Lexicalidade biomédica e sua mensuração em um corpus sobre COVID-19 em língua portuguesa

Karhyne S. Padilha de Assis[1], Camila das Merces Silva[1], Janaína da Silva Leite[1], Wellington Araujo Nogueira[1], Kenji Nose Filho[1], André K. Takahata[1], Margarethe Steinberger-Elias[1].

[1]Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC (UFABC) - Santo André - SP - Brasil

{karhyne.assis, camila.merces, janaina.leite, wellington.araujo, kenjinose, andre.t, margarethe.elias}@ufabc.edu.br

Abstract *We analyzed the biomedical lexicon of a corpus of Portuguese texts on Covid-19 extracted from the Pubmed database with the help of classic measures like lexical density and lexical diversity. Preliminary results could not show the lexical distribution in different texts genres and clinical specialties present in the corpus. Based on the concept of "biomedical lexicality", a new indicator, Lex-BioMed, was proposed and tested demonstrating good performance.*

Resumo. *Analisamos o léxico biomédico de um corpus de textos em língua portuguesa da base Pubmed sobre a Covid-19. A adoção inicial de medidas clássicas de densidade e diversidade lexical não foi capaz de evidenciar a distribuição lexical nos diferentes gêneros e especialidades clínicas de que se compõe o corpus. Com base no conceito de "lexicalidade biomédica", foi proposto e testado um novo indicador, o Lex-BioMed, com bons resultados.*

1. Introdução e motivação

A COVID-19, inicialmente identificada em Wuhan na China em dezembro de 2019, teve aumento de casos em diversos países e continentes, até ser decretada pandêmica pela Organização Mundial da Saúde (OMS) em 11 de março de 2020 [Cucinotta e Vanelli, 2020]. A partir desta data, textos biomédicos publicados na base Pubmed foram filtrados por Leite et al. (2020) para a criação de um corpus em português a respeito da COVID-19. Com o objetivo de prover dados para uma pesquisa dos processos de simplificação da linguagem da Saúde, o corpus COVID-19 visa atender a demanda informacional de um público brasileiro leigo sobre a pandemia. Compõe-se de 254 textos do período inicial da pandemia, compreendido entre março e setembro de 2020. O corpus é heterogêneo, distribuído entre 23 gêneros textuais e 16 especialidades clínicas. Sendo o nível lexical o de complexidade mais visível no corpus, optou-se no presente trabalho pela análise do léxico biomédico e procedeu-se à contagem de *types* e *tokens* e à mensuração da diversidade ou riqueza lexical (DiL) dada pela razão entre número de *type*s e *tokens* (TTR, *type-token ratio*) e à densidade lexical (DeL) dada pela razão entre o número de palavras de conteúdo semântico (nomes, adjetivos e verbos) e o número total de palavras do corpus [Santos et al., 2018].

O nosso problema inicial de pesquisa foi como identificar automaticamente termos biomédicos de difícil compreensão e convertê-los em expressões acessíveis. Tomou-se como ponto de partida a hipótese de que os índices de densidade e de diversidade lexical seriam capazes de apontar os gêneros de maior complexidade, isto é, onde haveria maior concentração de *types*. Métodos clássicos como Flesch-Kincaid, partições morfológicas e outros foram temporariamente deixados de lado, colocando-se o texto especializado como

foco da pesquisa. “Tendo-se o texto como foco, deixa de fazer sentido que se continue estudando somente os termos, de forma que se passa a englobar os modos de dizer peculiares de cada área de conhecimento” [Finatto, 2004a, p. 348].

Linguagens de especialidade, como é o caso da biomédica, tem um comportamento diferenciado nos estudos de corpora: “(...) a partir da observação da linguagem especializada em corpora que se percebe mais francamente como a observação de termos é somente um pequeno passo na observação do texto especializado” [Perna, Delgado e Finatto, 2010 p.138]. Estudo de Zilio (2009) sobre textos científicos de Cardiologia e Radiologia compara a distribuição lexical entre as duas especialidades e atribui as convergências a fatores textuais: “(...) se não houvesse no corpus de Radiologia um artigo que se ocupasse do coração, somente dois dos compostos estudados seriam comuns aos dois corpora” (p.142).

