Nessa direção, este estudo tem como objetivo apresentar as diretrizes metodológicas adotadas na

compilação e manipulação do corpus-fonte da COVID-19, pautado em estudo multidisciplinar - nas abordagens da Teoria Comunicativa da Terminologia (TCT) Cabré (1999, 2002), Estudos do Léxico, (Biderman, 1998, 2001, 2006) Organização do Conhecimento (OC), Café; Sales, (2010) Brascher; Café, (2008), e do Processamento de Línguas Naturais (PLN), Dias da Silva et al (2007), (Sousa, 2007) - de natureza aplicada para estabelecer uma estrutura conceitual (ontologia), legível por máquina e por humano, para o domínio da COVID-19, de modo que possa não apenas fornecer subsídios para auxiliar os estudos e pesquisas de abrangência de grupos de usuários com interesses e propósitos comuns, mas também guiar a elaboração de produtos, serviços e uso de tecnologias no processamento e na organização do conhecimento acerca do tema, nos diversos ambientes informacionais.

2 DIRETRIZES METODOLÓGICAS DE COMPILAÇÃO E MANIPULAÇÃO DE CORPUS

A OC se ocupa com os estudos e pesquisas acerca da sistematização de métodos e técnicas para fins temáticos de representação e recuperação da informação, e, também, com a criação de Sistemas de Organização do Conhecimento (SOC) (Albrechtsen, 1993; Hjørland, 1997).

De modo geral, considerando o pressuposto da interdisciplinaridade presente na Ciência da Informação, a aplicação de técnicas e métodos de outros domínios do conhecimento tem sido cada vez mais comum, e além dos já mencionados, citamos o PLN, microcampo da Linguística Computacional e subcampo da Inteligência Artificial, que estuda a possibilidade de computadores simularem a compreensão e o entendimento humano no que se refere ao processamento automático da língua natural (Dias da Silva, 2007).

A análise e sistematização da terminologia do domínio da COVID-19 envolve compreensão, mapeamento e uso para diversos fins, incluindo a criação e revisão de

sistemas de organização do conhecimento "como, tesauros, listas de assuntos, ontologias, taxonomias", especialmente em ambiente digital.

A estrutura conceitual, ou seja, a ontologia, constitui uma representação da realidade no âmbito do domínio que se toma como objeto de estudo, neste caso, SARS-CoV-2 e COVID-19. Essa representação busca coletar e organizar o conhecimento e as ramificações que lhes são próprias, refletindo de forma esquematizada o modo de saber e conhecer da área em questão. Por conseguinte, necessário se faz estabelecer, as ocorrências de variações terminológicas linguísticas (morfológicas, lexicais e sintáticas) e de registro (discursivas) em potencial, com o propósito de aprofundar o entendimento das complexidades conceituais inerentes a essa área de estudo.

A importância de obtermos uma organização conceitual (ou ontológica) para uma área emergente é que tal organização propicia, dentre outros aspectos:

a) a comunicação mais eficaz entre os especialistas (disseminação de conhecimento)
b) a elaboração de produtos e serviços de organização, busca e recuperação da informação, a exemplo de confecção de glossários, vocabulários e índices (Cabré, 1999).

Esses produtos podem ser criados por uma equipe interdisciplinar, formada por bibliotecários, linguistas, enfermeiros (demais profissionais da área de saúde) e pessoal da TI (com a ajuda de especialistas), através de um sistema colaborativo de extração e elaboração de terminologias.

