JHI JOURNAL OF HEALTH INFORMATICS

CBIS'20 7 a 11 de Dezembro Foz do Iguaçú - Paraná Evento Virtual

Elaboração de corpus biomédico em Português sobre o Covid-19

Elaboration of a biomedical corpus in Portuguese about Covid-19

Elaboración de un corpus biomédico en Portugués sobre Covid-19

Janaína da Silva Leite[1], André Kazuo Takahata[2], Margarethe Steinberger-Elias[2]

RESUMO

Descritores: Gestão de Ciência, Tecnologia e Inovação em Saúde; Betacoronavirus

Objetivos: Apresentar os resultados iniciais para elaborar um *corpus* de textos biomédicos publicados em português relacionados à COVID-19. **Métodos:** Foram realizados a extração, compilação, armazenamento e categorização de textos em português relacionados à COVID-19 da base de dados científicos *Pubmed*. **Resultados:** A metodologia executada resultou em 254 textos. O processo de categorização indicou prevalência de cerca de 30% para textos relacionados às áreas de Saúde Coletiva e Epidemiologia, em detrimento de outras áreas da medicina e de pesquisa como virologia e genômica. **Conclusão:** Foi encontrada uma maior prevalência de objetos de estudo voltados às ações de vigilância estratégica que mitiguem a evolução da pandemia de COVID-19 e colaborem no seu combate. Há a necessidade de se cobrir mais bases, anotar os textos e obter formas de se atualizar o *corpus* em trabalhos futuros.

ABSTRACT

Keywords: Health Sciences; Technology; and Innovation Management; Betacoronavirus

Objectives: To present initial efforts in developing a corpus of biomedical articles published in Portuguese related to COVID-19. **Methods:** The methodology was based on the extraction, compilation, storage and categorization of texts from the *Pubmed* scientific database in Portuguese. **Results:** The adopted methodology resulted in 254 texts. The categorization indicated a prevalence of about 30% for texts related to the areas of Public Health and Epidemiology, in detriment to other medical specialties and areas of research such as virology and genomics. **Conclusion:** The scientific literature in Portuguese language presented a larger prevalence of study objects with the objective of promoting strategic vigilance in order to mitigate the evolution of the COVID-19 pandemics. There is still the need to cover a larger number of bases, annotate the texts and elaborate strategies to update the corpus in future works.

RESUMEN

Descriptores: Gestión de Ciencia; Tecnología e Innovación en Salud; Betacoronavirus

Objectivos: Presentar los primeros esfuerzos para desarrollar un corpus de artículos biomédicos publicados en portugués relacionados con el SARS-CoV-2. **Métodos:** La metodología se basó en la extracción, compilación, almacenamiento y categorización de textos en portugués de la base de datos científicos *Pubmed*. **Resultados:** La metodologia adoptada resultó en 254 textos. La categorisación indicó una gran prevalencia de alrededor del 30% para textos relacionados con las áreas de Salud Pública y Epidemiología, en perjuicio de otras especialidades médicas y áreas de investigación más exploratorias que tratan con datos virológicos o genéticos. **Conclusión:** Se encontró en la literatura científica en portugués una mayor prevalencia de objetos de estudio orientados a acciones estratégicas de vigilancia que mitiguen la evolución de la pandemia COVID-19 y colaboren en su lucha. Es necesario cubrir más bases, anotar los textos y obtener formas de actualizar el corpus en trabajos futuros.

[1] *Mestranda em Engenharia da Informação, Universidade Federal do ABC - UFABC, Santo André (SP), Brasil.*
[2] *Professor Doutor do Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas - UFABC, Santo André (SP), Brasil*

Autor Coorespondente: **Janaína da Silva Leite**
e-mail: **janaina.leite@ufabc.edu.br**