A observação inicial sobre o comportamento lexical das linguagens de especialidades nos textos do corpus e a indefinição das medidas de densidade lexical nesse contexto levou a busca de um novo indicador da lexicalidade no corpus. Propomos aqui o conceito de “lexicalidade biomédica” ou “densidade lexical biomédica” para identificar com maior segurança o espaço lexical que é das especialidades biomédicas e diferenciá-lo de um léxico fronteiriço revelado em gêneros menos técnicos. O problema de pesquisa foi revisto, tornando-se imperativo reconhecer no corpus um repertório de termos biomédicos, de modo a identificar sua distribuição no corpus sobre COVID-19. Investigamos o novo indicador de lexicalidade para verificar se seria capaz de cumprir uma função distribucional que as simples densidade e diversidade lexicais não lograram alcançar.

2. Materiais e Métodos

Como já descrito na Seção 1, seguindo [Leite et al., 2020], foi obtido um corpus de textos em língua portuguesa a respeito da COVID-19, a partir de textos indexados na base científica *Pubmed* do período entre março e setembro de 2020, totalizando 254 textos. Os textos foram categorizados manualmente conforme gêneros, utilizando informações fornecidas pela base, e conforme especialidades clínicas, de acordo com o nome da revista, título do artigo ou palavras-chave do texto. As distribuições dos textos nas classes obtidas se encontram nas Tabelas 2 e 3. Após a fase preliminar de filtragem, limpeza e compilação, buscou-se nomear os arquivos já em formato .txt, tokenizar e fazer a anotação morfossintática das classes de palavras (PoS, *parts of speech*) com o uso do analisador e corretor gramatical CoGrOO [Silva, 2013]. O resultado para as classes de palavras de conteúdo semântico (nome, verbo e adjetivo) se encontra na Tabela 1.

Com finalidade de caracterizar e descrever o léxico biomédico do corpus, foram calculados os índices da DeL e da DiL. A DeL descreve a proporção de palavras de conteúdo pelo número total de palavras em cada texto e de acordo com [Ure, 1971] [Johansson, 2008], o valor resultante desse indicador expressa a concentração de conteúdo lexical presente em um determinado texto. Um texto com alta densidade lexical contém mais palavras de conteúdo, enquanto um texto com baixa densidade lexical é composto por palavras funcionais (preposições, conjunções, artigos, pronomes, verbos modais e auxiliares). Já a DiL pode ser descrita pela TTR. Entretanto, a TTR tende a possuir valor menor em textos com maior número de *tokens* ou a possuir valor maior em textos com menor número de *tokens*, fazendo com que o seu uso torne enviesada a comparação entre dois textos ou corpora com número de *tokens* diferentes [Johansson, 2008]. Para mitigar esse efeito, utilizamos também o vocabulário teórico (VT) ou *theoretical vocabulary* [Broeder, Extra & van Hout 1986], em que a proporção de *types* é calculada para subconjuntos de tamanho fixo com N *tokens*, no nosso caso $N=100$, amostrados aleatoriamente. Em nosso trabalho,

mostramos resultados obtidos a partir da realização de $S=100$ amostragens aleatórias distintas para cada grupo de texto.