Propomos, para a organização conceitual que aqui vai se levar em conta como base, principalmente os pressupostos da pesquisa Terminológica de natureza descritiva, que se tem mostrado eficaz para sistematizar as linguagens de especialidade (Cabré, 2003; Almeida; 2003; Conceição; 2011; 2014). Segundo essa perspectiva teórica, a organização conceitual de uma área de especialidade pressupõe a realização das seguintes atividades de natureza teórico- metodológicas:

(1) delimitação da área de conhecimento (área objeto);
(2) identificação das instituições, associações e/ ou demais organismos que representam e/ ou fazem parte dos setores envolvidos com a área- objeto;
(3) identificação dos representantes de cada um dos setores acima mencionados (especialistas-interlocutores);

(4) seleção do corpus fonte (textos escritos, textos digitais, fontes orais, etc.);

(5) extração automática (do corpus -fonte) de “termos candidatos” para auxiliar a proposição de uma estrutura conceitual para a área em questão;

(6) edição e gerenciamento da estrutura conceitual de maneira semiautomática, utilizando ferramentas computacionais (inclusive via Web);

(7) verificação manual da estrutura conceitual gerada e seu aprimoramento por meio do feedback de especialistas do domínio.

Trata-se de um rol de atividades centradas nos estudos da linguística, lexicografia, e terminologia, abordando tópicos específicos para ontologias, bem como nas atividades relacionadas à aplicação de tecnologias da informação, com ênfase tanto no uso de ferramentas de software para extração automática de termos a partir de corpus e na definição da própria ontologia (Teline, 2004) e (Almeida, 2003). Em linhas gerais, envolve:

a) Copilar o corpus a partir de várias fontes e com características diversas, incluindo: títulos e resumos de artigos indexados pelo ISI Web of Knowledge (Web of Science), bem como de artigos completos de periódicos especializados, livros dedicados ao tema, títulos e resumos de patentes, sites especializados, dentre outras, com vistas a identificar e analisar a taxonomia existente em livros, bases de dados, páginas da Web e programas institucionais e governamentais dedicados a COVID -19;

b) extrair termos do corpus (acima descrito) por meio de técnicas de extração automática, estatística, técnicas bibliométricas e de Processamento de Línguas Natural (PLN);

c) elaborar e implementar no Editor de Ontologias uma estrutura conceitual do domínio da COVID-19, para fins de compartilhamento, busca e recuperação da informação de interesse comuns de pesquisadores; caracterizar o Domínio da COVID-19, por meio da categorização de assuntos que constituem o conhecimento nesse ramo do saber.

Tais procedimentos, auxiliam tanto na análise das variantes terminológicas linguísticas (aspectos morfológico, lexicais e sintáticos), como na análise das variantes terminológicas de registro (aspectos discursivos) e na própria elaboração do modelo conceitual léxico ontológico. Isto porque, como se sabe, o léxico é composto por um

acervo de unidades lexicais, termos e expressões disponíveis para os falantes de uma língua, além de ser responsável por constituir o patrimônio linguístico desses falantes.

3 METODOLOGIA

Este estudo é caracterizado como uma pesquisa exploratória de natureza descritiva aplicada em que utiliza como procedimentos técnicos e meios de investigação a pesquisa bibliográfica e documental. A pesquisa aplicada "objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos" (Prodanov; Freitas, 2013, p. 51), tendo em vista a viabilidade da aplicação prática dos resultados obtidos para contribuir em determinada realidade, neste contexto, objetiva-se gerar conhecimentos no âmbito da compilação e manipulação de corpus aplicados no desenvolvimento de um modelo léxico-ontológico da terminologia da COVID-19.