INTRODUÇÃO

Compartilhar informação tornou-se crucial para o combate ao novo Coronavirus e a linguagem não pode ser uma barreira para a difusão da ciência. No campo do processamento de linguagem natural biomédica (BioNLP em inglês), aplicações de tradução automática tentam criar versões de artigos científicos em várias línguas. Bases científicas como a *Scielo* ou a *Pubmed* passaram a incluir artigos biomédicos em outras línguas que não o inglês. Estudo publicado em fevereiro de 2020 por Soares e Yamashita[1] identifica textos chineses sobre, por exemplo, meios de proteção aos médicos nos primeiros focos da doença, que não estão na base *Pubmed/Medline*, onde a maioria dos textos é em inglês. Segundo os autores, cerca de 80% dos artigos da *Pubmed* dirigem-se a especialistas em Genômica e não são úteis para prevenção, diagnóstico ou tratamento do vírus. No Português, os primeiros artigos científicos referentes ao novo Coronavírus surgiram em março de 2020. Iniciativas de promoção e disseminação de informação no cenário pandêmico do Brasil vieram também de fontes jornalísticas como a *FolhaMed*, voltada aos profissionais da linha de frente no combate à pandemia. Contudo, textos científicos biomédicos no idioma português ainda são escassos, se comparados ao grande volume produzido em língua inglesa. Estudos baseados na produção científica biomédica em português são raros. Há apenas alguns trabalhos, como o de Peters et al., baseado em narrativas clínicas contidas em sumários de alta hospitalar[1], ou como o de Soares e Krallinger[2], que compilaram resumos de artigos biomédicos em três idiomas (Inglês, Espanhol, Português).

O objetivo deste artigo é apresentar os primeiros esforços para a elaboração de uma coleção (*corpus*) de textos biomédicos em português com foco no SARS-CoV-2. A construção de um *corpus* depende do propósito a que se destina e a escolha dos materiais depende dos dados que se quer obter. Um *corpus* com foco no Covid-19 pode servir a várias finalidades, por exemplo, mapear correlações entre comorbidades e quadro de evolução da doença. O *corpus* proposto nesta pesquisa tem como foco um estudo do léxico biomédico para a detecção automática de expressões de difícil compreensão (complexas) em textos científicos sobre o Covid-19, tendo em vista uma aplicação posterior de simplificação textual que facilite a leitura e o acesso à informação pelo público leigo. O reconhecimento de expressões da linguagem biomédica precede a detecção das expressões complexas e serve também para aplicações de extração de informação destinadas a profissionais no combate da doença. O *corpus* aqui proposto com base em textos em Português é similar ao *corpus* de documentos usado por Wang[3] para treinar algoritmos de aprendizado de máquina em tarefas de BioNLP como, por exemplo, o reconhecimento da complexidade de conteúdos biomédicos. As próximas seções deste artigo apresentam o Referencial teórico, a Metodologia, Resultados e discussão e Conclusão.

REFERENCIAL TEÓRICO

Elaboração de corpus

A elaboração de um *corpus* não é tarefa trivial, sua concepção se dá por meio de critérios estabelecidos previamente e que contemplem aspectos relacionados à representatividade de um idioma ou variedade linguística[4], que considerem também a autenticidade e a naturalidade dos textos que irão integrá-lo e sua adequação ao objeto da análise, além de estar em formato processável por meio eletrônico[5].Conforme definição proposta por Sardinha[4], a Linguística de *Corpus*

> *Ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.*

Conforme define Hasan[6] "'Para serem adequados, os *corpora* devem ser afinados com os objetivos da análise". Em síntese, um *corpus* é uma fonte a partir da qual se extraem e se descrevem recortes de evidências de elementos linguísticos que poderão ser convertidos em dados de pesquisa. Sua construção se origina de questionamentos cujas respostas virão *a posteriori* , salientam Evers e Finatto[7] "Um *corpus* é, antes de tudo, uma fonte de novas perguntas e de respostas muitas vezes inesperadas."

De acordo com Aluísio e Almeida[8], há três estágios para concepção de um *corpus*. O primeiro refere-se ao projeto do *corpus* e seleção dos textos pertinentes e relevantes para a pesquisa, sendo nesse estágio definido o tipo de *corpus* a ser compilado. O segundo estágio refere-se à compilação dos arquivos de textos selecionados no estágio anterior, podendo os arquivos ser extraídos de forma manual, automática ou semiautomática. E o terceiro estágio dedica-se à nomeação dos arquivos. Após convertê-los para o formato de texto simples, compatível com muitas ferramentas de processamento de linguagem natural, escolhem-se nomes para cada arquivo seguindo um padrão que permita recuperar depois os dados armazenados[8].