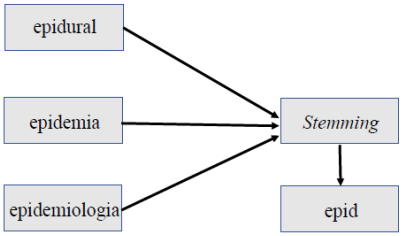


Figura 1: *Stemming*

Tabela 1: Resultados para as PoS por subcorpora

Especialidade	Nº de publicações	Total de *tokens*	n	adj	v	n-adj	n (%)	adj (%)	v (%)
Saúde Pública	55	83028	40997	11376	17707	117	49,38	15,03	21,33
Epidemiologia	36	50810	24542	7381	10267	88	48,30	14,53	20,21
Cardiologia	30	37891	16429	6690	7981	65	46,00	17,66	21,06
Enfermagem	24	46410	23265	6476	10083	68	50,13	13,09	21,73
Cirurgia	17	25098	10693	3548	5138	22	42,60	14,14	20,47
Outros	16	21446	11656	3827	5498	49	54,35	17,84	25,64
Nefrologia	14	14930	6882	2196	3122	21	46,10	14,71	20,91
Terapia Intensiva	11	15534	7424	2209	3219	20	47,79	14,22	20,72
Pneumologia	10	4791	2289	823	1000	3	47,78	17,18	20,87
Relatório Imagens	9	3717	1939	755	757	3	46,79	20,31	20,37
Não Classificados	8	6413	3062	876	1368	10	47,75	13,66	21,33
Anestesiologia	7	6672	3073	1071	1520	4	46,06	16,05	22,78
Clínica Geral	5	2733	1280	399	624	8	46,83	14,60	22,83
Fonoaudiologia	5	4131	2094	604	941	3	50,69	14,62	22,78
Pediatria	4	5679	2569	919	1158	7	45,24	17,94	20,39
Saúde Primária	3	2177	1192	330	386	3	50,16	15,16	17,73

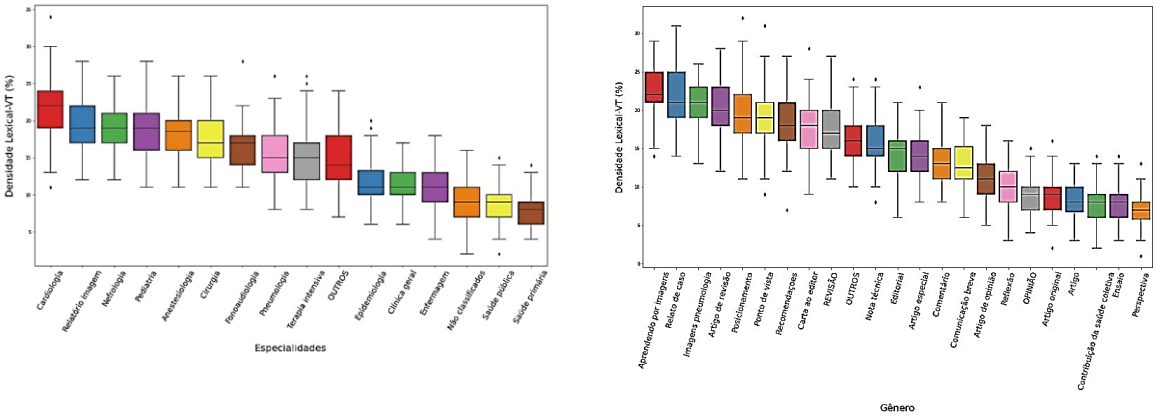
Para a identificação das palavras de domínio biomédico em toda extensão do corpus foi usado, como primeiro passo, o *stemming*, sendo empregado para implementação o RSLP Stemmer [Orengo e Huyck 2001] que tem como objetivo remover os sufixos, para reduzir o número de palavras, conforme ilustrado na Figura 1. Observamos que originalmente o corpus possui 31.123 *types*. Após o *stemming* esse número se reduziu para 11.513 raízes distintas. Após a obtenção das raízes, foi criada uma lista de pares do tipo (*raiz*, *type*), em que o segundo elemento consiste em um type escolhido ao acaso dentre as palavras do corpus que podem ser formadas com o uso da respectiva raiz. Assim, para a raiz "epid" na Figura 1, o par formado poderia ser ("epid", "epidural"). Os *types* de todos os 11.513 pares foram analisados por três pesquisadores de processamento de linguagem natural (PLN) em textos biomédicos, sendo o time formado por duas mulheres e um homem, em que dois são docentes universitários (com formação em engenharia elétrica e linguística) e uma é estudante de mestrado (com formação em análise e desenvolvimento de sistemas), com idades entre 34 e 69 anos. Procedeu-se à análise de modo a comparar a identificação de palavras com uso prevalente no domínio biomédico, resultando em 2.258 raízes associadas a palavras no domínio biomédico.