4 A COPILAÇÃO E MANIPULAÇÃO DE CORPUS DA COVID-19: RESULTADOS INICIAIS

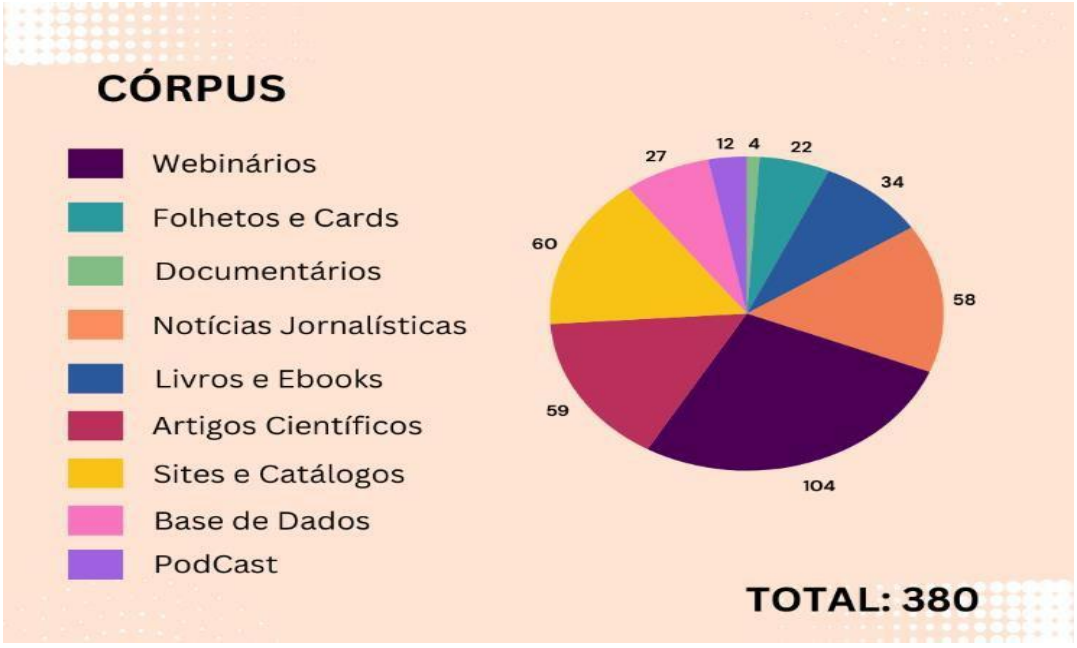
A organização do corpus foi realizada a partir das diretrizes de Almeida (2006) que expõe a necessidade de haver um conjunto de requisitos de forma a garantir a validade e confiabilidade do corpus. Assim, teve-se como ponto de partida, a concepção de que "Para organizar um corpus, parte-se, inicialmente, da seleção dos textos pertinentes e relevantes para a pesquisa, bem como dos gêneros aos quais eles pertencem." (Almeida, 2006, p. 88).

A autora destaca também, a importância da variação de gêneros textuais na composição do corpus, tendo em vista as possibilidades de representatividade na comunicação de determinados domínios." Portanto, mesmo em se tratando de uma pesquisa terminológica, o corpus deve ser balanceado e diversificado, contendo, pelo menos, textos dos gêneros: técnico-científico, científico de divulgação e instrucional." (Almeida, 2006, p. 88).

Logo, tomou-se como corpus escolhido para realização deste estudo, diversos tipos de fonte e gêneros textuais, como se pode observar nas figuras, 1, 2, 3 e 4, com seus respectivos índices, incluindo do artigo científico voltados para a temática do Covid-19, após buscas realizadas na base do Portal de Periódicos da Capes, com a utilização de termos como: COVID-19; Corona Vírus; Pandemia; Novo Corona Vírus, tendo como

critério de seleção, a ordem de relevância em que foram recuperados, como instituições, bases de dados, webinários e profissionais envolvidos com estudos sobre COVID-19. Como mostram as ilustrações:

