UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ESTUDOS DA LINGUAGEM

YULI SOUZA CARVALHO

INTELIGIBILIDADE E CONVENCIONALIDADE EM TEXTOS DE DIVULGAÇÃO DA ÁREA MÉDICA:

Uma análise à luz da Linguística de Corpus

PORTO ALEGRE
2020

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ESTUDOS DA LINGUAGEM

YULI SOUZA CARVALHO

INTELIGIBILIDADE E CONVENCIONALIDADE EM TEXTOS DE DIVULGAÇÃO DA ÁREA MÉDICA:

Uma análise à luz da Linguística de Corpus

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul como parte dos requisitos para obtenção do título de Mestre em Letras.
Orientadora: Profª. Drª. Rozane Rodrigues Rebechi.

PORTO ALEGRE
2020

CIP - Catalogação na Publicação

Carvalho, Yuli Souza
INTELIGIBILIDADE E CONVENCIONALIDADE EM TEXTOS DE DIVULGAÇÃO DA ÁREA MÉDICA: Uma análise à luz da Linguística de Corpus / Yuli Souza Carvalho. -- 2020.
152 f.
Orientadora: Rozane Rodrigues Rebechi.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2020.

1. Textos de Divulgação. 2. Tradução. 3. Acessibilidade Textual. 4. Convencionalidade. 5. Inteligibilidade. I. Rebechi, Rozane Rodrigues, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

YULI SOUZA CARVALHO

INTELIGIBILIDADE E CONVENCIONALIDADE EM TEXTOS DE DIVULGAÇÃO DA ÁREA MÉDICA:

Uma análise à luz da Linguística de Corpus

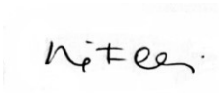
Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul. Área: Estudos de Linguagem. Linha de pesquisa – Lexicografia, Terminologia e Tradução: Relações Textuais.

Aprovada pela banca examinadora em 23 de setembro de 2020.

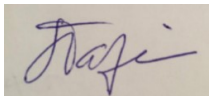
BANCA EXAMINADORA

Profª. Drª. Rozane Rodrigues Rebechi (Orientadora)

Profª. Drª. Maria José Bocorny Finatto



Profª. Drª. Bianca Franco Pasqualini



Profª. Drª. Stella Esther Ortweiler Tagnin

AGRADECIMENTOS

À minha mãe, por sempre me apoiar nas minhas decisões e, posteriormente, me aguentar sofrendo por causa delas. Por ter sempre me oferecido as melhores oportunidades alcançáveis. Por ter feito o papel de mãe e pai durante toda a minha vida e a da minha irmã. Por sempre ter colocado nosso bem-estar em primeiro lugar, por vezes abrindo mão de oportunidades para si mesma.

À minha irmã, que mesmo (fisicamente) longe está sempre presente nos meus dias, por sempre tentar me motivar a ver as coisas positivamente. Por, também, sempre me apoiar nas minhas decisões e, também, me aguentar sofrendo por causa delas.

À minha orientadora, por ter me auxiliado durante os 2 anos de Mestrado e me acompanhado na trajetória até este trabalho. Por sempre ter exigido mais de mim quando sabia que eu conseguia ir mais longe. Por ter se mantido presente mesmo durante este período conturbado de pandemia.

Às professoras maravilhosas que tive durante a minha formação, não só no período do Mestrado, mas também da graduação.

Aos meus amigos e à minha família, por terem me ajudado a manter a sanidade mental no período de término deste trabalho, que coincidiu justamente com o momento (provavelmente) mais atípico de nossas vidas.

RESUMO

Este trabalho aborda os temas de inteligibilidade e convencionalidade em textos de divulgação, a partir da análise de originais em inglês, de suas traduções para o português e de originais em português. Como pilares para o trabalho, serão utilizados conceitos relacionados aos Estudos de Tradução, à Linguística de Corpus, à acessibilidade textual e ao gênero texto de divulgação. O principal objetivo do trabalho é investigar se a tradução pode ser um meio que acaba por dificultar a compreensão de textos que têm como finalidade apresentar informações de maneira clara e simples ao leitor médio. Para isso, serão apresentados dados sobre analfabetismo e nível de instrução das populações estadunidense e brasileira. Com a finalidade de testar a hipótese, foram construídos dois corpora de estudo que foram as bases para a análise. O corpus paralelo é formado por textos de divulgação escritos originalmente em inglês e suas traduções para o português, e o corpus comparável é composto por um subcorpus de textos de divulgação escritos originalmente em português junto e pelo subcorpus de textos traduzidos para o português. A metodologia do estudo combina análises quantitativas à análise qualitativa. Na análise quantitativa, foram utilizadas as ferramentas Coh-Metrix e Coh-Metrix-Port, com o levantamento de Índice Flesch dos textos dos corpora e, posteriormente, calculando a média, a mediana, a variância e o desvio padrão. A análise qualitativa foi auxiliada pelas ferramentas AntConc e AntPConc, a partir da investigação de palavras-chave exclusivas do subcorpus de textos traduzido, recorrendo ao corpus paralelo para buscar a origem dessas palavras-chave, e ao subcorpus de originais em português para apontar se essas escolhas tradutórias fogem ao vocabulário tipicamente utilizado nos textos de divulgação escritos por brasileiros. Como resultado do levantamento do Índice Flesch, observou-se que os textos em inglês, cujo público tem mais alto nível de escolaridade, obtiveram o índice mais elevado de inteligibilidade, sendo considerados ‘razoavelmente fáceis’; em segundo lugar, vieram os textos traduzidos para o português, que tiveram sua média de índice categorizada como ‘razoavelmente difícil’; por fim, os textos escritos originalmente em língua portuguesa tiveram média de Índice Flesch considerada ‘difícil’. Por outro lado, os resultados da análise qualitativa apontaram que os textos de divulgação traduzidos apresentam várias quebras de convencionalidade, com opções de escolhas tradutórias que, por vezes, não são condizentes com o que recorre nos originais em português. Além disso, a partir das análises feitas utilizando o corpus paralelo, descobriu-se que os textos traduzidos apresentam alta recorrência de palavras cognatas do inglês, quando comparados aos originais em português, apontando preferência por equivalentes *prima facie*. A partir dos resultados, concluiu-se que, apesar de os textos traduzidos terem obtido índices de inteligibilidade mais altos que os originais em português, a quebra da convencionalidade pode ser um fator que influencia o entendimento do leitor brasileiro, gerando dificuldades na associação entre o vocabulário dos textos e os conceitos aos quais esse vocabulário se refere.

Palavras-chave: Textos de Divulgação; Tradução; Acessibilidade Textual; Convencionalidade; Inteligibilidade.

ABSTRACT

This research addresses themes of readability and conventionality in expository texts, based on the analysis of originals in English, their translations into Portuguese and originals in Portuguese. As pillars for the work, we will use concepts related to Translation Studies, Corpus Linguistics, textual accessibility and genre expository text. The main objective is investigating whether translation can be a means of hindering text understanding when they are intended to present information in clear and simple language to average readers. For this, data about literacy and education level of U.S. American and Brazilian populations will be presented. In order to test the hypothesis, two corpora were built to serve as basis for the analysis. The parallel corpus is comprised of expository texts originally written in English and their translations into Portuguese, and the comparable corpus comprises a subcorpus of expository texts originally written in Portuguese and the subcorpus of translations into Porguese. The methodology combines quantitative and qualitative analyses. Coh-Metrix and Coh-Metrix-Port tools were used for quantitative analysis in order to calculate Flesch Reading Ease. After that, mean, median, variance and standard deviation were calculated. Qualitative analysis was made using AntConc and AntPConc tools. Unique keywords from the translation subcorpus were analyzed, using the original texts in English to search for the source of those keywords, and original texts in Portuguese to point out whether these translation choices are far from the vocabulary typically used in expository texts written by Brazilians. As a result of Flesch Reading Ease indexes, it was observed that texts in English, whose public has higher education level, obtained the highest readability index, being considered 'fairly easy'; in second, we find the texts translated into Portuguese, which had their average index categorized as ‘fairly difficult'; at last, texts originally written in Portuguese had an Flesch Reading Ease average considered 'difficult'. On the other hand, qualitative analysis results pointed out that translated expository texts present several conventionality deviations, as translators chose equivalents that are not consistent with what is used in originals texts in Portuguese. Moreover, through the analysis of the parallel corpus, we discovered that translated texts present recurrence of cognate words from English, pointing to several *prima facie* translations. It was possible to conclude that, although translated texts have obtained higher readability indexes than originals in Portuguese, conventionality deviation may be an influence factor for Brazilian readers’ comprehension, leading to difficulties in associations between vocabulary and concepts this vocabulary refers to.

Keywords: Expository Texts; Translation; Plain Language; Conventionality; Readability.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

SUMÁRIO

INTRODUÇÃO

Os textos de divulgação têm papel informativo para o público geral. Portanto, é necessário que os dados apresentados nesses textos sejam fornecidos de maneira clara e acessível. Assim como elevadores e rampas visam a tornar as construções acessíveis ao público geral, a acessibilidade textual depende de características que tornam os textos mais (ou menos) claros para leitores conforme suas idades, classes sociais e escolaridades. Como consequência de um alto nível de complexidade textual, esses textos podem acabar não atingindo seu objetivo final, qual seja, o de instruir a população geral acerca de temas especializados, pois o público geral poderia ter dificuldade para compreender as informações neles contidas.

O interesse em investigar a questão da acessibilidade textual surgiu nas disciplinas de Estágio Supervisionado de Tradução do Inglês, que são obrigatórias nas etapas 7 e 8 do curso de bacharelado em Letras da Universidade Federal do Rio Grande do Sul. Nessas disciplinas, a proposta é desenvolver traduções, sob a supervisão de um professor orientador. A partir de traduções de textos em inglês da área de obstetrícia, observou-se um fenômeno recorrente durante a prática tradutória. Esse fenômeno não configurava erros de tradução ou falhas em reproduzir os traços do gênero do texto, mas sim um desequilíbrio percebido entre o nível de complexidade do texto original e do texto traduzido. A exemplo disso, ao traduzir para o português, a repetição de palavras era evitada, sempre buscando por sinônimos, e as frases acabavam ficando mais longas.

Atualmente, consumimos diversos produtos por meio da tradução, seja em forma de textos escritos ou de produções audiovisuais. Ao considerarmos que os textos de divulgação devem ser acessíveis para uma grande parcela da população, é necessário investigar se o processo de tradução resulta em textos acessíveis ao leitor, principalmente no que diz respeito a aspectos como a convencionalidade e a inteligibilidade.

A convencionalidade está relacionada ao domínio da fluência de determinada língua (TAGNIN, 2013). Isso quer dizer, em outras palavras, que os falantes dispõem de unidades linguísticas previamente armazenadas na memória. Assim, ao produzir um enunciado, reiteram-se sintagmas previamente usados, causando a cristalização

de padrões linguísticos. Na mesma direção, a inteligibilidade é a qualidade que caracteriza algo que é claro e de fácil compreensão (DUBAY, 2004).

Esta pesquisa está inserida na grande área de Estudos da Linguagem, na linha de pesquisa de Lexicografia, Terminologia e Tradução. Mais especificamente, os temas aqui tratados serão a acessibilidade textual e a tradução, para atingirmos o propósito de analisar a complexidade textual em traduções de textos de divulgação do inglês para o português.

A fim de traçarmos uma análise de textos de divulgação traduzidos, foram compilados dois corpora. O primeiro deles é composto por textos de divulgação escritos originalmente em língua inglesa e suas respectivas traduções para o português, e o segundo, por textos escritos em português brasileiro.

Textos de divulgação, bem como outros gêneros textuais, possuem características específicas, sendo a mais marcante delas o fato de estarem situados entre o falar científico e o falar comum, apresentando vocabulário desses dois registros (KRIEGER, 2009). Assim, será verificado se o nível de complexidade desses textos traduzidos está adequado para o seu público-alvo, que é o leitor médio brasileiro, quando comparados com os textos originais em português. Primeiro, será feita uma análise de inteligibilidade de todos os corpora utilizados neste estudo. Em seguida, será feito o contraste em relação à convencionalidade, baseando-se nos textos originalmente escritos em português e recorrendo aos originais em inglês, quando necessário, para atestar possíveis influências do texto-fonte no texto-alvo.

Portanto, serão feitas não só análises quantitativas, por meio dos índices de inteligibilidade e chavicidade, mas também análises qualitativas, auxiliadas por ferramentas de Linguística de Corpus. Para esgotar o objetivo da investigação proposta, a pesquisa não poderia se manter apenas no âmbito estatístico, mas sim tomá-lo como base para uma investigação mais aprofundada. Afinal, de acordo com Biderman (1967), os “primeiros senões facilmente apreensíveis são constituídos pelos dois aspectos irredutíveis da realidade linguística: o elemento qualitativo e o quantitativo”, reiterando a importância de se analisarem os textos por esses dois vieses.

As hipóteses a serem verificadas são de que: 1) os textos escritos originalmente em português são mais inteligíveis, quando em comparação com os traduzidos; 2) os textos traduzidos apresentam nível de complexidade textual mais alto do que os

escritos originalmente em língua portuguesa, apresentando quebras de convencionalidade que dificultam o entendimento; e 3) os textos traduzidos apresentam nível de complexidade textual mais alta do que seus originais em língua inglesa.

O acesso à informação é um direito assegurado na lei (BRASIL, 2015). Contudo, para que a população possa usufruir desse direito, devem-se garantir condições não só de acesso, mas também de compreensão desses materiais. Ao traduzir textos com a finalidade de atingir o público geral, o tradutor deve levar em conta informações sobre o leitor a quem as traduções se destinam (NORD, 2006). Com base nisso, a escolaridade e a proficiência desse público são indispensáveis para que se pense a tradução. Portanto, neste estudo, também será apresentado quem é o leitor médio brasileiro e estadunidense, e verificado se os textos estão adequados para esses públicos.

O objetivo geral da pesquisa é analisar como se dá a relação entre complexidade textual e tradução, fazendo uma análise da inteligibilidade e da convencionalidade dos textos. Para alcançar esse objetivo maior, os objetivos específicos são:

1) Comparar a complexidade textual dos textos escritos originalmente em português com a dos textos traduzidos para o português;
2) Comparar a complexidade textual dos textos escritos em inglês com suas respectivas traduções;
3) Fazer um levantamento da adequação dos textos para seus respectivos públicos-alvo;
4) Fazer um paralelo de levantamentos estatísticos de palavras-chave e n-gramas de textos originais e traduzidos, para comparar o vocabulário empregado pelos tradutores;
5) Apontar discussões sobre as consequências que os níveis de complexidade podem acarretar;
6) Levantar conclusões do efeito que esses resultados têm sobre a circulação desses textos.

A partir disso, a motivação é descrever a relação que se dá entre inteligibilidade textual e tradução, com o objetivo de dar subsídios para ajudar os profissionais do texto a refletirem sobre traços de convencionalidade, que são essenciais para o entendimento por parte do leitor médio. Esses traços podem ajudar na elaboração de um texto-alvo que seja compatível com a proficiência em leitura do público ao qual as traduções se destinam.

A presente dissertação está dividida em sete partes. Esta introdução compõe a primeira parte, apresentando o tema, situando em que área a pesquisa está inserida e apresentando as hipóteses. O capítulo 1 apresentará a fundamentação teórica, onde serão abordados os seguintes temas: Linguística de Corpus e tradução; acessibilidade textual; e o gênero textos de divulgação. O capítulo 2 dará insumos sobre dados de analfabetismo e graus de instrução da população estadunidense e brasileira, para situar quem é o leitor médio dos originais em inglês e de suas traduções para o português e dos originais em português. O capítulo 3 traçará a metodologia aplicada para desenvolver a pesquisa, apresentando os dois corpora de estudo e os *softwares* e as ferramentas utilizados para realizar as análises quantitativa e qualitativa. No capítulo 4, serão apresentados os resultados referentes aos levantamentos quantitativos, possibilitados pelo Coh-Metrix e pelo Coh-Metrix-Port, e aos levantamentos qualitativos, feitos no AntConc e no AntPConc. O capítulo 5 discutirá os resultados, categorizando-os em aspectos macroestruturais e aspectos microestruturais dos textos. Por fim, a última parte do trabalho apontará as considerações finais acerca das análises e as conclusões que foram delas tiradas.