A partir dessa identificação de raízes biomédicas, passamos para a etapa de análise de cada subcorpus formado por textos agrupados por gênero ou especialidade com cálculo do DeL, DiL-TTR e DiL-VT. Constatando que tais índices não foram capazes de mostrar a distribuição lexical por gênero e por especialidade, criamos um novo indicador que pudesse medir a lexicalidade biomédica (Lex-BioMed), que consiste na proporção do número de *types* que possuem raízes biomédicas em conjuntos de tamanho fixo (no nosso caso $N=100$)

obtidos aleatoriamente no grupo de textos em análise. Assim como para a DiL-VT, mostramos resultados para Lex-BioMed obtidos a partir da realização de $S=100$ amostragens aleatórias distintas.

3. Resultados

Nas Tabelas 2 e 3 são apresentados os resultados para os subcorpora formados a partir da categorização dos textos por especialidades clínicas e por gênero textual, respectivamente. Ao analisarmos o DiL-TTR, observamos que o principal fator de influência para essa métrica foi o número de *tokens*. A influência dos números de *tokens* na medida não se observou com o DeL e com o DiL-VT, mas, como mostrado nas Tabelas 2 e 3, os valores obtidos revelaram-se similares entre os subcorpora analisados, não permitindo uma discriminação clara a respeito do conteúdo biomédico presente nos respectivos subconjuntos de textos. Contrastando com as métricas mais tradicionais, ao analisarmos a Lex-BioMed, é possível observar que se torna possível diferenciar subcorpora ou grupos de subcorpora diferentes a partir dessa métrica, o que também é corroborado pela Figura 2. Por exemplo, para especialidades, as categorias com maior Lex-BioMed foram "Cardiologia" e "Relatório de Imagens" com 25,29% e 24,57% respectivamente, enquanto as categorias de menor Lex-BioMed foram "Saúde Pública" e "Saúde Primária", com, respectivamente, 9,17% e 8,54%. Na categorização por gêneros, os de maior Lex-BioMed foram "Aprendendo por Imagens" e "Imagens Pneumologia" com proporção de 29,84% e 27,35% respectivamente. Na Figura 2, os *box-plot* indicam que a distribuição da proporção de palavras com raízes biomédicas em conjuntos de $N=100$ palavras escolhidas aleatoriamente é diferente entre as categorias, principalmente dentre as com maior e menor valor de Lex-BioMed. Ao ordenar as subcategorias em ordem decrescente de Lex-BioMed, como na Tabela 2 e na Figura 2(a), observamos que podemos agrupar as especialidades em dois subconjuntos. O primeiro composto por "Cardiologia", "Relatório Imagens", "Nefrologia", "Pediatria", "Anestesiologia", "Cirurgia", "Fonoaudiologia", "Pneumologia" e "Terapia Intensiva"; e o segundo formado por "Outros", "Epidemiologia", "Clínica Geral", "Enfermagem", "Não Classificados", "Saúde Pública" e "Saúde Primária". Nota-se que, em linhas gerais, no primeiro grupo encontram-se especialidades médicas mais específicas, como Cardiologia, Radiologia e Nefrologia, enquanto, no segundo grupo, encontram-se categorias indefinidas, como "Outros" e "Não Classificados", categorias não médicas, como "Enfermagem" e "Saúde Pública", ou de caráter mais amplo, como "Clínica Geral".