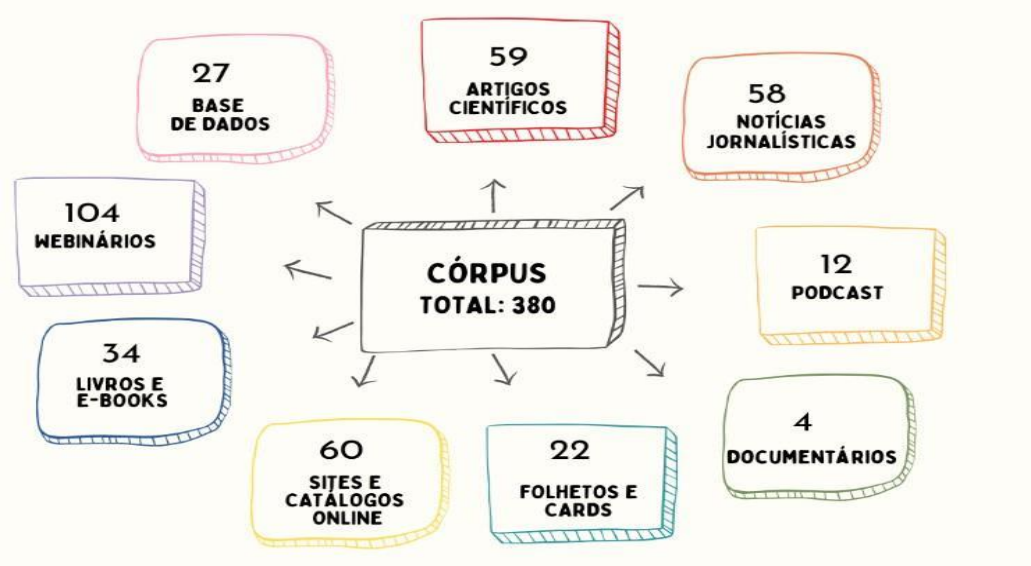
Figura 1 – Identificação quantitativa do Corpus



Fonte: As autoras (2024)

Descrição: #ParaTodosVerem. É uma figura com a identificação do corpus, com a descrição de quantos itens foram compilados por tipo de documento, apresentado por um gráfico colorido.

Figura 2 – Corpus identificado por tipologia



Fonte: As autoras (2024)

Descrição: #ParaTodosVerem. É uma imagem demonstrando o quantitativo de corpus analisados, por tipo de material, como por exemplo: em 104 webinários foram compilados termos relacionados a Covid-19.

Figura 3 – Compilação dos 380 itens

NOME DOS PALESTRANTES	TÍTULO	ASSUNTO	REALIZAÇÃO	DATA DE PUBLICAÇÃO	TEMPO DE DURAÇÃO	PLATAFORMA
Adam Jaffe	Understanding COVID 19 Webinar : Q&A for parents and carers of children with rare lung disease	Doença Pulmonar	Lung Foundation Australia	20 maio 2020	Video (58:44)	YouTube Canal : Lung Foundation Australia Link: https://www.youtube.com/watch?v=Vmaf_mVTwds&list=PLx5RMZtuxzUTAFrLPN75y7foY8jbNPaWN&index=6
Alois Jungbauer	Virus and virus-like particle and extracellular vesicles purification	Virús	Universidade Federal de São Paulo(Unifesp)	10 ago. 2020	Vídeo (1:19:54)	Youtube- Canal: Unifesp Link: https://www.youtube.com/watch?v=nbREFbdQoXs
1.Andrew Baum 2.Sabina Kalyan	Oxford Real Estate Webinar - The Macro Effects of Covid-19	Efeitos da Covid-19	University of Oxford e Said Business School	28 abr. 2020	Vídeo (51:37)	YouTube Canal:Said Business School, University of Oxford Link:https://www.youtube.com/watch?v=D3gulcFZ_OA
1.Cécile Viboud 2.Michael Levitt 3.Pedro Curi Hallal	Epidemiological monitoring and measurement of infectivity rates in key countries	monitoramento epidemilogico Taxas de infectividade	Agência Fapesp	21 maio 2020	2:12:01	YouTube Canal: Agência Fapesp Link: https://www.youtube.com/watch?v=_1u_yworTco
Dan Chambers	Understanding COVID 19 webinar: Information for Interstitial Lung Disease Patients	Doenças Pulmonares	Lung Foundation Australia	15 abr. 2020	Vídeo (58:13)	Youtube Canal Link: https://www.youtube.com/watch?v=SpKiqtWKt5U&list=PLx5RMZtuxzUTAFrLPN75y7foY8jbNPaWN&index=2
Debra Sandford	Understanding COVID 19 Webinar: Navigating life after COVID - 19 restrictions ease	Restrições	Lung Foundation Australia	17 de jul. 2020	Vídeo (57:58)	Youtube- Canal: Lung Foundation Australia Link: https://www.youtube.com/watch?v=1tEZrWF1_v4&list=