Elaboração de corpus para o domínio biomédico

No campo biomédico, tarefas como a mineração de textos requerem estágios de pré-processamento. Neves e Leser[9] lembram que abordagens de processamento de linguagem natural empregando *corpora* anotados permitem caracterizar alguns *corpora* como referência (padrão ouro). Eles são importantes para tarefas de aprendizado supervisionado ou modelos de classificação em fases de teste ou treinamento, podendo se concentrar tanto no nível de documento quanto no nível de texto. Entretanto, essa tarefa nem sempre é simples, dado que, por exemplo, uma sigla no domínio biomédico pode conter múltiplos significados[10].

Estudo conduzido por Angus et al.[11] descreveu a construção do primeiro *corpus* semanticamente anotado de registros clínicos de pacientes com câncer, para uso no desenvolvimento e avaliação de sistemas de extração de informação com foco no domínio biomédico. Denominado *CLEF*, este *corpus* foi criado com o objetivo de reconhecer automaticamente as entidades clínicas nomeadas (mencionadas nos textos) e localizá-las temporalmente. Por exemplo, para um paciente que sofreu

dois infartos, atribuem-se entidades clínicas a tratamentos e medicamentos administrados em cada uma das intercorrências. Para o *corpus CLEF* foram selecionados cerca de 20.000 registros clínicos submetidos a estratificação.

O desenvolvimento do *corpus CLEF* compreendeu cinco etapas: seleção dos textos que comporiam o *corpus,* criação de uma amostra estratificada decomposta por três grupos tipológicos (narrativas clínicas, relatórios histopatológicos, relatórios de imagem), treinamento dos profissionais que participaram do processo de anotação e, por fim, desenvolvimento de métricas comparativas para avaliar a consistência da anotação entre os anotadores[11]. A etapa da estratificação nos três grupos contribuiu para o desenvolvimento de diretrizes para o processo de anotação do *corpus* e para mitigar possíveis inconsistências no processo de anotação das entidades nomeadas. A anotação de um *corpus* é uma tarefa custosa, consome muito tempo e requer grande competência por parte dos anotadores. O *corpus CLEF* foi tratado com os procedimentos usuais de anotação de *corpora*. Os textos foram agrupados em lotes, sendo cada lote anotado por dois anotadores distintos. Para uma anotação aceitável, o trabalho dos dois anotadores deveria apresentar concordância superior a 66%, caso contrário a anotação seria revisada por um terceiro anotador[11].

Um outro *corpus* também submetido a método semelhante foi o *GREC*, constituído de 240 resumos de artigos científicos da base *Medline*. Foi semanticamente anotado com informações relacionadas a eventos biológicos, conforme orientação dos autores Thompson et al.[12], que recomendaram a identificação, para cada evento, de todos os argumentos estruturalmente relacionados. Experimentos iniciais baseados no *corpus GREC* mostraram que o tipo de anotação semântica adotada se revelou produtiva no reconhecimento de entidades nomeadas e outras tarefas de rotulagem de função semântica.

Para o idioma português, o estudo de Peters et al[1] já citado na seção Introdução baseou-se em textos clínicos identificados como sumários de alta hospitalar, condensando informações relativas ao tratamento de pacientes. Por meio de técnicas de processamento de linguagem natural, foram viabilizados a recuperação de informações e o cruzamento de dados médicos. Este *corpus* foi etiquetado automaticamente com auxílio do sistema M*ALT (Morphosaurus Active Learning Tool),* alcançando no final do estudo 91,5% de exatidão.