(a) (b)

Figura 2: *Box-plot* descrevendo a distribuição da proporção de raízes biomédicas em S=100 amostras aleatórias obtidas para cálculo de *Lex-BioMed* para cada (a) especialidade e (b) gênero textual.

Tabela 2. Resultados das características por especialidade

Especialidade	Nº publicações	Total de *tokens*	Total de *types*	*Tokens* biomédicos	*Types* biomédicos	DeL (%)	DiL-TTR (%)	DiL-VT (%)	Lex-BioMed (%)
Cardiologia	30	37891	8063	9311	912	43,24	21,28	76,28	25,29
Relatório Imagens	9	3717	1414	940	186	40,89	38,04	76,24	24,57
Nefrologia	14	14930	4568	3409	461	42,11	30,60	75,09	22,83
Pediatria	4	5679	2696	1200	390	39,67	47,47	76,31	22,54
Anestesiologia	7	6672	2765	1504	319	40,68	41,44	74,83	21,13
Cirurgia	17	25098	6320	5177	554	42,75	25,18	75,58	20,63
Fonoaudiologia	5	4131	1975	753	203	40,64	47,81	74,27	19,29
Pneumologia	10	4791	2168	924	208	40,94	45,25	74,95	18,23
Terapia Intensiva	11	15534	4147	2563	420	41,53	26,70	75,32	16,50
Outros	16	21446	6103	3485	448	42,41	28,46	75,95	16,25
Epidemiologia	36	50810	8813	6605	448	44,50	17,35	74,24	13,00
Clínica Geral	5	2733	1495	327	100	40,35	54,70	76,11	11,96
Enfermagem	24	46410	9033	5479	416	42,92	19,46	75,84	11,81
Não Classificados	8	6413	2597	660	132	40,78	40,50	75,64	10,29
Saúde Pública	55	83028	14091	7611	534	48,02	16,97	76,20	9,17
Saúde Primária	3	2177	1176	186	61	39,03	54,02	74,30	8,54

Tabela 3. Resultados das características por gênero

Gênero	Total publicações	Total de *tokens*	Total de *types*	*Tokens* biomédicos	*Types* biomédicos	DeL (%)	DiL-TTR (%)	DiL-VT (%)	Lex- BioMed (%)
Aprendendo por Imagem	3	868	484	259	98	45,28	55,76	74,20	29,84
Imagens Pneumologia	35	21521	6846	3652	514	43,90	31,81	74,79	27,35
Relato de Caso	3	8898	3820	972	179	41,88	42,93	76,55	24,85
Artigo de Revisão	14	33093	7350	8121	849	41,80	22,21	74,88	24,54
Recomendações	17	17876	5440	1955	249	42,93	30,43	74,65	22,82
Posicionamento	24	51996	9887	5716	347	42,71	19,01	75,29	22,32
Ponto de Vista	6	7594	3185	911	168	43,29	41,94	76,18	22,14
Cartas ao Editor	8	7607	2773	1684	325	41,38	36,45	75,99	21,13
Revisão	4	4437	2139	419	80	41,60	48,21	76,12	20,16
Nota Técnica	3	1888	992	309	89	44,16	52,54	75,79	19,29
Outros	6	5027	2195	751	124	48,52	43,66	75,50	18,11
Artigo Especial	11	11081	3727	2529	387	39,90	33,63	75,77	17,26
Editorial	20	23653	6410	4284	658	42,67	27,10	76,58	16,97
Comentário	28	11612	4330	2454	448	42,89	37,29	75,30	16,37
Comunicação Breve	6	13652	4125	2752	459	38,49	30,22	73,91	15,95
Artigo de Opinião	6	6854	2873	1703	466	41,39	41,92	73,46	14,94
Reflexão	26	51975	9074	5683	354	41,90	17,46	73,63	14,87
Opinião	8	8556	3037	1365	208	40,51	35,50	75,80	12,00
Artigo	3	479	329	131	62	41,85	68,68	74,11	10,99
Contribuições da Saúde Coletiva	8	9855	3497	1901	322	43,44	35,48	74,00	10,94
Artigo Original	3	6755	2630	1508	360	40,84	38,93	74,88	10,93
Ensaio	9	16900	4936	2917	449	43,17	29,21	75,44	10,92
Perspectiva	3	4863	2132	723	120	41,77	43,84	75,99	9,44