1 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados conceitos importantes da Linguística de Corpus e dos Estudos de Tradução para esta dissertação, além de estudos que combinam a metodologia de Linguística de Corpus a estudos descritivos sobre a tradução. A acessibilidade textual será definida, apontando os principais conceitos acerca desse tema e traçando, de forma breve, a trajetória das lutas em prol da acessibilidade textual. Por fim, o gênero texto de divulgação será detalhado e caracterizado.

1.1 LINGUÍSTICA DE CORPUS E TRADUÇÃO

Para tratar da análise da complexidade textual nos textos selecionados para os corpora de estudo desta dissertação, dois temas, de suma importância para o desenvolvimento da pesquisa, serão abordados no presente capítulo: a tradução e a Linguística de Corpus.

Atualmente, recebemos muitas informações por meio da tradução. Assistimos a filmes legendados ou dublados, lemos livros traduzidos, ou ainda recebemos produtos ao vivo por meio da tradução simultânea. Assim, a tradução desempenha um papel central no nosso cotidiano. Portanto, em primeiro lugar, será fornecida uma introdução aos estudos da tradução, com base em Even-Zohar (1990), Toury (1995) e Nord (2006), a fim de apresentar um breve panorama da trajetória da área até os estudos descritivos. Posteriormente, serão apresentadas as formas como a tradução se relaciona com a Linguística de Corpus, com base em Baker (1993) e Fantinuoli e Zanettin (2015). Por fim, serão descritas pesquisas que, de alguma forma, se aproximam da proposta do presente trabalho, como as de Frankenberg-Garcia (2006), Rebechi (2017), Fuchs (2018), Pasqualini (2012) e Lima (2013).

1.1.1 A abordagem descritiva da tradução

As análises de traduções muitas vezes caem na armadilha de traçar julgamentos de valor sobre a obra traduzida, apontando momentos em que o tradutor “errou” ou “poderia ter feito diferente”. Nesse sentido, Baker (1993) afirma que dois

temas centrais permearam, por muito tempo, o foco das discussões acerca da tradução: a fidelidade e a noção de equivalência. Apesar de serem temas que perpassaram a história dos estudos da tradução, são questões que se detêm a julgamentos qualitativos acerca do texto traduzido. Chegar o mais próximo possível do original e encontrar a tradução exata para alguma palavra ou termo podem ser questões-chave, mas não devem resumir os estudos de tradução.

Segundo Toury (1995, p. 1)[1,2], "nenhuma ciência empírica pode alegar completude e (relativa) autonomia a menos que possua um ramo descritivo adequado". No que diz respeito aos estudos de tradução, o autor afirma que a perspectiva descritiva só começou a ser desenvolvida por volta de 1995, enquanto em áreas afins – como literatura comparada, psicolinguística, linguística do texto etc. –, os estudos descritivos já estavam avançados.

Para ir além das análises puramente prescritivas da tradução, uma metodologia que ajuda a tornar o estudo mais descritivo, com base em dados reais, é a Linguística de Corpus. Corpus é entendido como uma coleção de textos autênticos, coletados criteriosamente e armazenados em formato eletrônico, com o objetivo de servirem para pesquisas linguísticas (REBECHI, 2017).

De acordo com Baker (1993), os textos traduzidos não eram considerados, na Linguística de Corpus, como expressões reais de língua. Por exemplo, para a compilação de um corpus de língua portuguesa, os textos traduzidos seriam ignorados, por não se tratarem de expressão natural da língua, e sim, algo "inferior" aos textos originalmente escritos na língua. Para a autora, os textos traduzidos possuem, sim, suas especificidades, o que não os torna inferiores àqueles produzidos originalmente em determinado idioma. Isso só prova que a tradução merece ser um objeto de estudo analisado a partir de dados quantitativos, para mapear de que forma os textos traduzidos se distinguem dos seus originais, ou de textos produzidos originalmente na língua. Também, apontar essas diferenças pode auxiliar na avaliação do quão acessíveis e convencionais são as traduções em comparação com seus originais.

Para Baker (1993), algumas mudanças nos estudos da tradução vinham abrindo espaço para a abordagem da Linguística de Corpus. A primeira delas foi

[1] Todas as traduções de citações apresentadas nesta dissertação foram elaboradas pela autora.
[2] No original: "[…] *no empirical science can make a claim for completeness and (relative) autonomy unless it has a proper descriptive branch*."

deixar de lado a visão de que o texto-fonte seria o modelo superior a ser alcançado, dando mais visibilidade à língua-alvo e ao texto-alvo. A segunda mudança que a autora cita é o abandono da premissa de que existe uma mensagem ou um significado inerente ao texto de partida que deveria obrigatoriamente ser transferido ao texto de chegada. Por fim, Baker (1993) apresenta como importantes, também, as mudanças na visão do texto-fonte como primazia e na visão do que seja equivalência. Essas mudanças levaram os estudos de tradução do viés prescritivo para o descritivo, bem como carregaram a perspectiva conceitual para uma mais situacional. Baker (1993) aponta como influências para essas mudanças nos estudos da tradução os trabalhos de dois pesquisadores: Even-Zohar (1990), com a teoria dos polissistemas, e Toury (1991 *apud* BAKER, 1993), com as reflexões sobre norma.

A teoria dos polissistemas, de Even-Zohar (1990), apresenta um novo olhar sobre a literatura traduzida. Se antes os textos traduzidos eram vistos como "colados" ao texto original, essa teoria trouxe a perspectiva de que as traduções deviam ser tratadas como uma produção individual e com certa independência. Na perspectiva do autor, o polissistema literário não deve ser visto como uma coleção de textos diferentes e estáticos, mas sim como um agrupamento de sistemas hierárquicos e dinâmicos. Por isso, o autor apresenta a divisão de periférico *vs.* central para tratar esses sistemas. Pensando idealmente, a literatura nacional produzida em um país ocuparia a posição central no polissistema, enquanto a literatura traduzida ocuparia uma posição mais periférica. Entretanto, essa disposição não é a regra. Há diferentes fatores que influenciam o deslocamento do periférico para o central, e vice-versa.

Para Even-Zohar (1990), as literaturas estrangeiras traduzidas podem inserir novas características no polissistema literário da língua-alvo, introduzindo novas linguagens e técnicas.

Os estudos descritivos de Toury (1995) também desempenharam importante papel na visão sobre a literatura traduzida. Enquanto Even-Zohar (1990) tratava dos sistemas literários, Toury (1995) apontava a influência do sistema e da cultura da língua-alvo sobre o texto traduzido.

De acordo com Toury (1995), a posição do texto traduzido na cultura receptora é um dos fatores que regem o processo da tradução. A quantidade de características do original que aparecem no texto-alvo é determinada pelo domínio alvo, pois, na visão da cultura receptora, há aspectos específicos mais importantes de serem

transmitidos do que outros. Essa decisão é guiada pelas normas, ditadas pela cultura receptora.

As normas, para Toury (1995), são valores gerais ou ideias compartilhadas por uma comunidade. Nessas normas, se enquadram as noções de certo e errado, adequado e inadequado, bem como os comportamentos que são aceitáveis ou não em determinado contexto. As normas não são leis escritas ou regras ditadas por alguém; na verdade, elas não dependem de qualquer formulação na língua. Elas são de conhecimento daqueles que compartilham o mesmo contexto – no caso da tradução, pensamos em pessoas que compartilham a mesma língua e cultura em um país.

Toury (1995) considera que a tradução pode ser guiada por duas tendências: adequação ou aceitabilidade[3]. A adequação seria o movimento em que o tradutor se sujeita ao texto original e suas normas. A aceitabilidade seria, então, sujeitar o texto às normas da cultura-alvo. As decisões reais na atividade de tradução acabam incluindo decisões que atendem ora às normas de adequação, ora às de aceitabilidade.

Os estudos de Even-Zohar (1990) e Toury (1995) formaram duas vertentes que começaram a abrir os caminhos para a perspectiva descritivista da tradução. Deixando de lado o ideal de equivalência, esses estudos começaram a tratar da situação real da tradução – apontando descrições a partir de observações do que, de fato, acontecia. Apesar de esses estudiosos tratarem especificamente de textos literários, os conceitos podem também ser aplicados a textos especializados, já que seus estudos possibilitaram que o âmbito descritivo se expandisse cada vez mais.

A abordagem funcionalista proposta por Nord (2006), com base em textos especializados, coloca como foco o propósito do texto traduzido, seu público-alvo e a função da tradução. A busca por equivalência se torna apenas um requisito para alcançar o texto ideal. Nesse sentido, a tradução é tida como uma ação que, como toda atividade humana, é orientada por um objetivo. Como as ações humanas estão situadas em determinados sistemas culturais, devem ser consideradas no processo de tradução a intenção (relacionada ao ponto de vista do emissor) e a função (associada ao receptor, que possui necessidades e expectativas em relação ao texto).

[3] No original: “*adequacy*” e “*acceptability*”.

Para a teórica, funcionalidade significa dizer que o texto 'funciona' para aqueles que irão recebê-lo, em uma situação comunicativa que o emissor deseja que ele funcione. Assim, caso "o propósito seja informação, o texto deve oferecê-la em uma maneira compreensível para o público leitor" (NORD, 2006, p. 31)[4]. O tradutor deve avaliar as capacidades de compreensão e cooperação de sua audiência, antecipando os possíveis efeitos que determinadas escolhas textuais poderão ter sobre o leitor. Vale ressaltar, também, que é o leitor, no momento de recepção, que confere (ou não) ao texto traduzido o *status* de funcional.

Para essa abordagem teórica, a finalidade da tradução é o que deve determinar a escolha do método e das estratégias tradutórias utilizadas, que dependem da função comunicativa almejada pelo texto-alvo. Dessa forma, não se pode garantir a mesma recepção do texto em ambientes culturais distintos.

1.1.2 A relação entre Linguística de Corpus e tradução

Sinclair (1992, p. 395 *apud* BAKER, 1993, p. 242)[5], já no início dos anos 90, afirma: "Espera-se que os novos recursos de corpus possuam profundo efeito nas traduções do futuro". O autor se referia principalmente ao âmbito da tradução automática, que com o auxílio do uso de corpora tem se tornado cada vez mais precisa e eficiente.

Apesar de a previsão de Sinclair (1992 *apud* BAKER, 1993) ter se provado correta, a tradução automática não é a única aplicação possível do uso de corpus na tradução. Fantinuoli e Zanettin (2015) afirmam que a metodologia de corpus permite aplicações práticas não só na tradução profissional humana, mas também nas ferramentas de tradução e na terminologia. A importância do uso dessas aplicações se dá principalmente porque:

> Uma metodologia baseada em Linguística de Corpus possibilita a pesquisa em textos autênticos da área de interesse, a análise de grandes quantidades de dados, o levantamento automático de candidatos a termos e seus colocados, assim como combinações recorrentes (*clusters*), além de facilitar a busca por equivalentes e/ou definições. (REBECHI, 2017, p. 205).

[4] No original: "*If the purpose is information, the text should offer this in a form comprehensible to the audience* […]."
[5] No original: "*The new corpus resources are expected to have a profound effect on the translations of the future*."

Em pesquisas baseadas na metodologia de Linguística de Corpus, é importante delimitar as distinções entre estudos com corpora paralelos ou corpora comparáveis. De acordo com Fantinuoli e Zanettin (2015), grosso modo, estudos com corpora paralelos tratariam do texto-fonte alinhado ao texto traduzido, enquanto os corpora comparáveis seriam a união de textos coletados a partir de critérios compartilhados específicos; entretanto, afirmam eles, isso não é regra. A definição mais adequada, e que será adotada nesta pesquisa, é que:

> Corpora paralelos podem, então, ser considerados como corpora onde dois ou mais componentes estão alinhados, ou seja, estão subdivididos em unidades composicionais e sequenciais (de diferente extensão e tipo) que estão conectadas e podem ser armazenadas como pares (ou trios etc.). Por outro lado, corpora comparáveis podem ser pensados como corpora que são comparados com base em suposta similaridade. (FANTINUOLI; ZANETTIN, 2015, p. 4)[6].

O uso de corpora paralelos é chave para encontrar soluções tradutórias. Grandes corpora podem servir como referência de apoio para o tradutor, principalmente aquele que estiver iniciando-se na profissão, porque possibilitam:

> (i) ir mais longe do que recursos tradicionais de língua, fornecendo os meios de suplementar o conhecimento limitado na cultura e língua-alvo por parte dos tradutores na L2;
> (ii) mostrar mais das produções convencionais da língua do que de costume por parte de seus usuários. (STEWART, 2000, p. 88)[7].

De acordo com Rebechi (2017, p. 204), a comparação entre textos originais lado a lado com suas traduções "possibilita a identificação de equivalentes tradutórios previamente utilizados de forma relativamente simples, por meio do alinhamento das sentenças do texto original com as do texto traduzido". Entretanto, por passar pela mediação de um tradutor, que às vezes não é um especialista da área, o uso apenas de corpora alinhados pode apontar para respostas inconclusivas. Por isso, corpora

[6] No original: "*Parallel corpora can thus be thought of as corpora in which two or more components are aligned, that is, are subdivided into compositional and sequential units (of differing extent and nature) which are linked and can thus be retrieved as pairs (or triplets, etc.). On the other hand, comparable corpora can be thought of as corpora which are compared on the whole on the basis of assumed similarity*."

[7] No original: "(i) *can go further than traditional language resources in furnishing the means to supplement the reduced knowledge of target language and culture on the part of L2 translators, and* (ii) *may bring about more conventional language production than usual on the part of its users*."

comparáveis podem revelar com mais clareza as terminologias e fraseologias que são, de fato, utilizadas nas línguas, ajudando também a identificar discrepâncias no tipo textual nas línguas e culturas com as quais se está trabalhando. Nesse sentido, os estudos que mapeiam características de textos traduzidos em comparação com textos escritos originalmente na língua também fazem uso de corpora comparáveis.

No que diz respeito à importância que o uso de corpora viria a ter nos estudos da tradução, Baker (1993, p. 243)[8] afirma que seria "consequência deles [corpora] nos possibilitarem a identificação de traços de textos traduzidos que nos auxiliariam no entendimento do que é a tradução e como ela funciona". Uma aplicação de corpora nos estudos de tradução seria mapear "universais" da tradução, ou seja, características que perpassam a atividade tradutória em grande parte das línguas.

Acerca desses "universais", a autora cita a tendência de traduções serem explicitadoras do conteúdo do texto original, fornecendo ao leitor da tradução interpretações que no original estavam subentendidas. Nessa direção, a tradução apresenta também movimentos de desambiguação e simplificação, como, por exemplo, o uso de formas pronominais mais precisas, possibilitando que o leitor identifique o referente mais facilmente. Outras tendências tradutórias apontadas em Baker (1993) dizem respeito à gramatização do que no original fugia das regras da gramática da língua, e à esquiva a repetir palavras, substituindo-as por sinônimos ou omitindo-as (BAKER, 1993).