O idioma português também foi contemplado na construção de um *corpus* biomédico paralelo. Um *corpus* paralelo baseia-se em textos de pelo menos duas línguas. Trabalho recente divulgado em 2019 por Soares e Krallinger[2] propôs-se a elaborar e avaliar um *corpus* paralelo de resumos científicos do domínio biomédico contendo fragmentos de textos em 3 pares de idiomas (Inglês/ Português, Inglês/Espanhol e Inglês/ Espanhol/ Português). Os textos foram extraídos da base de dados BVS (Biblioteca Virtual de Saúde) e a construção do *corpus* baseou-se no alinhamento manual e automático das sentenças dos conjuntos de textos. Esta etapa contou com a ajuda do score *BLUE*, ferramenta que avalia a adequação de textos traduzidos automaticamente. Obteve-se uma média de 96% de acertos, isto é, de sentenças alinhadas corretamente, e apenas 2% alinhadas parcialmente[2].

Um outro *corpus* paralelo é o *CLEAR*, construído por Cardon e Grabar[13] e utilizado para simplificação automática de textos do domínio biomédico. Três conjuntos de textos foram submetidos a comparação. Cerca de 238 pares de sentenças foram anotados por dois anotadores distintos para um posterior treinamento de modelo de alinhamento automático de sentenças. Como resultado, 98% das sentenças foram alinhadas corretamente[13].

A especificidade da construção de *corpora* biomédicos está relacionada ao domínio de conhecimento expresso pelo vocabulário próprio (léxico biomédico) e às relações semânticas que se estabelecem entre as entidades do universo de discurso em que a área se define. Como visto, métodos e técnicas de construção de *corpora* não se distinguem em função do domínio, mas pelo propósito da pesquisa. A seguir, a construção de um *corpus* biomédico sobre o Coronavírus em Português.

MÉTODOS

O acesso à ciência e à sua linguagem é a motivação geral que orientou a elaboração do *corpus* aqui proposto. A motivação específica é o acesso à informação sobre o Covid-19 em Português, através de registros cuja complexidade lexical possa ser reconhecida e convertida para formulações mais simples.

Antes de iniciar a elaboração do *corpus* denominado nesse trabalho como *Corpus Covid-19 UFABC*, efetuou-se uma pesquisa em 3 bases de dados científicas distintas com intuito de investigar se já havia algum *corpus* elaborado que compreendia o escopo e o mesmo propósito deste artigo. Para tanto selecionou-se as bases de dados Biblioteca Virtual de Saúde (BVS), Science Direct e Pubmed, as consultas foram realizadas no dia 28 de agosto de 2020. Adotou-se a seguinte string de pesquisa: "corpora in biomedical domains" OR ("corpus in biomedical domains") OR ("biomedical text mining") AND ("natural language processing"). Obteve-se os seguintes resultados: Biblioteca Virtual de Saúde (BVS) 67, Science Direct 102, Pubmed 218, totalizando 387 artigos retornados. A busca não trouxe nenhum artigo em cujo título constassem informações compatíveis com *corpus* ou *corpora* para o domínio biomédico no idioma português sobre o novo Coronavírus. Iniciou-se então a elaboração do *Corpus Covid-19 UFABC*, conforme as etapas descritas a seguir.

A primeira etapa refere-se à extração dos textos e para esse fim, selecionou-se apenas a base de dados científica *Pubmed*, seus respectivos textos foram extraídos no dia 15 de setembro de 2020. Adotou-se a seguinte string de pesquisa: ("coronavírus") OR ("coronavirus") OR ("corona virus") OR ("COVID-19"[MeSH Terms]) OR ("sarscov2") OR ("sars-cov-2") OR ("2019-ncov") OR ("sarscov") AND (portuguese[Filter]) AND (2020:2020[pdat]). Ao todo foram retornados 257 textos referentes aos meses de março a setembro do ano de

2020. Do total retornado, 3 artigos científicos foram excluídos por não apresentarem versão em língua portuguesa, portanto o conjunto de textos extraídos compõe-se de 254 textos científicos completos.