A partir dos resultados da Tabela 3 e da Figura 2(b), pode-se observar também que o léxico de diferentes gêneros pode ser diferenciado com uso da Lex-BioMed, em especial gêneros como “Aprendendo por Imagem” e “Perspectiva”. Na Figura 3 é mostrada a distribuição conjunta entre os gêneros e as especialidades clínicas, onde as linhas e as colunas estão ordenadas por ordem decrescente de Lex-BioMed. Pode-se observar que, *grosso modo*, gêneros com maior valor de Lex-BioMed são constituídos de textos de especialidades que também possuem maior valor do índice, como “Aprendendo por Imagem”, constituído em sua totalidade por textos da especialidade “Relatório Imagens” e “Relato de Caso”, constituído majoritariamente por textos de “Cardiologia”, ambas as especialidades com alta Lex-BioMed. Mesmo gêneros que sugerem textos mais opinativos, como “Posicionamento” e “Ponto de Vista”, possuem relativa alta Lex-BioMed, por serem constituídos exclusivamente por textos de Cardiologia. Além disso, gêneros com baixa Lex-BioMed estão, em geral, associados às especialidades com o mesmo comportamento. Por exemplo, gêneros como “Artigo Original” e “Artigo” possuem maior prevalência de especialidades como “Epidemiologia”, “Enfermagem” e “Saúde Pública”. Os gêneros que possuem os menores valores de Lex-BioMed são “Ensaio” e “Perspectivas”, com textos que descrevem o contexto da pandemia, ao invés de assuntos técnicos do ponto de vista clínico. É o caso

dos textos “COVID-19, as *fake news* e o sono da razão comunicativa gerando monstros: a narrativa dos riscos e os riscos das narrativas” [Vasconcellos-Silva e Castiel, 2020] e “Da Tuberculose ao COVID-19: Legitimidade Jurídico-Constitucional do Isolamento/Tratamento Compulsivo por Doenças Contagiosas em Portugal” [Peixoto et al., 2020]. Os gêneros “Ensaio” e “Perspectivas” possuem maior prevalência de especialidades associadas à baixa Lex-BioMed. A Figura 3 ainda indica uma justificativa para a “Pediatria”, especialidade mais geral, possuir valor maior de Lex-BioMed, do que especialidades mais específicas, como, por exemplo, “Pneumologia”. Observa-se nesses casos participação majoritária de textos de gêneros como “Relato de Caso”, “Artigo de Revisão” e “Imagens Pneumologia”, que possuem alta Lex-BioMed como em “Pediatria”, enquanto em “Pneumologia” se observa a predominância de textos dos gêneros “Cartas ao Editor” e “Editorial”, de menor índice.

Gênero \ Especialidade	Cardiologia (%)	Relatório Imagens (%)	Nefrologia (%)	Pediatria (%)	Anestesiologia	Cirurgia (%)	Fonoaudiologia	Pneumologia (%)	Terapia Intensiva (%)	Outros (%)	Epidemiologia (%)	Clínica Geral (%)	Enfermagem (%)	Não Classificados (%)	Saúde Pública (%)	Saúde Primária (%)
Aprendendo por Imagem		3														
Relato de Caso	4			1			1									
Artigo de Revisão	4			1		5			1		2		1			
Imagens Pneumologia	4			1		5										
Posicionamento	3															
Ponto de Vista	8															
Recomendações			11													
Cartas ao Editor	3				3	2	4	5		4		3		2	1	1
Revisão		1	1							2			1		1	
Outros	5		1	1	2	1		1		2	3		2		2	
Nota Técnica						8										
Editorial	6	1	1	2	2			4	1	3	6	1	3	1	4	
Artigo Especial									3		4			1		1
Comentário									3							
Comunicação Breve	1	1									2				4	
Artigo de Opinião											6					
Reflexão													3			
Opinião											1			1	3	1
Artigo Original						2			3	1	4		12		4	
Artigo										3	5			2	14	
Contribuições da Saúde Coletiva															17	
Ensaio															3	
Perspectivas												1		1	2	