Frankenberg-Garcia (2006) apresenta resultados similares para a língua portuguesa. A autora utilizou como base a ferramenta COMPARA, que, apesar de possuir corpora paralelos do português e inglês, também possibilita o uso da metodologia de corpora comparáveis. A ferramenta tem textos traduzidos em ambas as direções, do português para o inglês, e do inglês para o português. A partir das análises feitas com o uso do COMPARA, Frankenberg-Garcia (2006, p. 147)[9] atesta que "as traduções tendem a ser mais longas do que os textos-fonte, tanto na direção inglês-português como na direção português-inglês", corroborando o que foi apresentado em Baker (1993).

---

[8] No original: "*The profound effect that corpora will have on translation studies, in my view, will be a consequence of their enabling us to identify features of translated text which will help us understand what translation is and how it works*."

[9] No original: "[…] *translations tended to be longer than source texts in both the English-Portuguese and the Portuguese-English directions*."

Além das tendências tradutórias, Baker (1993) menciona estudos sobre a chamada “terceira língua”[10] na tradução. A terceira língua consiste no resultado do confronto entre língua-fonte e língua-alvo, dando ao texto traduzido algumas características que distanciam a tradução tanto do texto-fonte quanto de textos originalmente produzidos na língua-alvo. Portanto, essa terceira língua se refere à influência que a língua do texto-fonte acaba tendo no resultado do texto-alvo.

1.1.3 Pesquisas contrastivas em Linguística de Corpus

Nesta seção, serão relatadas três pesquisas que utilizaram a Linguística de Corpus como metodologia, a fim de descrever as diferenças e as semelhanças entre textos em inglês e português de diferentes gêneros. Além disso, serão apontados os impactos que essas características têm sobre a tradução.

Com base em um corpus comparável, Rebechi (2017) analisou livros de receitas brasileiras escritos originalmente em português brasileiro e em inglês estadunidense. De acordo com a autora, o gênero textual receita culinária em português prescinde de muito detalhamento, sendo que diversas informações ficam implícitas. A partir de sua análise, foi possível destacar quatro principais diferenças no grau em que a língua portuguesa transmite informações para o seu leitor quando comparada à língua inglesa: I) maior imprecisão no que diz respeito às medidas; II) menor detalhamento das etapas que devem ser seguidas; III) menor grau de tecnicidade das informações e IV) diferença na carga semântica dos termos utilizados.

Com base em corpora de resenhas de hotéis, Fuchs (2018) chegou a conclusões similares. A autora concluiu que a cultura estadunidense é, na maioria das vezes, caracterizada pela comunicação de baixo contexto. Um dos fatores que ajudaram a chegar a essa conclusão foi a falta de equilíbrio entre seus corpora de estudo: mesmo com o mesmo número de resenhas, o corpus do inglês tem o dobro de palavras do corpus do português. Além disso, a autora afirma que as resenhas escritas pelos estadunidenses são mais completas e detalhadas. Por outro lado, as resenhas brasileiras são mais sucintas, com uma linguagem mais expressiva e menos concreta, indo ao encontro dos resultados de Rebechi (2017).

[10] No original: “*third code*”.

Em estudo de Pasqualini (2012), é feita uma análise por meio tanto de corpora paralelos como de corpora comparáveis. O estudo analisa a complexidade textual de contos literários, tanto do inglês como do português, e suas respectivas traduções. Para proceder com a análise dos corpora, a autora utilizou ferramentas que aplicam fórmulas que ajudam a estimar a complexidade de um texto. A partir disso, os resultados mostraram “não só uma maior complexidade das traduções em relação aos seus textos-fonte em inglês, como também uma maior complexidade dos textos traduzidos para o português em comparação com textos originalmente escritos em língua portuguesa” (PASQUALINI, 2012, p. 113).

Além disso, a autora apontou que nas traduções feitas do português para o inglês aconteceu o movimento contrário: “traduções para o inglês apresentam nível de complexidade inferior ao de seus textos de origem em português” (PASQUALINI, 2012, p. 118). Assim, pode-se constatar uma tendência nas traduções feitas para o português de tornarem os textos mais complexos do que os originalmente escritos na língua. Por outro lado, a tendência predominante nas traduções feitas para a língua inglesa é mais simplificadora, ou seja, as traduções feitas da língua portuguesa para a língua inglesa facilitaram os textos para seus leitores.

Por fim, Lima (2013) conduziu um estudo de padrões de uso linguístico em corpora comparáveis de textos da área de Medicina (na subárea de triagem neonatal para anemia falciforme). Os textos que compuseram os corpora foram artigos acadêmicos, manuais técnicos e cartilhas de divulgação. Os resultados apontam diferenças em relação ao registro, à variação lexical, densidade lexical, frequência de ocorrência de itens lexicais e itens gramaticais, além do mapeamento da forma como esses itens estão distribuídos em classes de palavras.

Nesta seção, propunha-se apresentar, além de um panorama sobre a tradução, as relações que ela desempenha com a Linguística de Corpus. Por meio dos estudos apresentados, foi possível contemplar diferentes pesquisas aplicando a metodologia da Linguística de Corpus a estudos que podem ser aplicados à tradução. Os estudos apresentados fizeram uso, majoritariamente, de corpora comparáveis, sendo essa metodologia combinada à metodologia de corpora paralelos em alguns casos.

1.2 ACESSIBILIDADE TEXTUAL

Cada vez mais o cidadão brasileiro tem acesso a informações em formato textual por meio da popularização da Internet. Isso não quer dizer, necessariamente, que o acesso à informação esteja diretamente relacionado ao conhecimento. Isso ocorre porque as informações, muitas vezes, não são dadas de forma clara e acessível.

Para compreender do que trata a acessibilidade textual, anteriormente é necessário que se entenda o conceito de 'acessibilidade'. Na legislação brasileira, acessibilidade é definida como:

> possibilidade e condição de alcance para utilização, com segurança e autonomia, de espaços, mobiliários, equipamentos urbanos, edificações, transportes, **informação e comunicação**, inclusive seus sistemas e tecnologias, bem como de outros serviços e instalações abertos ao público, de uso público ou privados de uso coletivo, tanto na zona urbana como na rural. (BRASIL, 2015, on-line, grifo nosso).

Isso mostra o quão abrangente é o termo 'acessibilidade', não estando apenas relacionado a fatores físico-espaciais que garantam o acesso de todas as pessoas a construções e a meios de transporte, por exemplo. Há seis modalidades de acessibilidade que foram categorizadas: I) atitudinal, que diz respeito à percepção do outro sem preconceitos, estigmas, estereótipos e discriminações; II) arquitetônica, que visa à eliminação de barreiras ambientais físicas; III) comunicacional, que tem por objetivo eliminar barreiras na comunicação; IV) instrumental, que visa à superação de barreiras no uso de instrumentos, utensílios e ferramentas; V) metodológica, que prega a ausência de barreiras em metodologias e técnicas de estudo; e VI) programática, que tem por objetivo a eliminação de barreiras presentes nas políticas públicas (BRASIL, 2013).

A acessibilidade comunicacional visa a eliminar barreiras no que diz respeito à comunicação interpessoal, escrita e virtual. Para isso, além da acessibilidade textual, é relevante mencionar recursos que são imprescindíveis para promover o acesso à informação, como interpretação em Língua Brasileira de Sinais, textos em braile, audiodescrição e, ainda, legenda para surdos e ensurdecidos.

Assim como as rampas de acesso para cadeirantes ou elevadores, que podem ser observadas em prédios públicos, e o piso tátil para pessoas cegas, que são vistos

pelas ruas, a acessibilidade atua, no âmbito textual, como um facilitador para o entendimento do texto pelo leitor (FINATTO, 2020). A partir desse conceito de acessibilidade, desenvolvido com amparo legal, instituiu-se o conceito de 'barreiras', que podem ser consideradas:

> qualquer entrave, obstáculo, atitude ou comportamento que limite ou impeça a participação social da pessoa, bem como o gozo, a fruição e o exercício de seus direitos à acessibilidade, à liberdade de movimento e de expressão, **à comunicação, ao acesso à informação, à compreensão**, à circulação com segurança, entre outros [...]. (BRASIL, 2015, on-line, grifo nosso).

Dessa forma, a acessibilidade textual pode ser entendida como uma condição desejada de qualidade de texto, evitando que este conte com barreiras linguísticas. Essa qualidade deve ser almejada quando se escreve com o objetivo de atingir um público-leitor, para que este tenha condições de compreender as informações apresentadas no texto.

Nesse sentido, um segundo conceito que vai ao encontro da acessibilidade textual precisa ser introduzido: o da complexidade textual. A complexidade textual é uma característica de um texto, sendo manifestada por meio da presença de alguns componentes ou traços – como o uso de terminologia, de vocabulário rebuscado, organizações complexas de frases, entre outros – que acabam atuando como dificultadores da compreensão do texto para alguns grupos de leitores.

A fim de estimar o nível de complexidade, foram desenvolvidas fórmulas de inteligibilidade de textos. Essas fórmulas servem para demonstrar matematicamente o quão difícil este pode para determinado público. Vale ressaltar que esses índices devem ser utilizados apenas como auxiliares, não como resultados absolutos.

Inteligibilidade não deve ser confundida com legibilidade. A primeira está relacionada ao que é inteligível, ou seja, aquilo que é de fácil compreensão. A segunda está relacionada ao que é legível, ou seja, que está claro e nítido (DUBAY, 2004). Ambas estão ligadas a formas de compreensão, respectivamente, à compreensão de forma mental e à compreensão de forma visual. Vale ressaltar que a legibilidade tem relação com design e *layout* – como a diagramação, o tipo e o tamanho de fonte etc. –, estando mais relacionada, consequentemente, à acessibilidade visual.

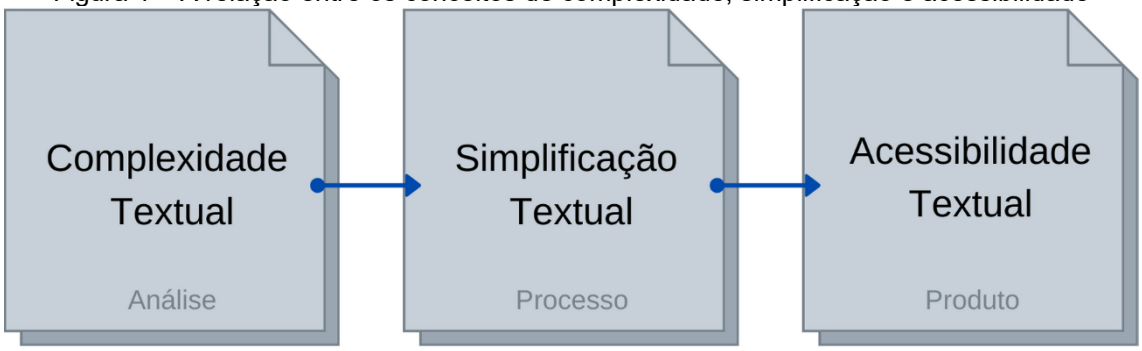
Há outros métodos, além de fórmulas de inteligibilidade, que auxiliam a testagem do nível de complexidade de textos, dependendo de quem é o seu público-alvo. Por exemplo, outro indicador pode ser testes em que pessoas enquadradas

como o público-alvo do seu texto o leiam, apontando onde estariam os dificultadores de compreensão. Esse método, aliado ao uso dos índices, indicaria com maior precisão a avaliação da complexidade de determinado texto.

A simplificação textual vem, então, como um mediador para solucionar o problema da complexidade textual. O objetivo da simplificação é diminuir ou até eliminar as barreiras existentes nos textos, que fazem com que ele não seja entendido por um público específico. De acordo com Finatto (2020, p. 81), "o processo de simplificação [...] poderá ser guiado por uma série de procedimentos e de critérios científicos, previamente estabelecidos e mensurados". Espera-se, portanto, que o produto da simplificação seja um texto acessível – ou seja, com baixo índice de complexidade textual.

A Figura 1 demonstra a trajetória e a relação existente entre os três conceitos previamente apresentados: complexidade, simplificação e acessibilidade.

Figura 1 – A relação entre os conceitos de complexidade, simplificação e acessibilidade



Fonte: reproduzida com base em Paraguassu (2018, p. 127).

Para ilustrar a importância do uso da linguagem acessível, será apresentado o exemplo de um estudo feito no ano de 2003, nos Estados Unidos, relatado por DuBay (2004). Nessa época, acidentes de trânsito vinham causando cerca de 46% das mortes acidentais de crianças. O uso correto do bebê conforto, da cadeirinha ou do assento para as crianças, entretanto, reduziria os riscos de fatalidades em 71%. Todavia, a maioria desses assentos vinha sendo instalado incorretamente.

A partir disso, um estudo foi conduzido para analisar os manuais de instalação desses assentos para crianças. Descobriu-se que os 107 manuais analisados eram escritos numa linguagem considerada difícil para 80% dos adultos no país. Essa porção da população tinha conhecimentos de leitura mais ou menos a nível

equivalente à sétima série do Ensino Fundamental, e os manuais eram escritos com linguagem a nível equivalente ao Ensino Médio (DUBAY, 2004).

1.2.1 A trajetória da luta em prol da acessibilidade textual

Com base no que foi exposto, pode-se depreender que o uso de uma linguagem rebuscada, com palavras desnecessariamente complexas, não instrui o interlocutor, apenas cria uma sobrecarga em sua mente com um acúmulo desnecessário de informações. Nesse sentido, Martinho Lutero, já em 1506, movimentou-se em apoio à tradução da Bíblia para um alemão que o povo conseguisse compreender. Entretanto, foi a partir dos anos 1920 que as pesquisas em inteligibilidade começaram a criar força, partindo da necessidade de adequar materiais de leitura a públicos específicos (PASQUALINI, 2012).

No período após a Segunda Guerra Mundial, com a quantidade de trabalhadores imigrantes com baixo nível de proficiência em inglês que as fábricas estadunidenses receberam, vislumbrou-se a necessidade de métodos que estimassem a inteligibilidade de um texto (DUBAY, 2004).

Flesch (1948 *apud* DUBAY, 2004) foi o pesquisador que desenvolveu uma das fórmulas mais antigas e mais utilizadas para estimar a inteligibilidade de textos. Ele tinha como proposta a escrita clara, simples e acessível à população geral, independente do grau de escolaridade. Flesch defendia o acesso amplo à informação em prol dos Direitos Civis. Isso porque textos de leis comunicam fazendo uso de uma linguagem complexa, muitas vezes inacessível para o público geral. Ele acreditava que as pessoas tinham que entender os seus direitos assegurados por lei, a fim de exigir que eles fossem cumpridos (DUBAY, 2004). Em 1948, ele desenvolveu o Índice Flesch, que estima a complexidade de um texto. Na seção 2.2.1.1, essa medida de inteligibilidade será aprofundada.

Outra figura de destaque na luta a favor da acessibilidade textual foi DuBay (2004). Além de ser a favor da linguagem acessível, ele lutou pelos Direitos Civis, saindo em defesa das minorias e se posicionando contra a segregação racial e a homofobia. Depois de ser expulso da Igreja Católica por publicar um livro repleto de críticas à instituição, o autor seguiu divulgando suas ideias progressistas por meio de jornais, palestras em universidades e livros. Em sua obra sobre estudos de letramento

e de inteligibilidade, o autor compila uma lista de diretrizes, que denomina “regras de ouro para escrever documentos”[11]:

> Usar palavras curtas, simples e familiares;
> Evitar jargões;
> Usar linguagem neutra em relação a cultura ou gênero;
> Usar gramática, pontuação e grafia corretas;
> Usar frases simples, voz ativa e tempo presente;
> Começar instruções no modo imperativo, iniciando as sentenças com verbo de ação;
> Usar elementos gráficos simples, como listas de itens e passos numerados, a fim de tornar a informação visualmente acessível. (DUBAY, 2004, p. 2)[12].