A etapa seguinte à extração, refere-se ao armazenamento e indexação dos textos. Cada arquivo foi armazenado nas extensões *.pdf* e *.txt*. Reforça-se que em virtude da relevância e abrangência do objeto de pesquisa acerca do Covid-19 no contexto atual, os textos que compõem a base de dados *Pubmed* estão disponíveis gratuitamente para leitura. Para a indexação dos arquivos, cada arquivo extraído foi organizado em uma planilha eletrônica conforme descrito na tabela 1.

A terceira etapa consistiu na limpeza e padronização dos dados, retirando dos arquivos de texto as referências bibliográficas, informações autorais, tabelas, figuras e as versões em língua estrangeira dos resumos. Com auxílio da biblioteca de expressões regulares (*RE- Regular expression*) da linguagem *Python*, efetuou-se a retirada de datas e termos acompanhados de números decimais e *URL* de sites. Já com o auxílio da biblioteca *NLTK* da linguagem de programação *Python*, efetuou-se a aplicação de filtros de normalização à caixa baixa, remoção de *stopwords* e a retirada de palavras em inglês com auxílio do *corpus Word* da mesma biblioteca. Além disso, foi realizada a análise descritiva do *corpus* quanto a suas características gerais, tais como número de *tokens*, número de *types* e densidade lexical (razão entre número de *types* e número de *tokens*). Isto permitiu analisar a distribuição dos *tokens*, formando-se um *ranking* dos *types* em ordem decrescente de frequência. A partir dos dados obtidos, foi ajustada uma curva segundo a lei de Zipf[14-16].

$$f(r) \propto \frac{1}{r^{\alpha}}, \qquad (1)$$

em que $f(r)$ é a frequência do r-ésimo *type* no *ranking* e $\alpha \approx 1$ constante. O ajuste foi feito com o uso do método dos mínimos quadrados para a expressão.

$$\log(f(r)) = k - \alpha \log(r), \qquad (2)$$

em que k é uma constante. No caso, para se evitar o peso excessivo dos *types* com baixa frequência na regressão, para cada frequência $f(r)$ foi considerada a média das posições no *ranking* dos *types* com a frequência correspondente como sendo o valor de r

A quarta etapa dedicou-se à categorização manual dos arquivos por especialidade clínica e por gênero textual. Para identificar a especialidade clínica, o nome da revista científica serviu como referência e, quando o nome não indicasse a especialidade, optou-se por analisar o título do artigo ou as palavras-chave do texto. A identificação do gênero textual adveio de indicações da própria base de dados.

RESULTADOS E DISCUSSÃO

Por meio da análise descritiva realizada no *Corpus Covid-19 UFABC,* identificou-se 27.816 *types,* 311.201 *tokens* e 8,98 de densidade lexical. Na Figura 1, cada cruz representa um *type* e a linha vermelha representa a curva de melhor ajuste com uso da equação(2). Uma vez que os gráficos possuem eixos com escala logarítmica, a curva ajustada aparece como uma reta. No caso, os valores para o melhor ajuste foram k=4,83 e α=1,01 , seguindo assim, uma distribuição Zipfiana. Em uma análise mais detalhada, na Figura 1(a) é possível observar que os dados seguem a curva da Lei de Zipf para *types* com altas e médias frequências. Por outro lado, *types* com frequência abaixo de 20 a 30 ocorrências possuem frequência menor que a prevista pela Lei de Zipf, como mostrado na imagem ampliada na Figura 1(b). Como observado em[15], esse é um comportamento típico para um *corpus* de tamanho reduzido, lembrando que a Lei de Zipf foi descoberta pela análise da obra *Ulysses* de James Joice, cujo número de *types* (29.899) e *tokens* (260.430) é semelhante ao *Corpus Covid-19 UFABC* do atual trabalho.

As cruzes verdes representam os *types* e a linha vermelha a curva de melhor ajuste referente à Lei de Zipf. A Figura 1 (a) mostra todos os *types* e Figura 1 (b) mostra o detalhe da distribuição para *types* de baixo valor de $f(r)$.

A categorização dos textos do *Corpus Covid-19 UFABC* por especialidade clínica, como mostrado na tabela 2, permitiu observar a distribuição de artigos publicados em língua portuguesa sobre o novo Coronavírus em categorias clínicas específicas.