Fonte: As autoras (2024)

Descrição: #ParaTodosVerem. É uma imagem de um quadro, de tonalidade lilás e branco, com dados da compilação da terminologia da Covid-19 em Webinários Internacionais, com as seguintes colunas: nome dos palestrantes; titulo; assunto; realização; data de publicação; tempo de duração e plataforma.

Figura 5 – Sites e catálogos online

MATÉRIA	ASSUNTO	ACESSO
	Covid-19, Coronavírus	https://www.paho.org/pt/covid19
	Covid-19, Coronavírus	https://covid19.who.int/
	Covid-19, Coronavírus	https://coronavirus.saude.gov.br/

Fonte: As autoras (2024)

Descrição: #ParaTodosVerem. É uma imagem de um quadro, de cor amarela e branca, com a descrição da matéria, assunto e acesso de sites e catálogos online, referente aos termos identificados na Covid-19.

Estabelecemos então, como método de trabalho, uma sequência de etapas que devem integrar um projeto terminológico e que tenham como embasamento teórico norteador, como destacamos anteriormente, a Terminologia de orientação descritiva, fundamentada em princípios da Linguística e que servem de base para a elaboração dos instrumentos de Organização do Conhecimento/Linguagem Documentária, a exemplo de tesauros, vocabulários controlados e glossários, como recomendam Cabré (2003), Almeida, Oliveira e Aluísio (2006), Conceição (2014), Barbosa, Mey e Silveira (2005) especificamente na elaboração de vocabulário controlado, da seguinte forma:

4.1 Compilação, manipulação do corpus e nomeação de arquivos

A compilação consiste no armazenamento em arquivos predeterminados de todos os textos pertinentes e relevantes para a pesquisa. Para essa compilação, serão utilizados os seguintes termos de busca: COVID-19, pandemia da COVID-19, síndromes gripais, sintomas da COVID-19, tratamento da COVID-19, sequelas da COVID-19, vacinas

da COVID-19, dentre outros indicados na literatura para formar o corpus em LI. A manipulação do corpus consiste nas seguintes atividades, a saber: a) conversão manual e automática (Pacote XPDF) de formatos "doc", "html" e "pdf" para "txt", assim como a utilização de um concordanceador (e demais ferramentas de – PLN que se mostrarem pertinentes) e a extração automática, e limpeza e formatação, de maneira a preparar o corpus para o processamento computacional (Almeida; Aluísio; Oliveira, 2006).

a) Nomeação de arquivos e geração de cabeçalhos. Depois que todos os textos forem convertidos em formato “txt”, eles devem receber um nome. Ressalte- se que essa nomeação deve seguir determinado padrão de forma a facilitar a recuperação posterior de cada texto. Após a nomeação dos arquivos, é gerado (de forma semiautomática) um cabeçalho para cada texto. A geração semiautomática desse cabeçalho será feita por meio de um editor (programa computacional “com interface gráfica” para criar ou modificar arquivos) que auxilia o bibliotecário na especificação de diversos metadados sobre os textos, a exemplo de: titulo, subtítulo, fonte, editor, local de publicação, data, assunto, autoria, tipo de autoria (individual ou coletiva), sexo do autor, tipo de texto, meio de distribuição e comentários (introduzem-se nesse campo informações adicionais sobre o texto). O preenchimento de todos esses campos é útil para esta pesquisa porque a partir desses dados será possível fazer constatações tais como: o repertório vocabular tem alguma relação com a temática do texto, com o gênero, com a autoria ou com o meio de distribuição? Identifica a instituição com maior produção? Enfim, considera-se que será possível fazer constatações relevantes sobre a língua de especialidade e a produção técnica-científica da área em questão.