Os movimentos que apoiavam o uso de linguagem acessível eram fortemente ligados ao âmbito do Direito. Em 1972, nos Estados Unidos, foi decretado que o Diário Oficial estadunidense fosse escrito fazendo uso de linguagem para leigos. Posteriormente, em 1978, foram emitidas ordens que exigiam que as regulamentações governamentais fizessem uso de uma linguagem acessível e compreensível, para que aqueles que precisavam cumpri-las conseguissem compreendê-las (MAZUR, 2000).

Na Inglaterra, em 1979, o movimento foi introduzido com um ato radical, onde uma ativista rasgou centenas de documentos oficiais na Parliament Square, em Londres. Esse ato serviu para mostrar a insatisfação com a forma como as informações chegavam à população. A partir disso, surgiu a *Plain English Campaign* (2020), que batalha contra o falar complicado, os jargões e as informações enganosas. Desde então, a organização presta serviço para diversos órgãos do governo.

No Brasil, o movimento em prol da acessibilidade textual não se estabeleceu como produto do ativismo. Foi apenas no final dos anos 1980 que, no âmbito da pesquisa acadêmica, surgiu a necessidade de atender a diferentes tipos de leitores, principalmente no cenário do ensino de línguas. Perini (1982 *apud* FINATTO, 2011)

---

[11] No original: “*golden rules of documentation writing*”.
[12] No original: “*• Use short, simple, familiar words;*
• Avoid jargon;
• Use culture-and-gender-neutral language;
• Use correct grammar, punctuation, and spelling;
• Use simple sentences, active voice, and present tense;
• Begin instructions in the imperative mode by starting sentences with an action verb;
• Use simple graphic elements such as bulleted lists and numbered steps to make information visually accessible.”

alavancou os estudos de compreensão de leitura, com a proposta de que os estudantes tivessem acesso a materiais didáticos desenvolvidos de acordo com sua escolaridade e idade.

Seguindo a mesma linha, Fulgêncio e Liberato (1992) tiveram uma preocupação com a situação de alunos com problemas de aprendizagem. Com base nisso, as autoras descrevem de que forma as informações de um texto são compreendidas durante a leitura. De acordo com elas, um leitor pode não compreender um texto, mesmo que esteja escrito em uma língua que ele domina, se esse texto tratar de um assunto sobre o qual ele não tem informações prévias. As pesquisadoras verificaram, então, passos que permitem chegar à interpretação do texto e aos tipos de informação que um leitor precisa para compreender a mensagem.

Posteriormente, Leffa (1996) apontou três aspectos essenciais para se levar em conta ao falar de compreensão textual: o texto, o leitor, e a relação dada entre os dois. Para Leffa (1996, p. 72), o processo da leitura “envolve vários aspectos, incluindo não apenas características do texto e do momento histórico em que ele é produzido, mas também características do leitor e do momento histórico em que o texto é lido”. Ao discorrer sobre a relação entre o texto e o leitor, o autor pressupõe que, assim como há leitores que têm mais facilidade para ler, também há textos que são mais fáceis de serem lidos. Por essa razão, enfatiza-se a importância de considerar o público-alvo para considerar a acessibilidade de um texto.

No Brasil também houve a preocupação em facilitar a compreensão da linguagem jurídica. Em 2005, surgiu a campanha para a Simplificação da Linguagem Jurídica, da Associação dos Magistrados Brasileiros (2005). A campanha visa a propor uma linguagem mais simples e objetiva no âmbito do Direito, ampliando o acesso da sociedade à Justiça. Um dos maiores problemas do sistema Judiciário, apontado em pesquisa, é a difícil compreensão da linguagem jurídica. Portanto, por se tratar de um serviço público, os fundadores da campanha pregam que a comunicação nessa área deve ser feita de forma acessível a todo cidadão.

Entre 2007 e 2010, na Universidade de São Paulo, um grupo de cientistas da computação e linguistas produziu um projeto chamado PorSimples, que propõe o desenvolvimento de tecnologias que facilitem o acesso à informação. Nesse projeto, o Núcleo Interinstitucional de Linguística Computacional (2010) desenvolveu o Simplifica, um sistema semiautomático para tornar a linguagem escrita mais fácil para

pessoas com dificuldade de compreensão, fazendo com que os textos possam ser entendidos por um número maior de leitores.

Desde 2019, na Universidade Federal do Rio Grande do Sul, está sendo desenvolvido o projeto MedSimples (PARAGUASSU *et al.*, 2020), com o propósito de construir uma ferramenta de simplificação, promovendo mais acessibilidade na comunicação nas áreas da saúde (TEXTECC, 2020). Atualmente, o MedSimples utiliza quatro bancos de dados para auxiliar na adequação de textos dos temas de doença de Parkinson, COVID-19 e Pediatria, visando atingir três possíveis perfis de leitores, levando em conta seus anos de escolaridade.

Nesta seção, os principais conceitos relacionados à acessibilidade textual que serão importantes para esta pesquisa foram apontados e definidos. Além disso, de forma breve, foi traçada a trajetória de trabalhos a favor da acessibilidade textual ao longo dos anos.

1.3 O GÊNERO TEXTO DE DIVULGAÇÃO

Os textos que compõem os corpora de análise são textos de divulgação escritos originalmente em língua inglesa, suas traduções para o português e textos de divulgação escritos originalmente em língua portuguesa (mais informações sobre esses corpora serão detalhadas no capítulo de metodologia). Serão apresentadas, a seguir, características desse gênero e sua importância no que diz respeito à instrução do público leigo.

As atividades de divulgação científica têm origem no discurso oral, junto com o surgimento da Ciência Moderna. No século XVIII, em anfiteatros, demonstrações de fenômenos pneumáticos, elétricos e mecânicos eram apresentadas ao público, bem como exposições e palestras de Física, Química e Medicina percorriam as cidades e, às vezes, países (SILVA, 2006). Essa era uma forma de levar a informação à população que, na época, estava interessada nesses temas.

Textos de divulgação, assim como outros gêneros textuais, possuem características específicas. A mais marcante delas é estar situado entre o falar científico e o falar comum, apresentando vocabulário desses dois registros:

> seria redutor pensar a divulgação científica apenas como uma redução ou adaptação de textos científicos, elaborados para a leitura de pares dotados de uma mesma competência profissional. Ao contrário, a divulgação científica em seu amplo universo, ainda carente de descrições, afigura-se como uma categoria textual autêntica, com regras próprias de produção de significação e de recursos que visam a uma comunicação eficiente. (KRIEGER, 2009, p. 9-10).

O termo ‘divulgação científica’, de acordo com Silva (2006, p. 53), “está relacionado à forma como o conhecimento científico é produzido, como ele é formulado e como ele circula numa sociedade como a nossa”. Assim, o principal contraponto e o mais evidente entre textos científicos e textos de divulgação é a diferença com base em seus públicos-alvo.

Massarani e Moreira (2005) distinguem três categorias na comunicação científica: I) os discursos científicos primários, escritos por pesquisadores para pesquisadores; II) os discursos didáticos, como os manuais científicos para ensino; e III) os discursos de divulgação científica. Enquanto o destinatário do texto científico é um par com conhecimento especializado sobre o tema, o texto de divulgação é voltado para pessoas leigas, sem conhecimento prévio construído sobre o que será abordado no texto (SANTIAGO, 2007). Dessa forma, ao escrever um texto de divulgação, é necessário que o especialista transmita sua mensagem de maneira diferente do que faria a especialistas no assunto. Portanto, um dos pontos importantes é como a terminologia será comunicada a esse público leigo.

Bem como os textos científicos, os textos de divulgação acabam contendo terminologias. Entretanto, nesse gênero, as terminologias devem ser adaptadas conforme o seu público-alvo, sendo utilizadas, então, variantes populares dos termos técnicos (KRIEGER, 2009). Como um exemplo disso, podemos citar ‘lepra’, nome popularmente utilizado para se referir a ‘hanseníase’.

De acordo com Ciapuscio (1998), com a crescente alfabetização científico-tecnológica da sociedade, os conceitos acabam transcendendo a limitação do âmbito especializado, sendo incorporados na comunicação cotidiana. Assim, para que seja feita a comunicação entre especialistas e leigos, aqueles necessitam recorrer a alguns recursos de tratamento linguístico. A autora aponta como recorrentes os usos de tratamento parafrástico (definição e explicações) e tratamento não parafrástico (inclusão de informações enciclopédicas). O tratamento linguístico configura uma

forma de compatibilizar a linguagem do especialista, responsável por redigir o texto, a seu público-alvo (KRIEGER, 2009).

Além de artigos sobre temas de ciência e tecnologias divulgados em mídia escrita, podem-se presenciar expressões de divulgação científica em exposições, museus ou até vídeos divulgados na Internet. Atualmente, apesar de haver inúmeras informações espalhadas na Internet sobre os mais diversos temas, é necessário que seja checada a confiabilidade do material. Diversos *sites* apresentam matérias e textos com informações duvidosas ou não confirmadas pela comunidade científica. Portanto, o ideal é buscar informações apresentadas por veículos oficiais, seja de notícias, universidades ou canais que tratam seriamente dos assuntos.

Com isso em mente, a seleção dos textos para os corpora, e das entidades por eles responsáveis, se deu de forma criteriosa, a partir de dois portais confiáveis. Os textos brasileiros foram extraídos do *site* do Ministério da Saúde, órgão oficial do governo brasileiro responsável pela administração e manutenção da Saúde pública. Os textos em inglês e suas traduções foram extraídos do *site* MedlinePlus, mantido pela U.S. National Library of Medicine com o intuito de ajudar na localização de informações oficiais sobre saúde. Mais detalhes sobre a seleção e a compilação dos corpora de estudo estão disponíveis no capítulo de metodologia.

Neste capítulo, foram apresentados conceitos importantes para o desenvolvimento deste trabalho. Primeiro, foi estabelecida a trajetória dos Estudos de Tradução e sua relação com a Linguística de Corpus. Além disso, foram relatados trabalhos de Linguística de Corpus que tratam dos impactos que as diferenças entre as línguas têm sobre a tradução.

Após, o tema da acessibilidade textual foi abordado, detalhando os principais conceitos da área e traçando a trajetória da defesa da acessibilidade textual. Por último, foram apontadas características pertinentes do gênero de textos de divulgação.

No capítulo a seguir, serão detalhados dados sobre analfabetismo e nível de instrução nos Estados Unidos e no Brasil, a fim de verificar o quão capacitado está o leitor médio de cada país para compreender os textos que fazem parte dos corpora de estudo.

2 DADOS DE ANALFABETISMO E INSTRUÇÃO NOS ESTADOS UNIDOS E NO BRASIL

Para demonstrar a importância de textos escritos com um nível de inteligibilidade adequado para o público geral, apresentam-se, a seguir, dados sobre analfabetismo e grau de instrução nos Estados Unidos e no Brasil, para fins de comparação. Os dados apresentados são provenientes do National Center for Education Statistics (2019), e de pesquisas elaboradas pelo U.S. Census Bureau (2017), para os Estados Unidos, e de levantamentos do Instituto Paulo Montenegro (2018) e do IBGE (2019), para o Brasil.

2.1 ANALFABETISMO E INSTRUÇÃO NOS ESTADOS UNIDOS

Os dados apresentados pelo National Center for Education Statistics são provenientes da pesquisa *Program for the International Assessment of Adult Competencies*. Essa pesquisa levanta informações das habilidades da população em três domínios: alfabetização, conhecimentos matemáticos e resolução de problemas. Nesses levantamentos, considera-se como alfabetização[13] a habilidade de avaliar, utilizar e compreender textos, com a finalidade de participar na sociedade, alcançando objetivos e desenvolvendo conhecimentos. Por desacreditarem que esta seja uma condição que os indivíduos possuem ou não, baseiam-se em uma escala de alfabetização para classificar o nível de proficiência na língua da população (NATIONAL CENTER FOR EDUCATION STATISTICS, 2019).

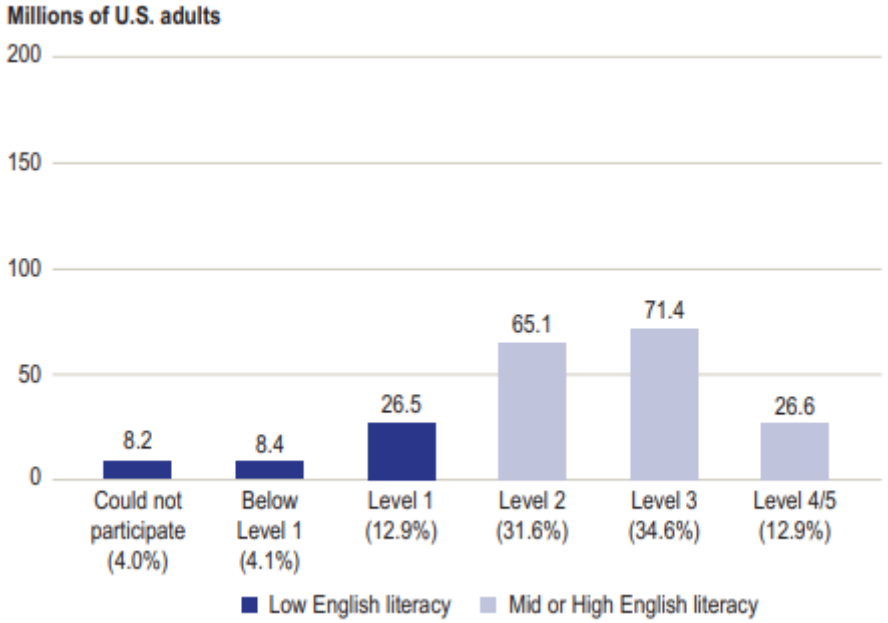
As classificações para os níveis de alfabetização estão divididas em 5 níveis. O baixo nível de alfabetização (*Low English literacy*) inclui aqueles que não puderam participar (*Could not participate*) devido à barreira linguística ou a alguma dificuldade física ou cognitiva, abaixo do nível 1 (*Below Level 1*) e nível 1 (*Level 1*). O nível médio ou alto de alfabetização (*Middle or High English literacy*) abrange nível 2 (*Level 2*), nível 3 (*Level 3*) e nível 4/5 (*Level 4/5*).

A população abaixo do nível 1 foi categorizada dessa forma porque “não foi capaz de determinar o significado de frases, de ler textos relativamente curtos para encontrar uma informação, ou de preencher formulários simples” (NATIONAL

[13] No original: “*literacy*”.

CENTER FOR EDUCATION STATISTICS, 2019, p. 1)[14]. De acordo com a descrição do National Center for Education Statistics (2019), adultos que se enquadram na classificação abaixo do nível 1 são considerados analfabetos funcionais[15]. O Gráfico 1 demonstra os resultados dessa pesquisa.

Gráfico 1 – Taxas de alfabetização conforme o National Center for Education Statistics



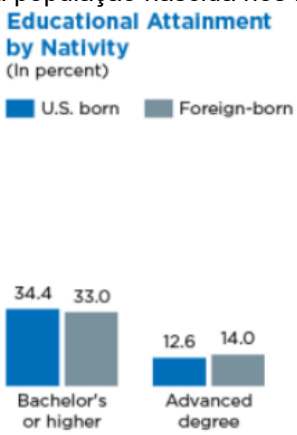
Fonte: National Center for Education Statistics (2019, p. 1).

Como pode ser observado no gráfico, 4,1% da população estadunidense é composta por analfabetos funcionais. A maior parte da população se enquadra no nível 3, de alfabetização média ou alta.

O U.S. Census Bureau (2017) é o órgão federal estadunidense responsável pelo sistema estatístico federal do país, produzindo, a cada dez anos, levantamentos sobre a população e a economia. No que diz respeito à escolaridade, de acordo com esse órgão, a porcentagem da população acima de 25 anos que concluiu o *high school* (ou graus mais altos de instrução) chegou aos 90% – nível de escolaridade similar ao Ensino Médio brasileiro. O Gráfico 2 mostra a distribuição da população conforme graus de instrução, dividido entre população nascida nos Estados Unidos (em azul) e nascida em outros países, mas morando nos Estados Unidos (em cinza).