Note-se que cerca de 21% das publicações científicas que compõem o *Corpus Covid-19 UFABC,* abordam temática relacionada à Saúde Pública e 14% referem-se a textos relacionados à especialidade de Epidemiologia. A discussão destes primeiros resultados pode ser realizada tomando-se como referência o *Corpus CORD-19* disponível em língua inglesa[17]. Enquanto os artigos científicos biomédicos escritos em Português valorizam a dimensão da Saúde Pública e da distribuição de

Tabela 1- Descrição da indexação do armazenamento dos arquivos *Corpus Covid-19* UFABC.

Campo	Descrição
Id artigo	Índice de indexação do artigo na base de dados.
PMID	Código do artigo da base de dados
Titulo	Título do artigo em português
Categoria	Categoria de gênero conforme base científica consultada
Nome do arquivo	Nome do arquivo armazenado localmente
Palavra-chave do artigo	Palavras - chaves, caso o artigo tivesse.
Base de dados extração	Base de dados em que o artigo foi extraído
Especialidade	Especialidade clínica do artigo
Autores do artigo	Nomes dos autores do artigo
Journal publicado	Periódico ou jornal em que o artigo foi publicado
Data de publicação	Data em que o artigo foi publicado na base de dados científica.
DOI	Padrão de identificação de documentos.

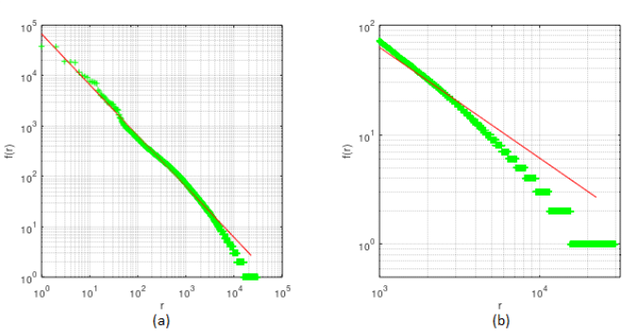

Figura 1- Frequência dos types, *f(r)*, em função da posição no ranking, *r* Corpus Covid-19 UFABC.

Tabela 2- Publicações por especialidade médica *Corpus Covid-19 UFABC.*

Categoria	Quantidade de publicações	Categoria	Quantidade de publicações
Saúde pública	55	Relatório imagens	9
Epidemiologia	36	Anestesiologia	7
Cardiologia	30	Clínica geral	5
Enfermagem	24	Fonoaudiologia	5
Cirurgia	17	Pediatria	4
Nefrologia	14	Saúde primária	3
Terapia intensiva	11	Outros	16
Pneumologia	10	Não classificado	8

Tabela 3- Percentual de publicações do corpus CORD-19[17]

Subcampo de conhecimento	Percentual de publicações
Virologia	42,3%
Imunologia	20,7%
Biologia Molecular	12,7%
Genética	8,0%
Medicina Intensiva	6,7%
Outros	9,6%

Tabela 4- Número de publicações por gênero textual no *Corpus Covid-19 UFABC*

Gênero_Textual	Nº Publicações	Gênero textual	Nº Publicações
Editorial	35	Opinião	6
Carta ao editor	28	Relato de caso	6
Artigo original	26	Revisão	6
Artigo	24	Perspectivas	4
Contribuições da saúde coletiva	17	Aprendendo por imagens	3
Artigo de revisão	14	Comentário	3
Recomendações	11	Ensaio	3
Artigo especial	9	Imagens pneumologia	3
Comunicação breve	8	Posicionamento	3
Nota técnica	8	Reflexão	3
Ponto de vista	8	Outros	20
Artigo de opinião	6		

tratamentos, medicamentos e serviços de saúde aos cidadãos, os *corpora* em língua estrangeira, como, por exemplo, no idioma inglês, parecem tender a abordar temáticas mais focadas nos subcampos de conhecimento da biologia ou da genética, ou seja, mais compatíveis com a pesquisa da natureza do vírus e a cura da doença[17].