4.2 Categorização dos conceitos e Inserção dos termos na ontologia

Do domínio e sua validação pelos especialistas (Mapa conceitual); elaboração e preenchimento das fichas terminológicas para coleta de dados e informações enciclopédicas (quando for o caso);

Inserção dos termos na ontologia. Os termos obtidos devem ser inseridos na ontologia por meio do editor Protegé, por isso ela deve ser organizada preliminarmente ou concomitantemente à extração dos termos, já que, à medida que os termos vão

sendo obtidos, é que se pode ter uma visão real de quais serão os campos nocionais/ categoriais que deverão integrar a ontologia. A ontologia é uma organização semântica da área-objeto, semelhante ao que se entende por árvore de domínio, a diferença é que os conceitos/termos estão ali armazenados (Lima-Marques, 2006).

Organiza-se uma estrutura constituída de campos nocionais, de forma que essa estrutura reflita os conceitos da área-objeto bem como as relações entre eles.

A ontologia é uma estrutura conceitual legível por máquina fundamental para

(1) possibilitar uma abordagem mais sistemática de um campo de especialidade do saber humano;
(2) circunscrever a pesquisa, já que todas as ramificações da área-objeto, com seus campos, foram previamente consideradas;
(3) delimitar o conjunto terminológico;
(4) determinar a pertinência dos termos;
(5) prever os grupos de termos pertencentes à área-objeto, como também os que fazem parte de matérias conexas;
(6) definir as unidades terminológicas de maneira sistemática, e, finalmente,
(7) controlar a rede de remissivas (Conceição, 2014).

É conveniente destacar que a estrutura conceitual final sobre a COVID-19, será implementada e disponibilizada num ambiente computacional denominado e-Termos da USP, tal qual se fez com o vocabulário do domínio musical em 2015, com vistas a auxiliar a pesquisa e desenvolvimento de terminologias especializadas e, por conseguinte, o registro, a organização e a recuperação do conhecimento. Trata-se de um Ambiente Web Colaborativo, composto por seis módulos de trabalho independentes, mas interrelacionados, cujo propósito é automatizar ou semiautomatizar as tarefas de criação e gerenciamento do trabalho terminológico (Núcleo Interinstitucional de Linguística Computacional, 2022).

5 CONSIDERAÇÕES FINAIS

Como parte da pesquisa intitulada “Sistematização da Terminologia da Covid-19”, o estudo ora em tela, tem como objetivo analisar o processo de compilação e manipulação do corpus da terminologia da COVID-19 para auxiliar no desenvolvimento

de modelagem léxico-ontológica, os estudos, compartilhamento de informações e na recuperação da informação.

Por meio do estudo foi possível conhecer a configuração das pesquisas, autores, instituições e produções técnicas científicas sobre a Covid-19; ter melhor compreensão sobre o PLN (Processamento de Língua Natural) e seu impacto na produção do conhecimento, bem como compreender as etapas do processo de Sistematização da Terminologia de domínio e consequentemente do arcabouço teórico- metodológico do seu fazer.

Vale destacar a importância da diversidade de fontes e variação de gêneros textuais na composição do corpus, o que potencializa as possibilidades de representatividade na comunicação do domínio, como foi enfatizado no estudo ao incluir no mapeamento das fontes para montagem do corpus blogs, sites, webinários, e-books, material instrucional, artigos técnicos científicos dentre outros.

As diretrizes levadas em consideração para sistematização da terminologia em língua portuguesa sobre a COVID -19 neste estudo prevê, detalhadamente, várias etapas, que incluem desde a seleção de fontes até a compilação, geração de um corpus, validação, implementação no editor de ontologias e disponibilização na web. Para realização do mapeamento foi utilizado o google acadêmico, bancos e bases de dados, Portal de Periódicos da Capes, Ministério da Saúde, Fundação Oswaldo Cruz, ANVISA, webinários e profissionais envolvidos com estudos sobre COVID-19.