[14] No original: "*Adults classified as below level 1 may be considered functionally illiterate in English: i.e., unable to successfully determine the meaning of sentences, read relatively short texts to locate a single piece of information, or complete simple forms*."
[15] No original: "*functionally illiterate*".

Gráfico 2 – Grau de instrução da população nascida nos EUA e em outros países



Fonte: U.S. Census Bureau (2017, on-line).

Cerca de 35% da população possui *bachelor's degree* – que se assemelha ao nosso formato de graduação. Enquanto isso, mais de 10% dessa população possui *advanced degree*, grau de instrução superior ao *bachelor's degree* – que pode ser comparado aos cursos de pós-graduação, como especialização, Mestrado e Doutorado no Brasil. Além disso, o grau de instrução da população nascida fora dos Estados Unidos é bem similar ao daquela nascida nos Estados Unidos.

2.2 ANALFABETISMO E INSTRUÇÃO NO BRASIL

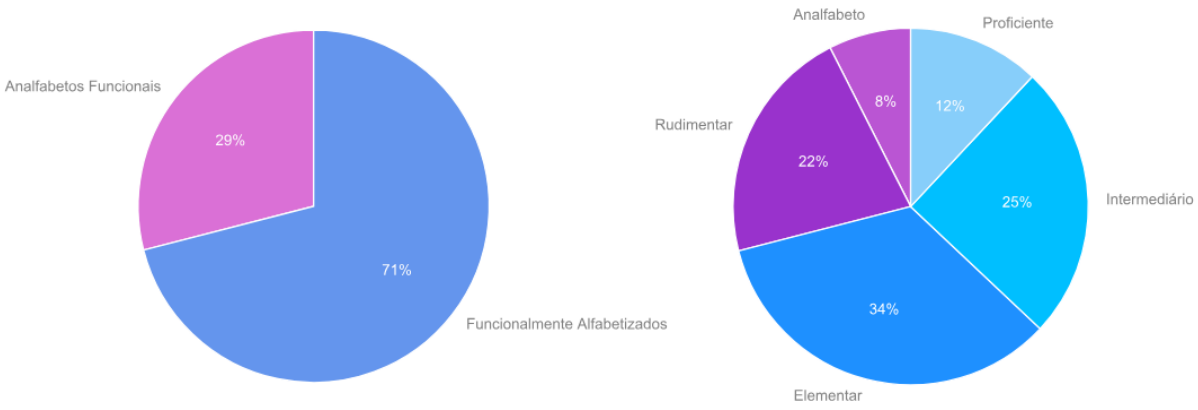
No Brasil, o Instituto Paulo Montenegro é responsável pelo levantamento do Indicador de Alfabetismo Funcional (INAF), cujo objetivo é mensurar o nível de alfabetização da população entre 15 e 64 anos. Esse indicador avalia as habilidades e práticas de leitura, escrita e matemática aplicadas ao cotidiano. O IBGE é responsável pela Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua. Esse levantamento é feito por meio de questionário, apresentando informações sobre as características básicas de educação de pessoas de 5 anos ou mais.

Os resultados obtidos com o último levantamento do INAF, no ano de 2018, demonstram que 29% da população brasileira é formada por analfabetos funcionais, e 71%, por pessoas funcionalmente alfabetizadas. Ainda, essas duas categorias se subdividem: analfabetos funcionais se dividem em analfabeto e rudimentar; e funcionalmente alfabetizados se dividem nas classificações elementar, intermediário e proficientes. Estas são as definições das categorias:

Analfabetos Funcionais
Analfabeto - Corresponde à condição dos que não conseguem realizar tarefas simples que envolvem a leitura de palavras e frases ainda que uma parcela destes consiga ler números familiares (números de telefone, preços etc.);
Rudimentar - Corresponde à capacidade de localizar uma informação explícita em textos curtos e familiares (como um anúncio ou um bilhete), ler e escrever números usuais e realizar operações simples, como manusear dinheiro para o pagamento de pequenas quantias ou fazer medidas de comprimento usando a fita métrica;
Funcionalmente Alfabetizados
Elementar - As pessoas classificadas neste nível podem ser consideradas funcionalmente alfabetizadas, pois já leem e compreendem textos de média extensão, localizam informações mesmo que seja necessário realizar pequenas inferências, resolvem problemas envolvendo operações na ordem dos milhares, resolvem problemas envolvendo uma sequência simples de operações e compreendem gráficos ou tabelas simples, em contextos usuais. Mostram, no entanto, limitações quando as operações requeridas envolvem maior número de elementos, etapas ou relações;
Intermediário - Localizam informações em diversos tipos de texto, resolvem problemas envolvendo percentagem ou proporções ou que requerem critérios de seleção de informações, elaboração e controle de etapas sucessivas para sua solução. As pessoas classificadas nesse nível interpretam e elaboram sínteses de textos diversos e reconhecem figuras de linguagem; no entanto, têm dificuldades para perceber e opinar sobre o posicionamento do autor de um texto.
Proficientes - Classificadas neste nível estão as pessoas cujas habilidades não mais impõem restrições para compreender e interpretar textos em situações usuais: leem textos de maior complexidade, analisando e relacionando suas partes, comparam e avaliam informações e distinguem fato de opinião. Quanto à matemática, interpretam tabelas e gráficos com mais de duas variáveis, compreendendo elementos como escala, tendências e projeções. (INSTITUTO PAULO MONTENEGRO, 2017, on-line).

Os gráficos a seguir demonstram, de forma visual, os resultados referentes à pesquisa do INAF de 2018. As fatias em cor roxa representam a população de analfabetos funcionais; já as fatias em cor azul representam a população funcionalmente alfabetizada.

Gráfico 3 – Taxas de alfabetização conforme a pesquisa do INAF de 2018

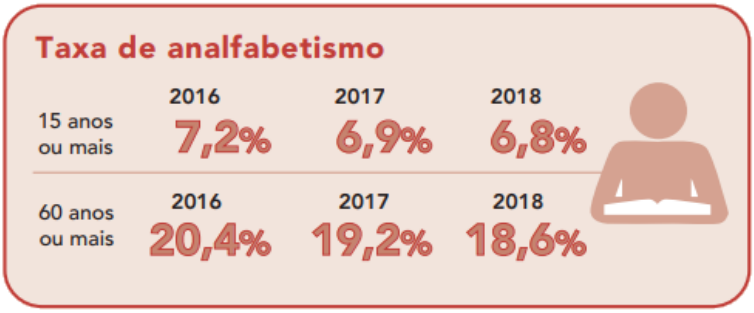


Fonte: elaborado pela autora com base em Instituto Paulo Montenegro (2018).

Constata-se, a partir dos gráficos, que a classificação de analfabetos funcionais se divide em 22% de rudimentares e 8% de analfabetos. Já a classificação de funcionalmente alfabetizados se divide em 34% da população como elementares, 25% intermediários e 12% proficientes.

O resultado da PNAD Contínua desse mesmo ano, publicado pelo IBGE em 2019, apresentou resultado similar. Conforme a pesquisa, no ano de 2018, no Brasil, 11,3 milhões de pessoas com 15 anos ou mais eram analfabetas, o que equivale a uma taxa de analfabetismo de 6,8% (IBGE, 2019). O Gráfico 4 apresenta a taxa de analfabetismo nos anos de 2016, 2017 e 2018.

Gráfico 4 – Taxa de analfabetismo conforme a PNAD Contínua de 2018

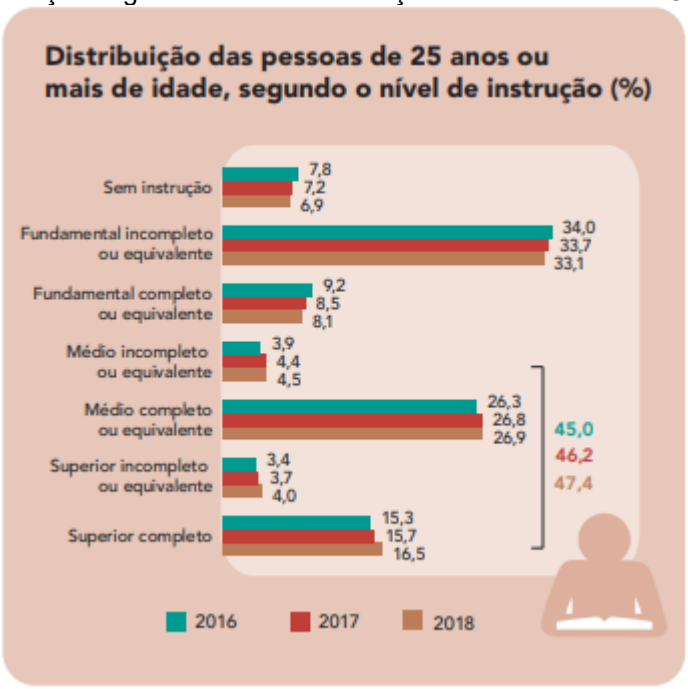


Fonte: IBGE (2019, p. 1).

Esse índice teve uma queda baixa no intervalo entre 2017 e 2018, em comparação com o intervalo entre 2016 e 2017. Em relação à população idosa, o índice também caiu menos levando-se em conta esses dois intervalos de tempo. Isso demonstra que, felizmente, a taxa de analfabetismo no país vem caindo, porém de forma lenta e gradual.

Outra informação importante levantada pela PNAD Contínua de 2018 foi a distribuição da população conforme o nível de instrução, que está detalhada no Gráfico 5.

Gráfico 5 – Distribuição segundo o nível de instrução conforme a PNAD Contínua de 2018



Fonte: IBGE (2019, p. 3).

Da população brasileira acima de 25 anos, quase 7% não possui instrução formal. A maior parcela da população acima dessa idade, aproximadamente 33%, possui Ensino Fundamental incompleto. A segunda maior parcela da população acima de 25 anos, equivalente a quase 27%, é composta por indivíduos com Ensino Médio completo. Vale ressaltar que esse gráfico comparativo confirma, por meio do contraste entre os anos de 2016, 2017 e 2018, que o grau de instrução da população vem, aos poucos, aumentando.

Ainda assim, é preciso considerar como maior parcela da população indivíduos que não possuem o Ensino Fundamental completo. Isso chama atenção quando em comparação com os dados apresentados sobre a população estadunidense. Enquanto lá a população com grau similar ao Ensino Médio (*high school*) é de 90%, aqui, menos de 27% da população possui o mesmo grau de instrução.

Neste capítulo foram apresentados os dados de analfabetismo e de instrução da população brasileira (Instituto Paulo Montenegro, 2018; IBGE, 2019) e da população estadunidense (U.S. CENSUS BUREAU, 2017; NATIONAL CENTER FOR EDUCATION STATISTICS, 2019). Esses dados são de extrema importância, pois é a partir deles que o nível de complexidade dos textos de divulgação deve ser avaliado,

considerando que o objetivo desses textos é atingir ao público geral, tanto nos Estados Unidos quanto no Brasil.

No próximo capítulo, a metodologia usada para desenvolver a pesquisa é relatada, apontando quais *softwares* foram utilizados para cada parte da análise.

3 METODOLOGIA

Neste capítulo, será detalhada a metodologia utilizada na dissertação. Em primeiro lugar, serão descritos os processos de compilação dos corpora de estudo, mostrando os resultados de número de textos, *types* e *tokens*. Posteriormente, serão apresentados *softwares* e ferramentas utilizados nas análises.

3.1 OS CORPORA DE ESTUDO

A seguir, será descrito o processo de compilação dos corpora especializados que são objetos de estudo desta análise. De acordo com Hunston (2002), um corpus especializado é um conjunto de textos de um tipo específico (como editoriais de jornais, livros didáticos, artigos acadêmicos etc.). Seu objetivo é ser representativo o suficiente para que, por meio dele, se investigue o uso da língua. Portanto, por serem compostos de materiais de divulgação, referentes a um domínio específico, os textos que fazem parte dos corpora de estudo são considerados especializados.

Em primeiro lugar, será apresentado o corpus paralelo, composto de originais e traduções, compilado a partir do site MedlinePlus (U.S. NATIONAL LIBRARY OF MEDICINE, 2020). Em seguida, será apresentado o subcorpus de textos originalmente escritos em língua portuguesa, compilado a partir da Biblioteca Virtual em Saúde (BRASIL, 2018), que compõe o corpus comparável juntamente ao subcorpus do MedlinePlus (PT).

3.1.1 Corpus de textos de divulgação originais e traduzidos

O MedlinePlus é uma página produzida pela U.S. National Library of Medicine (2020) com o intuito de disponibilizar informações sobre saúde para pacientes, seus familiares e amigos. Apresenta informações sobre doenças em diversas línguas, abrangendo sintomas e tratamentos, compondo um material que, de acordo com o site, é “de fácil leitura”[16]. No momento de compilação desse corpus, a página contava com traduções dos textos escritos originalmente em inglês para 60 idiomas, sendo o português brasileiro um deles.

[16] No original: “*Easy-to-Read Materials*”.

Os textos de divulgação disponibilizados pelo *site* são redigidos com foco em um público-alvo leigo, ou seja, que não possui formação médica ou na área da saúde. Além disso, esse material é produzido para leitores com proficiência limitada em língua inglesa (U.S. NATIONAL LIBRARY OF MEDICINE, 2020). O site possui um guia[17] para ser seguido durante a produção desses textos, descrevendo características a serem empregadas para desenvolver esse material. Dentre eles, é importante destacar os itens a seguir:

> O objetivo da página é, primordialmente, educacional e as informações dadas são neutras.
> As informações fornecidas são fáceis de entender, fáceis de explorar e bem organizadas. (U.S. NATIONAL LIBRARY OF MEDICINE, 2020, on-line)[18].

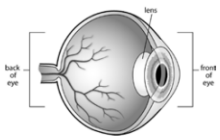
Na época da compilação, a plataforma contava com 66 textos em língua inglesa e suas respectivas traduções para a língua portuguesa, em formato PDF. As páginas contendo os textos originais e suas traduções são intercaladas, ou seja, o arquivo apresenta a página escrita originalmente em inglês, seguido desse conteúdo traduzido para o português, então apresenta a próxima página escrita em inglês, e sua tradução para o português, e assim sucessivamente – um exemplo de um trecho de texto pode ser observado na Figura 2. No Apêndice A, há a lista de arquivos que compõem esse corpus, com o título de cada um dos textos originais, seguidos da lista de arquivos e dos títulos de suas traduções. Vale ressaltar que esse conjunto de textos compõe um corpus paralelo, ou seja, são textos-fonte alinhados a suas traduções (FANTINUOLI; ZANETTIN, 2015).

[17] Disponível em: https://medlineplus.gov/languages/criteria.html. Acesso em: 23 ago. 2019.
[18] No original: "*The primary purpose of the website is educational, and the information is unbiased. The information provided is easy to understand, easy to navigate, and well organized*."

Figura 2 – Exemplo de trecho nos arquivos em PDF do MedlinePlus

Cataract

A cataract is the clouding of the lens of the eye that makes it hard to see. Cataracts can affect one or both eyes. Cataracts are common in older people.



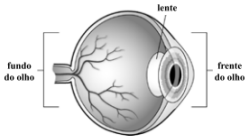
Risk Factors of a Cataract

The risk of a cataract increases with age. Other risk factors include:

- Some diseases such as diabetes
- Smoking
- Alcohol use
- Prolonged exposure to sunlight

Catarata

Catarata é o processo de opacificação do cristalino, que prejudica a visão. A catarata pode acometer um ou os dois olhos. Trata-se de um processo comum em pessoas idosas.