Figura 3: Distribuição conjunta dos textos do corpus por especialidades e gêneros

4. Conclusões

Mostramos neste trabalho que as medidas clássicas de diversidade e densidade lexical não são adequadas para mensurar o léxico de linguagens de especialidade como a biomédica. Com base no conceito de “lexicalidade biomédica”, o novo indicador proposto Lex-BioMed foi capaz de revelar a distribuição lexical nos diferentes gêneros e especialidades clínicas presentes no corpus Covid-19 tomado como referência. Os resultados mostraram que os índices de lexicalidade biomédica caem em contextos fronteiriços e mais genéricos em relação a áreas mais técnicas da medicina como a Cardiologia. O recurso de vocabulário teórico utilizado também se revelou interessante para contornar o problema da variação de número de *tokens* entre os textos do corpus.

Dando seguimento à pesquisa, deverá ser buscada uma ampliação do corpus Covid-19 ora composto de 254 textos e restrito à fase inicial de publicações sobre a doença, eventualmente acrescendo-se períodos posteriores que facultem uma análise de teor mais diacrônico. A ampliação do corpus também contempla a possibilidade de estudos comparados com corpora de outra natureza, por exemplo, construídos com base em um léxico mais popular dirigido ao público leigo. Tais resultados sugerem que outras linguagens de especialidade poderão ser investigadas em trabalhos futuros, tendo como horizonte a

hipótese de que seu léxico também seria sensível a uma mensuração de maior aderência ao texto especializado.

Quanto à identificação manual dos termos biomédicos no corpus, cabem algumas considerações finais. Em primeiro lugar, a filtragem dos termos poderá no futuro ser melhor testada mediante consulta ampla a especialistas da área biomédica com experiência em temas associados à Covid-19. Em segundo lugar, a exploração dos termos biomédicos relacionados ao corpus da pandemia poderá ser enriquecida pelo contraponto com vocabulários eletrônicos, glossários e ontologias disponíveis sobre a Covid em português, ou mesmo aqueles em língua estrangeira que possam ser submetidos a tradução. Em terceiro lugar, a caracterização inicial do léxico da Covid baseado apenas em unigramas, tal como apresentado aqui, certamente ganhará se vier a incorporar multipalavras e *collocations* biomédicas. A testagem do indicador Lex-BioMed, de modo a abranger itens nocionais de formação complexa, abrirá caminho para uma abordagem da complexidade vocabular biomédica em associação a outros níveis linguísticos – sintático, semântico, pragmático e discursivo. Por fim, em quarto lugar, um estudo acurado do comportamento das PoS nas várias especialidades e gêneros de que se compõe o corpus estudado poderá refinar a distribuição lexical mostrada aqui. Para além da constatação feita aqui de que os nomes são as "âncoras nocionais" em todas as especialidades clínicas, um estudo de verbos biomédicos como "acometer" ou "diagnosticar", ou de adjetivos como "anestésico" ou "pulmonar", poderá diferenciar classes com regime mais definido ou indefinido de distribuição no corpus.

Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES.

Referências

V. Kannan e S. Gurusamy (2014), "Preprocessing Techniques for Text Mining – An Overview", International Journal of Computer Science & Communication Networks, V. 5, p. 7-16.