O *Corpus CORD-19* segundo citam os autores Wang et al., foi construído com base em textos científicos, compõe-se de 52 mil artigos científicos sobre Covid-19, extraídos das bases científicas *Pubmed, BioRxiv e MedRxiv.* Projetado para facilitar o desenvolvimento de sistemas de mineração, extração e recuperação de informação, conectando a comunidade de aprendizado de máquina com especialistas em domínio biomédico e formuladores de políticas públicas, na corrida para identificar tratamentos eficazes e políticas de gerenciamento para o novo Coronavírus. Contempla anotações em nível de sentença e anotações sensíveis ao contexto de publicações cientificas relacionadas ao novo Coronavírus. No *CORD-19* não há uma distribuição dos artigos por especialidade clínica, como no *corpus* proposto nesse artigo, ao invés disso a tabela 3 mostra o recurso à classificação automática por subcampo de conhecimento, realizada pela *MAG* (Microsoft Academic Graph)[17].

Observa-se que cerca de 42% dos artigos que compõem o *CORD-19*, referem-se a textos relacionados

a Virologia. É possível que parte desses textos sejam de autores lusófonos, considerando, por exemplo, o estudo de Rosselli[18], segundo o qual cerca de 84% dos textos biomédicos brasileiros são publicados em inglês. Então o *CORD-19* pode incluir outros autores que sejam falantes não-nativos do inglês. Em contrapartida, o *Corpus Covid-19 UFABC* proposto nesse artigo, contempla uma categorização por gênero textual, tal como mostrado na tabela 4.

Geralmente, bases de dados cientificas não compreendem a filtragem por gênero textual, podendo as buscas retornar textos de diferentes gêneros. Observa-se que, do total de 254 textos retornados na base científica *Pubmed*, cerca de 13% referem-se a artigos do gênero textual editorial, 11% do tipo carta ao editor e 10% a artigos com teor científico (artigo original). A identificação dos 23 gêneros textuais adveio da própria base em que os artigos foram extraídos, ao passo que no *CORD-19* essa diversidade foi reduzida a 10 categorias. Gêneros textuais diferentes podem estar associados a variações na riqueza vocabular, conforme estudo de Kubat e Milièka[19] e também a diferentes níveis de complexidade textual, como indicou trabalho anterior do grupo de pesquisa em processamento de linguagem natural da Universidade Federal do ABC[20]. Riqueza vocabular e complexidade textual são variáveis relevantes para detectar expressões complexas no domínio biomédico.

CONCLUSÃO

Apresentou-se aqui os primeiros passos para criar um *corpus* regional de informação biomédica em língua portuguesa sobre o Covid-19. A categorização dos 254 textos por especialidade clínica e gênero textual identificou que cerca de 30% são artigos relacionados às especialidades de Saúde pública e Epidemiologia. Tal resultado destaca a prevalência de objetos de estudo voltados a ações de vigilância estratégica que mitiguem a evolução do vírus e colaborem no seu combate. A construção do *corpus* anotado em nível de documento é o esforço inicial de uma pesquisa mais abrangente sobre detecção automática de expressões complexas no domínio biomédico[21]. Trabalhos futuros poderão ampliar as bases científicas aqui analisadas durante o processo de extração de artigos. E poderão expandir a anotação do *corpus* para o nível de texto, classificando os itens lexicais complexos e apontando para sistemas de simplificação de textos no domínio biomédico, tal como proposto por Wang[3]. Pretende-se também implementar um processo automático de atualização de textos do *corpus*, semelhante ao *CORD-19*[17], garantindo a inclusão de novos textos científicos em português sobre tratamento, prevenção e diagnóstico do novo Coronavírus. Um subproduto dessa tarefa seria a criação de um histórico de tendências e pensamentos que prevaleceram ao longo da pandemia. Finalmente, não foge ao horizonte desta pesquisa a opção de explorar um *corpus* paralelo de textos científicos em inglês e sua tradução para o português.