Fatores de risco da catarata

O risco de catarata aumenta com a idade. Outros fatores de risco:

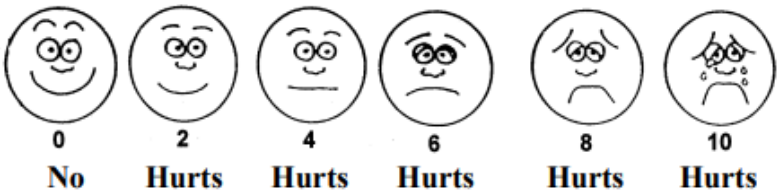
- Algumas doenças, como o diabetes
- Fumo
- Ingestão de bebidas alcoólicas
- Exposição prolongada à luz solar

Fonte: U.S. National Library of Medicine (2020, on-line).

Os textos apresentam títulos separando cada grupo de informações, como por exemplo ‘Sinais’ (ou ‘Sintomas’) e ‘Cuidados necessários’. Diversos textos incluem tabelas ou figuras, como o exemplo abaixo, do texto *Your Hospital Care after Surgery*. A Figura 3 apresenta uma escala de mensuração da dor, que vai de 0, representando a ausência de dor, a 10, que se refere a uma dor intensa e insuportável.

Figura 3 – Exemplo de figura presente nos textos

Wong-Baker FACES Pain Rating Scale



0	2	4	6	8	10
No Hurt	Hurts Little Bit	Hurts Little More	Hurts Even More	Hurts Whole Lot	Hurts Worst

From Wong, D.L., Hockenberry-Eaton, M., Wilson, D., Winkelstein, M.L., Ahmann, E., DiVito-Thomas, P.A. Whaley and Wong's Nursing Care of Infants and Children, ed. 6, St. Louis, 1999, p. 2040. Copyrighted by Mosby, Inc. Reprinted by permission.

Fonte: U.S. National Library of Medicine (2020, on-line).

Os textos presentes nas tabelas ou ao redor de figuras foram desconsiderados no processo de compilação, por se tratarem, muitas vezes, de termos ou palavras

soltas, que poderiam causar ruído textual no corpus e, consequentemente, problemas no que diz respeito aos métodos utilizados no trabalho. Por exemplo, deixar uma palavra “isolada” de texto influenciaria o cálculo do Índice Flesch (que será detalhado na seção 2.2.1.1), podendo fazer com que o resultado sugerisse maior nível de dificuldade do texto. Pelo mesmo motivo da exclusão das tabelas e textos das figuras, mensagens repetidas diversas vezes, como no exemplo abaixo, foram desconsideradas no momento de compilação:

> *Talk to your doctor or nurse if you have any questions or concerns.*
> [Fale com o seu médico ou enfermeiro para sanar quaisquer dúvidas ou preocupações.] (U.S. NATIONAL LIBRARY OF MEDICINE, 2020, on-line).

Os números totais do corpus estão apresentados na Tabela 1.

Tabela 1 – Informações do corpus paralelo do MedlinePlus

Subcorpus	**EN**	**PT**
Textos	66	66
Tokens	34 765	39 476
Types	3 088	4 554

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

Os 66 textos em língua inglesa somam aproximadamente 35 mil palavras (*tokens*) no total, das quais cerca de 3 mil palavras são diferentes (*types*). Já os 66 textos em língua portuguesa somam aproximadamente 40 mil *tokens* e quase 5 mil *types*.

A seguir, será apresentado o subcorpus formado por textos escritos originalmente em língua portuguesa, que compõe o corpus comparável juntamente ao subcorpus de textos traduzidos para o português, que foi descrita acima.

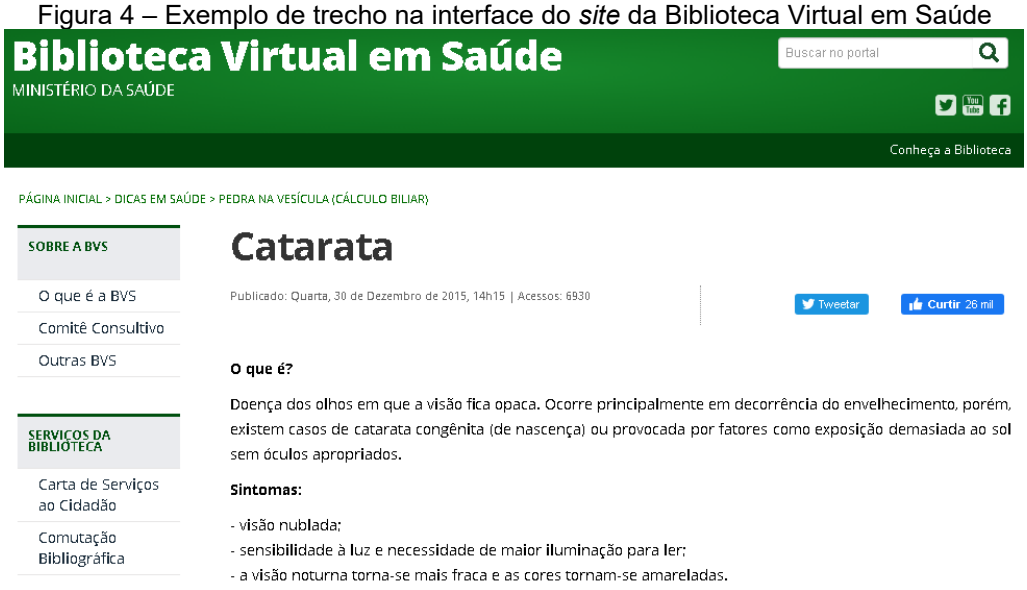
3.1.2 Subcorpus de textos de divulgação escritos originalmente em português

O Ministério da Saúde (BRASIL, 2018) criou, em 1953, uma biblioteca especializada na área da saúde. Posteriormente, no ano de 2001, essa biblioteca foi disponibilizada on-line, para facilitar o acesso da comunidade geral. Na Biblioteca

Virtual em Saúde[19], podem-se encontrar informações importantes acerca de doenças, tais como seus sintomas, tratamentos e cuidados. Esse material tem como objetivo disseminar informações em saúde, oferecendo consulta a qualquer cidadão. Portanto, a biblioteca se propõe a ser um instrumento acessível de informação.

O produto que compõe o subcorpus é denominado Dicas de Saúde, pois observou-se que esse era o que mais compartilhava características com os textos do MedlinePlus, configurando, junto ao MedlinePlus (PT), um corpus comparável (FANTINUOLI; ZANETTIN, 2015). Em suma, são textos de divulgação, escritos de forma curta e objetiva, tratando de assuntos gerais sobre saúde. Além disso, supõe-se que façam uso de linguagem acessível, por ter como objetivo auxiliar a população geral em relação aos temas de que o guia trata. Um exemplo de texto disponível na página pode ser observado na Figura 4.

Figura 4 – Exemplo de trecho na interface do *site* da Biblioteca Virtual em Saúde



Fonte: Brasil (2018).

No momento de compilação do subcorpus, esse produto era composto de 191 textos, havendo dentre eles páginas informativas sobre doenças, acidentes, alimentação e dúvidas frequentes. No Apêndice B, há a lista de arquivos que compõem esse subcorpus, com o título de cada um dos textos.

Durante a compilação, foi feita uma limpeza nos textos. As primeiras informações excluídas foram a data e o horário das publicações e o seu número de

[19] Disponível em: http://brasil.bvs.br/. Acesso em: 23 ago. 2019.

acessos, como no exemplo do texto *Acidente Vascular Cerebral (AVC)*: “Publicado: Quarta, 30 de Dezembro de 2015, 13h00 | Acessos: 6789”. O segundo item excluído foram tabelas e demonstração de cálculos, por acreditarmos que esse tipo de informação causaria ruídos no corpus já que não é apresentada em formato de texto. Conforme a Figura 5, essas tabelas apresentavam palavras soltas (muitas vezes termos) ou listas, seguidas de informações de índices, como valores ou porcentagens. Como mencionado na descrição da compilação de textos do MedlinePlus, deixar uma palavra “isolada” de texto poderia causar influência no cálculo do Índice Flesch.

Figura 5 – Tabela extraída do texto *Alimentação saudável*

IMC (Kg/m²)	Estado Nutricional
Menor que 18,5	Você está com baixo peso
18,5 a 24,99	O seu peso está adequado
25 a 29,99	Alerta: sobrepeso
Maior que 30	Alerta: obesidade

Fonte: Brasil (2018, on-line).

Como o interesse no presente trabalho é analisar o texto, e não fazer a extração de termos, optou-se por retirar do corpus as informações das tabelas. Além desses dois itens, uma mensagem final presente em todos os textos foi retirada:

> IMPORTANTE: Somente médicos e cirurgiões-dentistas devidamente habilitados podem diagnosticar doenças, indicar tratamentos e receitar remédios. As informações disponíveis em Dicas em Saúde possuem apenas caráter educativo. (BRASIL, 2018, on-line).

Pelo mesmo motivo da exclusão das tabelas, a repetição dessa mensagem em todos os textos do corpus causaria um ruído, além de aumentar o número total de *tokens* do subcorpus em quase cinco mil. Como apresentado na Tabela 2, esses 191 textos somam pouco mais de 84 mil palavras *tokens*, com quase 10 mil *types*.

Tabela 2 – Informações do subcorpus Dicas em Saúde do Ministério da Saúde

Subcorpus	**PT**
Textos	191
Tokens	84.085
Types	9.666

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

Na próxima seção, será explicada a aplicação de cada uma das ferramentas de análise. Além disso, serão detalhadas as configurações aplicadas nos levantamentos dos dados.

3.2 AS FERRAMENTAS DE ANÁLISE

A seguir, serão apresentadas as ferramentas Coh-Metrix e Coh-Metrix-Port, utilizadas para desenvolver a análise da complexidade textual dos corpora. Após, serão descritas as ferramentas do *software* AntConc, utilizado para analisar textualmente os corpora de estudo. Para relatar a compilação do corpus de referência em inglês, será descrito o funcionamento do *software* AntCorGen. Por fim, serão apresentados os *softwares* utilizados para alinhar e explorar o corpus paralelo, o LF Aligner e o AntPConc, respectivamente.

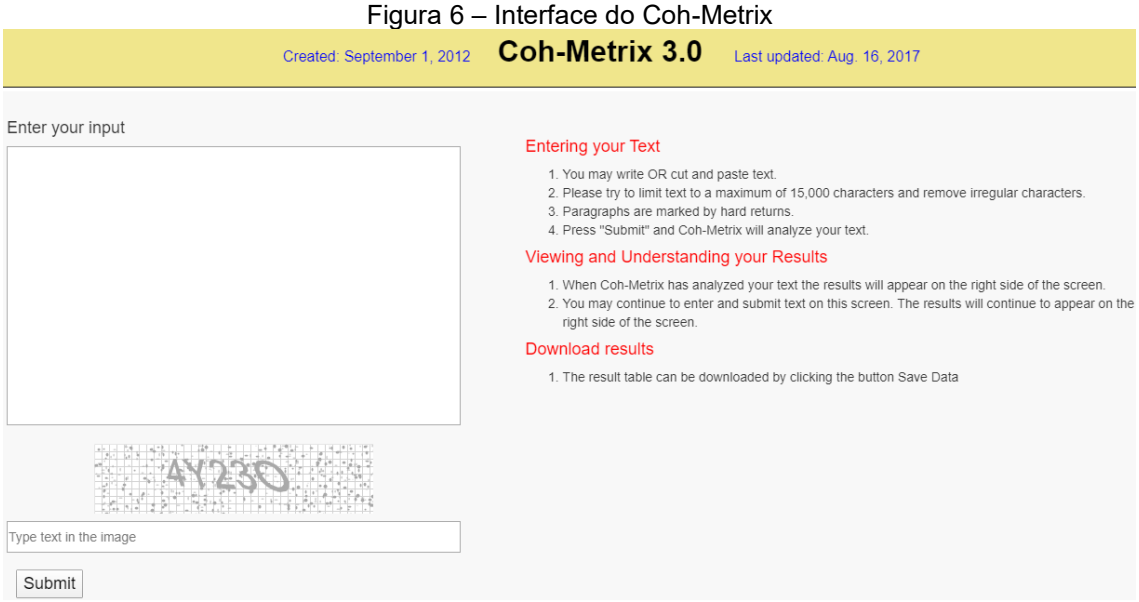
3.2.1 Coh-Metrix e Coh-Metrix-Port: análise de inteligibilidade

O Coh-Metrix (GRAESSER *et al*., 2017), desenvolvido por pesquisadores da Universidade de Memphis, foi disponibilizado no ano de 2012, com a finalidade de estimar, numericamente, a coesão e o nível de dificuldade de leitura de textos em língua inglesa com base nos níveis lexical, sintático, discursivo e conceitual. Seu nome, Coh-Metrix, vem de *cohesion metrics* (métricas de coesão), e a definição utilizada é de que a coesão "consiste em características do texto explícito que desempenham alguma função que ajude o leitor a conectar, mentalmente, as ideias no texto" (GRAESSER *et al*., 2017, on-line)[20].

[20] No original: "*Our definition of cohesion consists of characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text*."

Essa ferramenta está disponível on-line gratuitamente[21] e conta com 108 índices diferentes, desenvolvidos com base em recursos e ferramentas de Processamento de Linguagem Natural. Os índices são categorizados em 11 grupos, sendo eles: (1) *Descriptive*; (2) *Text Easability Principal Component Scores*; (3) *Referential Cohesion*; (4) *Latent Semantic Analysis*; (5) *Lexical Diversity*; (6) *Connectives*; (7) *Situation Model*; (8) *Syntactic Complexity*; (9) *Syntactic Pattern Density*; (10) *Word Information*; e (11) *Readability*. Na Figura 6, pode-se observar a página inicial da ferramenta.

Figura 6 – Interface do Coh-Metrix



Fonte: Coh-Metrix (GRAESSER *et al*., 2017).

Para utilizá-la, basta inserir o texto a ser processado pela ferramenta e digitar a sequência de caracteres mostrada na imagem. Então a ferramenta retornará, à direita do texto, os resultados das 108 métricas. As métricas e suas categorizações estão presentes no Anexo A.