Ademais, é possível depreender que o uso de PLN é de grande valia para elaboração de modelos léxico-ontológicos com fins de organização, representação e recuperação da informação em diversos domínios de especialidade.

Espera-se que a sistematização da terminologia da SARS-CoV-2 e COVID-19 possa contribuir com a comunicação eficaz entre os pesquisadores; o compartilhamento de conhecimento acerca da origem disciplinar das pesquisas geradas, identificando os domínios pré-existentes, autores que contribuíram para a produção científica para conceber, categorizar e tratar a doença que ainda assola o mundo; a elaboração de um mapeamento inicial e provisório do domínio em fase de consolidação, a fim de desenvolver, em curto prazo, ferramentas emergenciais de baixo custo que facilitem a organização temática do fluxo documental sobre a pandemia da COVID-19.

Outro aspecto de interesse é a inserção de dados na Wikipédia e no ambiente e-

Term/USP, privilegiando-se os termos da ontologia, que pode ser criada através da edição colaborativa, mas, monitorada com conceitos, para servir de material de consulta e ensino aos pesquisadores da área. Outra forma de se usar uma ontologia no cenário da COVID-19 é a sugestão de palavras-chave pertencentes à ontologia e que estão relacionadas a uma consulta feita por um usuário em uma máquina de busca específica e também na categorização e visualização de documentos recuperados, dentre outros.

REFERÊNCIAS

ALBRECHTSEN, H. **Subject analysis and indexing from automated indexing to dominion analysis**. The Indexer, London, v. 18, n. 4, p. 219-24, 1993.

ALMEIDA, G. M. B.; OLIVEIRA, L. H. M.; ALUISIO, S. M. A terminologia na era da informática. **Cienc. Cult**. [online], v. 58, n. 2, p. 42-5, 2006.

BARBOSA, Sidney; MEY, Eliane Serrão Alves; SILVEIRA, Naira D. **Vocabulário controlado para indexação de obras ficcionais**. Brasília: Briquet de Lemos, 2005.

BARITÉ, M.; FOTNAS, E. Un mapeo terminológico del dominio covid-19 con base en bibliometría y garantía académica. In: **Congresso Isko Espanha-Portugal**, 5. Universidade de Lisboa, Faculdade de Letras, 2021. Anais...

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. Nota Técnica - **Alimentação dos sistemas de informação pelos profissionais das equipes de atenção primária**. Brasília: MS, 2015.

CABRÉ, M. T. **Theoris of terminology:** their description, prescription and explanation. Terminoly, v. 9, n. 2, p. 163-200, 2003.

CABRÉ, M. T. **La terminología**: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada, 1999.

CONCEIÇÃO, V. P. Modelagem léxico-ontológica do domínio. Tese (Doutorado em Linguistica e Lingua Portuguesa) - Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, 2011.

CONCEIÇÃO, V. P. **O léxico do patrimônio cultural de São Luís**. 1. ed. Novas Edições Acadêmicas, 2014. v. 1. 244 p.

DIAS DA SILVA, Bento Carlos et al. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. São Carlos: NILC - ICMC-USP, 2007. 121 p. (Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional). Disponível em: http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0710-DiasDaSilvaEtAl.pdf. Acesso em: 15 mar. 2024.

GRUPO DE ESTUDOS E PESQUISAS EM TERMINOLOGIA (GETerm), Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil [2] Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP), São Carlos, SP, Brasil.

HJØRLAND, B. **Information seeking and subject representation**: an activity-theoretical approach to information science. Westport: Greenwood Press, 1997. 213 p.

LIMA-MARQUES, Mamede. **Ontologias**: da filosofia à representação do conhecimento. Brasília: Thesaurus, 2006. 72 p.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL (NILC). Universidade de São Paulo (USP), São Carlos, SP, Brasil.

ZHU, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. **New England Journal of Medicine**, v. 382, n. 8, p. 727-33, 2020.