REFERÊNCIAS

1. Soares F, Yamashita GH. On the crucial role of multilingual biomedical databases in epidemic events (SARS-CoV-2 analysis). International Journal of Infectious Diseases. 2020 Julho 01: p. 352-354.
2. Peters AC, Pacheco E, Moro C, Oleynik M. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. In In:XII Congresso Brasileiro de Informática em Saúde; 2010; Porto de Galinhas. p. 1-5.
3. Soares F, Krallinger M. BVS Corpus/ : A Multilingual Parallel Corpus of Biomedical Scientific Texts. CoRR,arXiv:1905.01712. 2019 Maio 05: p. 1-8.
4. Wang Y. Automatic Recognition of Text Difficulty from Consumers Health Information. In In:19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06); 2006; Salt Lake City: IEEE. p. 131-136.
5. Sardinha TB. Lingüística de Corpus Histórico e Problemática. Delta - Documentação de Estudos em Linguística Teórica e Aplicada. 2000; vol.16(2).
6. Sinclair J. Developing Linguistic Corpora. [Online].; 2004 [cited 2020 Maio 1. Available from: HYPERLINK "http://users.ox.ac.uk/~martinw/dlc/chapter1.htm" http://users.ox.ac.uk/~martinw/dlc/chapter1.htm .
7. R.Hasan. Rationality in everyday talk: From process to system. In Svartvik J. Directions in Corpus Linguistics. Berlin: Mounton de Gruyter; 1992. p. 257-307.
8. Evers A, Finatto MJB. Linguística de Corpus, Léxico-Estatística Textual e Processamento de Linguagem Natural: perspectiva para estudos de vocabulário em produções textuais. Revista GTlex. 2016 Janeiro; 1(2): p. 271-295.
9. Aluísio SM, Almeida GMDB. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. Calidoscópio. 2006 Setembro/Dezembro: p. p.155-177.
10. Neves M, Leser U. A survey on annotation tools for the biomedical literature. Briefings in Bioinformatics. 2014 março; 15(2): p. 327-340.
11. Goulart RR, LIma VSD. O Contexto no Reconhecimento de Entidades Nomeadas em Textos de Biomedicina. In In:7th Brazilian Symposium in Information and Human Language Technology; 2009; São Carlos. p. 9.
12. Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics. 2009 janeiro 23: p. 950-966.
13. Thompson P, Iqbal SA, McNaught J, Ananiadou S. Construction of an annotated corpus to support biomedical information extraction. BMC Bioinformatics. 2009 Outubro 23: p. 349.
14. Cardon R, Grabar N. Parallel Sentence Alignment from Biomedical Comparable Corpora. In Pape-Haugaard LB, Lovis C, Madsen IC, Weber P, Nielsen PH, Scott P, editors. In:Proceeding of MIE 2020; 2020 Junho; Berlin: IOS Press. p. 362-366.
15. Zipf GK. Human Behavior and the Principle of Least Effort. 1949.
16. Ha LQ, Sicilia-Garcia EI, Ming J, Smith FJ. Extension of Zipf's Law to Words and Phrases. In COLING 2002: The 19th International Conference on Computational Linguistics; 2002. p. 1-6.
17. Araújo LCd, Sansão JPH, Yehia HC. Influência da lei de Zipf na escolha de senhas. Revista Brasileira de Ensino de Física. 2016 Abril 05; 38(1): p. 1-14.
18. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J. CORD-19: The Covid-19 Open Research Dataset. [Preprint] ArXiv. 2020;arXiv:2004.10706v2. 2020 Abril: p. 1-10.
19. Rosselli D. The language of biomedical sciences. The Lancet. 2016 April; 387.
20. Kubát M, Milièka J. Vocabulary Richness Measure in Genres. Journal of Quantitative Linguistics. 2013 Nov 1: p. 339-349.
21. Leite JDS, Takahata AK, Steinberger- Elias M. Criação e análise de amostras de corpora em Português Brasileiro para detecção automática de expressões complexas em textos sobre covid-19.In: XXVII Congresso Brasileiro de Engenharia Biomédica (CBEB2020). 2020. Vitória,ES: p. 721-722.