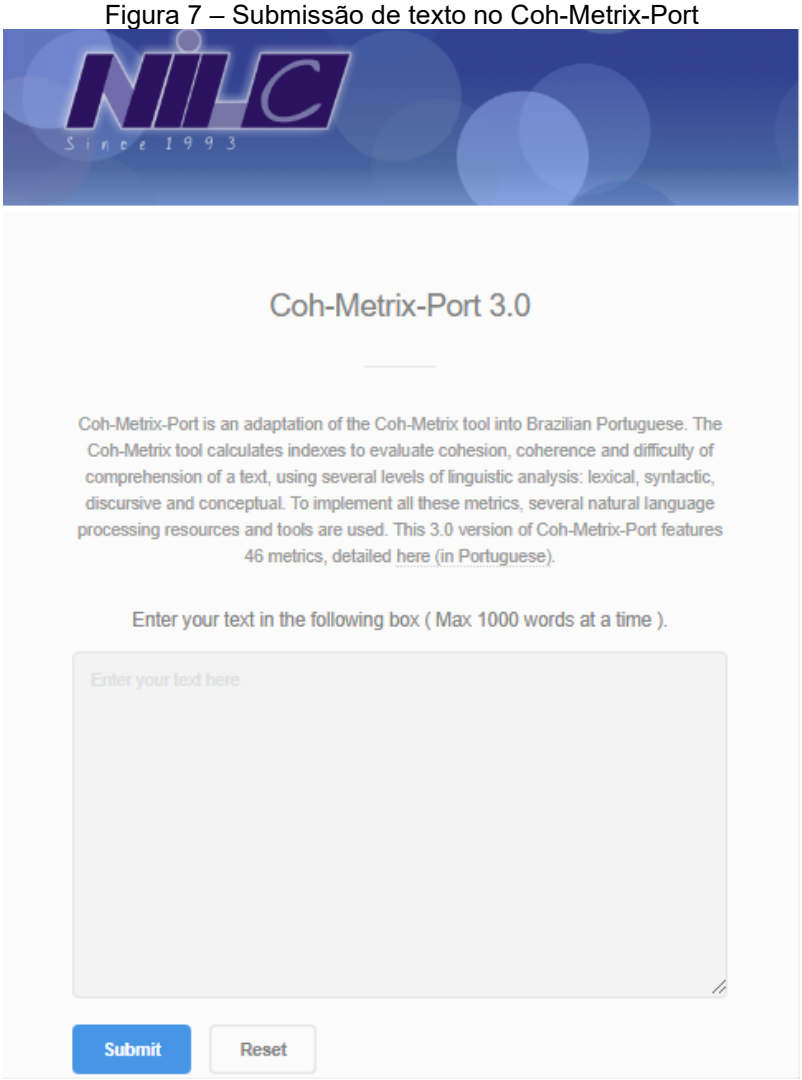
O Coh-Metrix-Port (NILC, 2020) é uma adaptação do Coh-Metrix para a língua portuguesa. Trata-se de uma ferramenta on-line e gratuita[22] que possui o mesmo objetivo da ferramenta desenvolvida para a língua inglesa, porém com um número reduzido de métricas disponíveis para a avaliação dos textos (SCARTON; ALMEIDA; ALUÍSIO, 2009). O Coh-Metrix-Port conta com apenas 46 métricas adaptadas para o

[21] Disponível em: http://cohmetrix.com/. Acesso em: 7 jan. 2020.
[22] Disponível em: http://fw.nilc.icmc.usp.br:23380/cohmetrixport. Acesso em: 3 fev. 2020.

português brasileiro, categorizadas em 10 grupos: (1) *Basic Counts*; (2) *Logic operators*; (3) *Content word frequencies*; (4) *Hypernyms*; (5) *Tokens*; (6) *Constituents*; (7) *Connectives*; (8) *Ambiguity*; (9) *Coreference*; e (10) *Anaphoras*.

De acordo com Scarton, Almeida e Aluísio (2009), a ferramenta foi adaptada para a língua portuguesa com o propósito de "colaborar com a inclusão social no âmbito do direito ao acesso à informação" (p. 8). A adaptação foi desenvolvida por pesquisadores do Núcleo Interinstitucional de Linguística Computacional, da Universidade de São Paulo. A interface da versão 3.0 da ferramenta é bastante similar à do Coh-Metrix (Figura 7).

Figura 7 – Submissão de texto no Coh-Metrix-Port



Fonte: Coh-Metrix-Port (NILC, 2020).

Não são todas as 46 métricas do Coh-Metrix-Port que possuem um equivalente dentre as 108 métricas do Coh-Metrix (a lista completa de métricas disponíveis no

Coh-Metrix-Port está presente no Anexo B). De acordo com Pasqualini (2012), de todas as métricas disponíveis para a língua portuguesa, apenas 31 delas são equivalentes às disponíveis para a língua inglesa. Como o objetivo aqui é tratar de inteligibilidade, ater-nos-emos à métrica que as duas ferramentas possuem em comum nessa esfera: o Índice Flesch.

3.2.1.1 Índice Flesch

Rudolf Flesch, austríaco naturalizado estadunidense, foi um refugiado da Segunda Guerra Mundial nos Estados Unidos e se tornou acadêmico na Universidade de Columbia. Como apontado na seção 1.2.1, esse pesquisador possuía como objetivo promover o uso de uma linguagem mais clara, simples e acessível à população geral, independentemente do grau de escolaridade. No ano de 1943, ele foi o pesquisador responsável pelo desenvolvimento do Índice Flesch, ou *Flesch Reading Ease*, uma das fórmulas mais antigas e mais utilizadas para estimar a inteligibilidade de textos.

O índice, desenvolvido originalmente para o inglês, leva em consideração duas variáveis: o comprimento médio das frases, representado por ASL (*Average Sentence Length*), que é o número de palavras do texto dividido pelo número total de sentenças; e a média de sílabas por palavras, representado por ASW (*Average of Syllables per Word*), resultado do número total de sílabas dividido pelo número de palavras do texto. A fórmula está descrita a seguir:

$$FRE = 206{,}835 - (1{,}015 \times ASL) - (84{,}6 \times ASW)$$

O Índice Flesch foi adaptado para a língua portuguesa no ano de 1996 por pesquisadores do Instituto de Ciências Matemáticas e de Computação (MARTINS *et al.*, 1996), da Universidade de São Paulo, que tinham como objetivo adequá-lo à realidade das palavras e sílabas da língua escrita em português do Brasil, já que seus comprimentos diferem bastante daqueles em língua inglesa. Além de as palavras em português serem mais longas, também as frases são mais compridas, quando contrastamos com textos de mesmo gênero escritos em inglês. A fim de ilustrar essa diferença, comparamos os textos sobre asma dos subcorpora originalmente escritos

nas línguas. Em português, a média de palavras por sentença é de 17,464; já a média de sílabas por palavra é de 3,056. Em inglês, a média de palavras por sentença é de 7,649; e a média de sílabas por palavra é de 1,424.

Para a fórmula no português, será utilizada, para representar o número de palavras por frase, a sigla PPF; e para o número de sílabas por palavra, a sigla SPP. A fórmula para a língua portuguesa pode ser expressa por:

$$IF = 248{,}835 - (1{,}015 \times PPF) - (84{,}6 \times SPP)$$

Como resultado do cálculo, obtém-se um índice que pode ir de 0 a 100, sendo que quanto mais próximo de 0, mais difícil seria o texto; e quanto mais perto de 100, mais fácil. Essa é a primeira e – ainda é – a única métrica de inteligibilidade que foi adaptada para o português brasileiro. Contudo, essa métrica é considerada superficial porque mede características "rasas" do texto, uma vez que conta apenas número de palavras, de frases e de sílabas (SCARTON; ALMEIDA; ALUÍSIO, 2009).

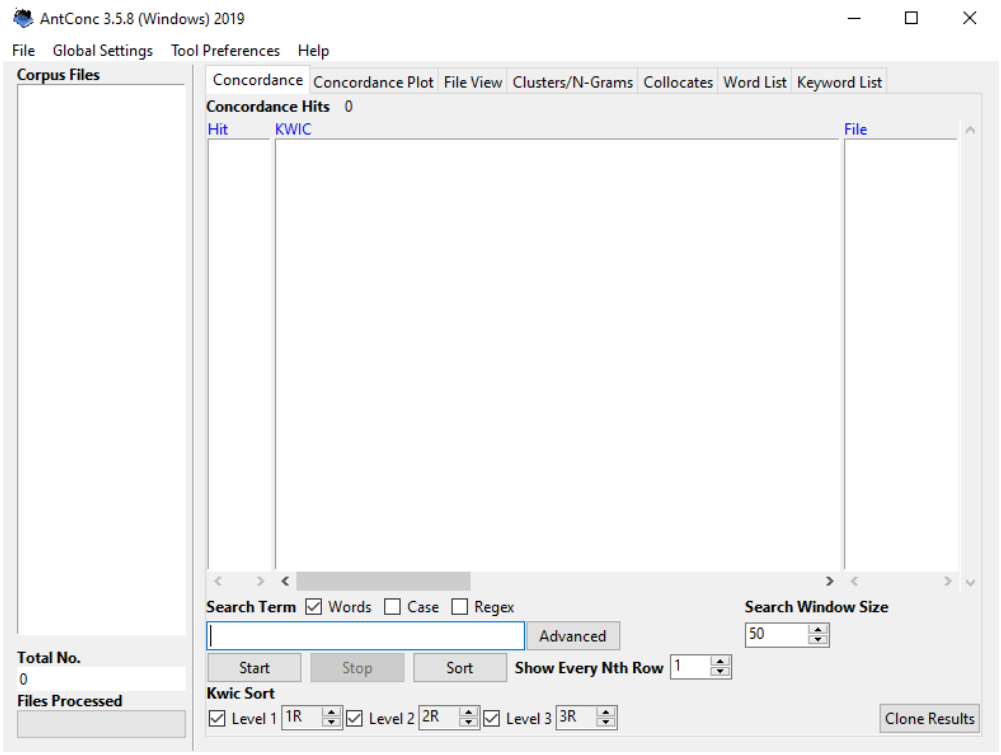
Para os fins da análise quantitativa, todos os textos que compõem o corpus do MedlinePlus (66 textos em inglês e 66 em português) foram inseridos na ferramenta. Devido à diferença no número de textos entre o corpus paralelo e os textos originalmente escritos em português, do Ministério da Saúde (191), para esse levantamento, foi utilizada apenas uma amostra dos textos do subcorpus do Ministério da Saúde (ou seja, 66) para equipará-lo ao corpus do MedlinePlus. Portanto, essa amostra corresponde a pouco mais de $^1/_3$ do subcorpus.

Há outras informações que seriam relevantes para ajudar a estimar a dificuldade de leitura de um texto. Por exemplo, número de palavras diferentes, frequência de uso das palavras e sua regularidade ou irregularidade seriam aspectos interessantes de serem apontados em um cálculo de inteligibilidade. Apesar de não contar com tais parâmetros, essa fórmula possui grande índice de confiabilidade e é utilizada em diversas pesquisas sobre a língua portuguesa. Porém, ela deve ser entendida apenas como ponto de partida, não sendo o único referente no que diz respeito a essa análise. Por isso, além de utilizarmos como base essa medida quantitativa, também partiremos para a análise textual, a fim de descobrir se o que foi indicado pela ferramenta de fato se comprova.

3.2.2 AntConc e a análise textual

Para proceder com a análise desses corpora de estudo, o *software* utilizado para o processamento dos textos foi o AntConc (ANTHONY, 2019), um *software* gratuito[23], que pode ser baixado e utilizado off-line. A Figura 8 mostra sua tela inicial.

Figura 8 – Interface do AntConc



Fonte: AntConc (ANTHONY, 2019).

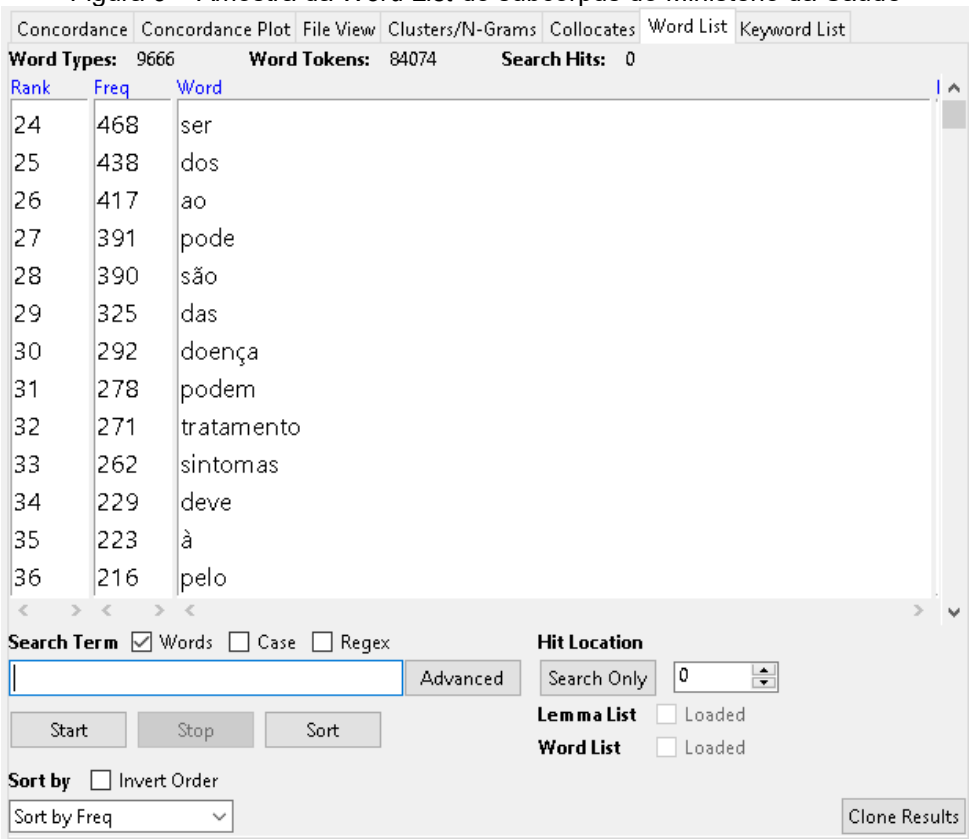
No AntConc, há sete ferramentas, que oferecem diferentes possibilidades de análise de corpus, baseando-se em índices estatísticos. Como se pode depreender da figura, são elas: (1) *Concordance*; (2) *Concordance Plot*; (3) *File View*; (4) *Clusters/N-grams*; (5) *Collocates*; (6) *Word List*; e (7) *Keyword List*. A seguir, será descrito o funcionamento das ferramentas utilizadas neste trabalho.

[23] Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 10 jan. 2020.

3.2.2.1 Word List

A ferramenta *Word List* foi utilizada já no capítulo de Metodologia, para anotar o número de *types* e *tokens* dos corpora. Na Figura 9, é possível observar um exemplo de lista de palavras (*Word List*), em ordem decrescente de frequência, gerada pelo AntConc a partir do subcorpus de textos do Ministério da Saúde (BRASIL, 2018).

Figura 9 – Amostra da *Word List* do subcorpus do Ministério da Saúde



Fonte: AntConc (ANTHONY, 2019).

É possível escolher se as listas geradas serão apresentadas em ordem de frequência, em ordem alfabética ou conforme a última letra da palavra. Visto que o subcorpus tem 9.666 *types* (palavras diferentes), esse é o número de palavras gerado nessa lista.

Conforme observado na maioria das análises textuais, as primeiras palavras da lista são gramaticais, como preposições (‘para’, ‘de’, ‘com’, ‘a’ etc.) e artigos (‘o’, ‘a’, ‘um’, ‘uma’ e seus plurais). A 30ª palavra dessa lista é a primeira palavra de conteúdo, o substantivo ‘doença’.

3.2.2.2 Collocates

A ferramenta *Collocates* é utilizada para fazer o levantamento de colocados de uma palavra de busca (*node*). A colocação é uma relação de habitual coocorrência de palavras, ou seja, palavras que tendem a ocorrer próximas em uma janela específica com frequência estatisticamente relevante em determinado corpus de estudo.