Leite, J.S., Takahata, A.K., Steinberger-Elias, M.(2020) Elaboração de corpus biomédico em Português sobre o Covid-19. Journal of Health Informatics: Número Especial CBIS Congresso Brasileiro de Informática em Saúde. Dezembro p.242-247. http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/821

Leite, J.S., Takahata, A.K., Steinberger-Elias, M. (2020) "Criação e análise de amostras de corpora em Português Brasileiro para detecção automática de expressões complexas em textos sobre Covid-19". In: XXVII Brazilian Congress on Biomedical Engineering. Proceedings of CBEB 2020, October 26-30, Vitoria, Brazil. https://www.springer.com/gp/book/9783030706005

Orengo, V. & Huyck, C. (2001) "A stemming algorithm for the Portuguese language". In Proceedings of the Eighth International Symposium on String Processing and

Information Retrieval (SPIRE 2001), (p. 186-193). Laguna de San Rafael, Chile: IEEE Computer Society Press.

Aluísio, S. M.; Almeida, G. M. D. B. (2006) "O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística Calidoscópio", São Leopoldo, V. 4, n.3. p.155-177.

Celso Romão Cardoso De Almeida Júnior (2017). “Proposta de um Sistema Automático de Avaliação de Redações do Enem, Foco na Competência 1: Demonstrar Domínio da Modalidade Escrita Formal da Língua Portuguesa”. Dissertação de Mestrado.

Cucinotta, Domenico, and Maurizio Vanelli. (2020) “WHO declares COVID-19 a pandemic”. Acta Bio Médica: Atenei Parmensis 91.1 (2020): p.157.

Vasconcellos-Silva, P. R., & Castiel, L. D. (2020) “COVID-19, “As fake news e o sono da razão comunicativa gerando monstros: a narrativa dos riscos e os riscos das narrativas”. Cadernos de Saúde Pública, V. 36, n. 7, p.1-6.

Peixoto, V. R., Mexia, R., Santos, N. D. S., Carvalho, C., & Abrantes, A. (2020) “Da tuberculose ao COVID-19: legitimidade jurídico-constitucional do isolamento/tratamento compulsivo por doenças contagiosas”. In Portugal. Acta Médica Portuguesa, V. 33, p.225-228.

Krieger, M. da G, Finatto, M. J. B. (2004) “Introdução à terminologia: teoria & prática”. São Paulo:Contexto, p.348.

Ure, J. (1971) “Lexical density and register differentiation”. In: G.E. PERREN; J.L.M. TRIM (eds.), Applications of linguistics. Selected papers of the Second International Congress of Applied Linguistics. Cambridge/Londres, Cambridge University Press, p. 443-452.

Johansson, V. (2008) “Lexical diversity and lexical density in speech and writing: a developmental perspective”. Lund University, Department of Linguistics and Phonetics: Working Papers, V. 53, p.61-79.

Broeder, P., Coenen, J., Extra, G., van Hout, R., & Zerrouk, R. (1986) “Ontwikkelingen in het Nederlandstalig lexicon bij anderstalige volwassenen: Een macro- en microperspectief”. In J. Creten, G. Geerts, & K. Jaspaert (Eds.), Werk-in-uitvoering: Momentopname van de sociolinguïstiek in België en Nederland, p.39-57.

Perna, L. Cristina; Delgado, K. Heloísa; Finatto, J. Maria. (2010) “Linguagens Especializadas em CORPORA. Modos de Dizer e Interfaces de Pesquisa”. EDIPUCS-Editora Universitária da Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, p.138.

Santos, E. S. et al. (2018) “Diversidade e densidade lexical em textos escritos por alunos recém-alfabetizados: um estudo descritivo de produções individuais e em díades”. Calidoscópio Revista Unisinos, V. 16, n.1, p.25-32.

Silva W.D.C.M. (2013) “Aprimorando o corretor gramatical CoGrOO”, Dissertação de Mestrado em Ciência da Computação, IME-USP, São Paulo, SP.

Zilio, L. (2009) “Colocações especializadas e Komposita: um estudo contrastivo alemão-português na área de cardiologia”. Porto Alegre: UFRGS. Dissertação de Mestrado. PPG-LETRAS/UFRGS.