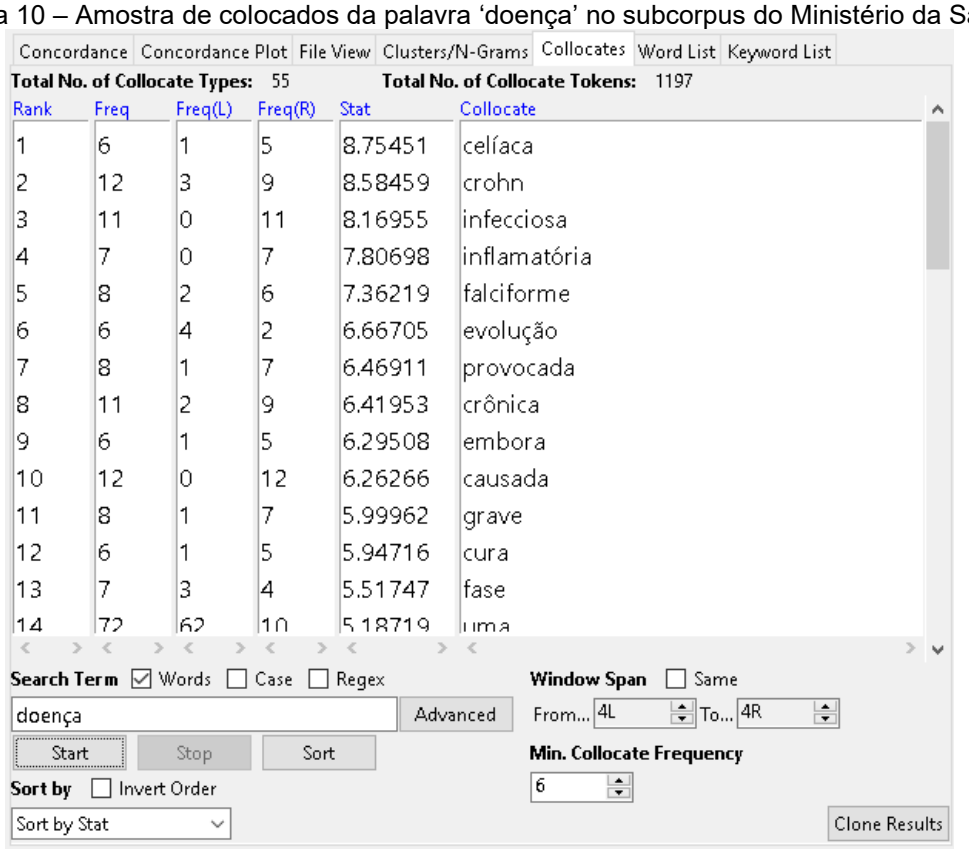
Para que seja feito o levantamento de colocados, além de selecionar uma palavra de busca, é necessário estabelecer o intervalo em que seus colocados deverão ocorrer – esse intervalo é chamado de janela (*window*). O AntConc faz apenas levantamento de palavras, desconsiderando pontuações ou quebras de parágrafo, ou seja, mesmo que a palavra esteja em outro segmento, se estiver na janela estabelecida, ela será considerada colocado. Por isso, faz-se necessária a análise manual para que se confirme a colocação.

Será aplicada a janela de busca de quatro palavras à direita e quatro palavras à esquerda, que é o número geralmente utilizado (BERBER SARDINHA, 2004). Acredita-se que, utilizando essa janela, seja possível respeitar os limites dos segmentos.

Para determinar as palavras que se configuram como colocados do termo de busca, a ferramenta oferece duas medidas estatísticas: *Mutual Information* e *T-Score*. O cálculo de *Mutual Information* compara a frequência de coocorrência da palavra de busca e do colocado com suas frequências de forma isolada. Já o *T-Score* considera a frequência absoluta de ocorrência dessas palavras (STUBBS, 1995). Para os fins deste trabalho, será utilizada a estatística de *Mutual Information*, pois ela leva em consideração a probabilidade de que as palavras ocorram juntas ao comparar com seus números totais de ocorrências.

Pode-se observar, na Figura 10, um exemplo de levantamento de colocados utilizando a palavra de busca ‘doença’. O subcorpus utilizado para o levantamento é o do Ministério da Saúde, sendo estabelecida uma frequência mínima de 6 ocorrências para uma palavra ser considerada colocado.

Figura 10 – Amostra de colocados da palavra ‘doença’ no subcorpus do Ministério da Saúde



Fonte: AntConc (ANTHONY, 2019).

3.2.2.3 Keyword List

A ferramenta *Keyword List* faz o levantamento de palavras-chave do texto. Para a Linguística de Corpus, diferentemente do conceito de palavra-chave mais comumente conhecido – aquelas palavras que o autor seleciona como descritoras dos temas principais abordados em um texto –, o termo tem outro significado. Palavras-chave, na perspectiva da Linguística de Corpus, são aquelas cuja frequência é estatisticamente relevante, em comparação com algum referente. Ou seja, para fazer o levantamento de palavras-chave é necessário utilizar um corpus de referência.

O resultado do levantamento é uma lista de palavras cujas frequências são estatisticamente mais elevadas no corpus de estudo em relação ao escolhido como referência. Pelo fato de palavras gramaticais serem frequentes em qualquer gênero textual, a tendência é que essas sejam neutralizadas no levantamento de palavras-chave, portanto, não aparecendo na lista.

Assim como na lista de palavras, também é possível organizar as palavras-chave conforme frequência, ordem alfabética ou da última letra da palavra. Adicionalmente, por se tratar de um levantamento estatístico, é possível, também,

organizar a lista conforme dois índices, que são utilizados para calcular as palavras-chave: *Keyness* (chavicidade) e *Effect* (efeito).

Pojanapunya e Todd (2018) listam dois tipos de medidas estatísticas para calcular o quão chave uma palavra é: testes de significância (*Keyness*, no AntConc), que investigam se há diferença no uso de uma palavra em dois ou mais corpora; e estatísticas de tamanho do efeito (*Effect*, no AntConc)[24], que investiga o tamanho dessa diferença de uso.

O *log-likelihood* é a medida de significância que será utilizada nesta pesquisa. O valor do cálculo de *log-likelihood* é alto quando há grande disparidade de frequência de uma palavra entre o corpus de estudo e o corpus de referência. Ou seja, quanto mais significativamente elevada for a frequência relativa de uma palavra em um corpus de estudo, maior será seu valor de *log-likelihood*.

No AntConc, esses valores são utilizados para organizar as palavras conforme *keyness*. Para esse teste, é utilizado um valor de p, que corresponde à probabilidade de um efeito ocorrer devido ao acaso. A ferramenta traz como uma das opções $p < 0{,}0001$, que será utilizado no levantamento. Esse índice admite uma margem de erro de 0,01%. De acordo com Brezina (2018), para pesquisas na área de Ciências Humanas, é aceitável utilizar até $p < 0{,}05$.

A medida de tamanho do efeito que será utilizada será *odds ratio*. Essa medida indica a relação entre as ocorrências de uma palavra no corpus de estudo e no corpus de referência. Os resultados podem variar de zero a infinito, sendo que o valor 1 aponta que não há diferença de uso, e valores perto de zero ou muito acima de 1 apontam diferença na frequência de uso.

Gabrielatos (2018) afirma que o nível de chavicidade de um item precisa ser estabelecido usando-se uma combinação de duas métricas que se complementam, já que o teste *log-likelihood* enfatiza palavras relativamente comuns que servem ao propósito de pesquisas orientadas pelo gênero, enquanto o *odds ratio* dá ênfase às palavras mais especializadas, que são mais adequadas para pesquisas voltadas à análise do discurso, por exemplo (POJANAPUNYA; TODD, 2018). Assim, serão utilizadas essas duas medidas para fazer o levantamento de palavras-chave dos corpora. Por esta ser uma pesquisa orientada pelo gênero, a organização dos resultados será feita conforme *log-likelihood* (*Keyness*, no AntConc) em ordem

[24] No original: "*significance test*" e "*effect size statistics*".

decrescente. O *odds ratio* será aplicado para estabelecer o ponto de corte para as palavras serem ou não consideradas chave.

A escolha do corpus de referência é determinante e deve ser feita levando-se em consideração a finalidade da pesquisa. Portanto, não se pode tratar de uma escolha aleatória, baseada apenas na ideia de que o corpus de referência deve ser n vezes maior que o corpus de estudo, como se costumava pensar. De acordo com Berber Sardinha (2004), os tamanhos críticos de corpora de referência são de 2, 3 e 5 vezes o tamanho do corpus de estudo, pois retornam um número de palavras-chave relativamente mais elevado do que corpora de tamanhos menores. Em razão disso, foi estabelecido que o corpus de referência fosse cerca de 5 vezes o tamanho do maior corpus de estudo.

Considerando o objetivo de discutir a acessibilidade dos textos que compõem os corpora de estudo, optou-se pela utilização de artigos da área de Medicina como referência. Acredita-se que, assim, o que a ferramenta apontará como palavras-chave será aquilo que é característico especificamente desse gênero de textos – o texto de divulgação –, e não necessariamente os termos médicos. Para tanto, dois corpora de referência foram compilados: um em língua portuguesa e outro em língua inglesa.

3.2.2.3.1 O corpus de referência em português

O corpus de referência em língua portuguesa foi compilado manualmente, selecionando-se artigos da área de Medicina no Portal de Periódicos Capes[25]. O formato em que os textos foram salvos foi txt – tal como os outros corpora. Para acessar o banco de dados de artigos, foi utilizado o acesso remoto via Comunidade Acadêmica Federada (CAFe). A pesquisa foi feita, inicialmente, partindo-se do termo de busca ‘medicina’. Posteriormente, foram acrescentados os filtros: artigo, como tipo de texto; e português, como a língua do texto. A partir desse método, foram compilados 117 artigos em português da área de Medicina. Os números que descrevem esse corpus estão contemplados na Tabela 3.

[25] Disponível em: https://www.periodicos.capes.gov.br/. Acesso em: 10 jan. 2020.

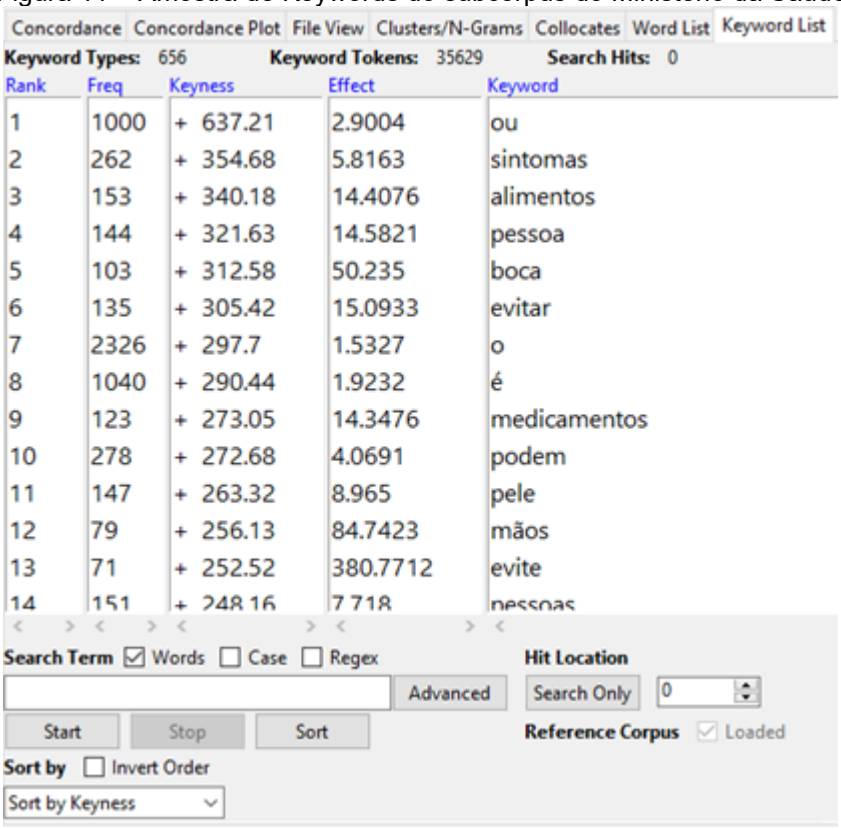
Tabela 3 – Informações do corpus de referência de artigos em português

Corpus	**PT**
Textos	117
Tokens	450.507
Types	32.799

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

Assim, a partir do contraste entre o corpus de estudo e o corpus de referência, é gerada uma lista de palavras-chave. A Figura 11 apresenta uma amostra dessa lista do subcorpus do Ministério da Saúde em contraste com o corpus de artigos científicos da área de Medicina escritos em português.

Figura 11 – Amostra de Keywords do subcorpus do Ministério da Saúde



Fonte: AntConc (ANTHONY, 2019).

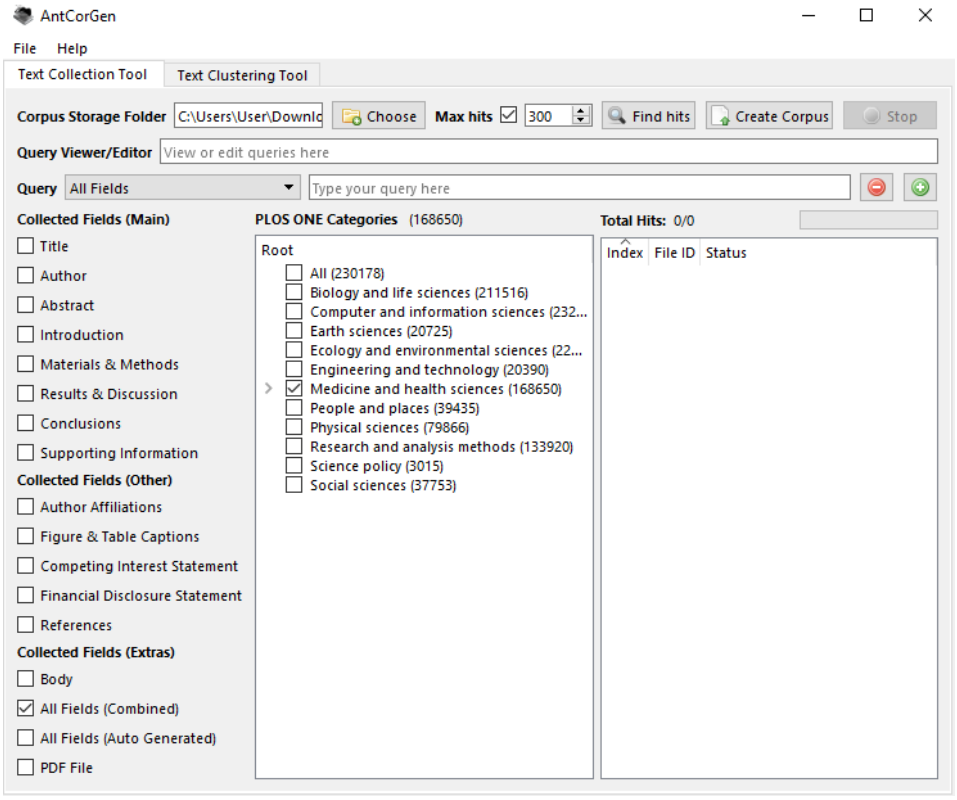
Como é possível observar na figura, o levantamento de palavras-chave lista as palavras que destacam nos textos de divulgação quando comparados com os artigos científicos, sendo ambos os gêneros da mesma área de especialidade, a Medicina. Por exemplo, no contraste com artigos científicos, as formas verbais ‘evitar’ e ‘evite’

se destacaram como estatisticamente relevantes nos textos de divulgação, pois devem se referir a orientações do que fazer em determinadas situações.

3.2.2.3.2 O corpus de referência em inglês

O corpus de referência em língua inglesa foi compilado por meio do *software* AntCorGen (ANTHONY, 2019b), que permite a compilação automática de artigos científicos de diferentes áreas (e subáreas) do conhecimento. Para que seja feita a compilação, é necessário selecionar as seções – *Abstract* [Resumo], *Introduction* [Introdução], *Materials and Methods* [Metodologia] etc. – dos artigos que serão coletados, ou selecionar *All Fields (Combined)* para que os artigos venham completos. Após, basta selecionar as áreas (ou subáreas) que integrarão seu corpus. Em *Max hits*, é possível, também, escolher um número máximo de textos para esse corpus. Para compilar o corpus de referência em inglês, então, foi selecionada a área *Medicine and health sciences* (Medicina e ciências da saúde), determinando como máximo de textos 300. A Figura 12 apresenta a seleção desses critérios.

Figura 12 – Interface do AntCorGen



Fonte: AntCorGen (ANTHONY, 2019b).

Posteriormente, esse corpus foi balanceado em relação ao número de *tokens* do corpus de referência em língua portuguesa, ficando, então, com 84 textos. Na Tabela 4, estão as informações do corpus de referência em inglês.

Tabela 4 – Informações do corpus de referência de artigos em inglês

Corpus	**EN**
Textos	84
Tokens	460.170
Types	23.785

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

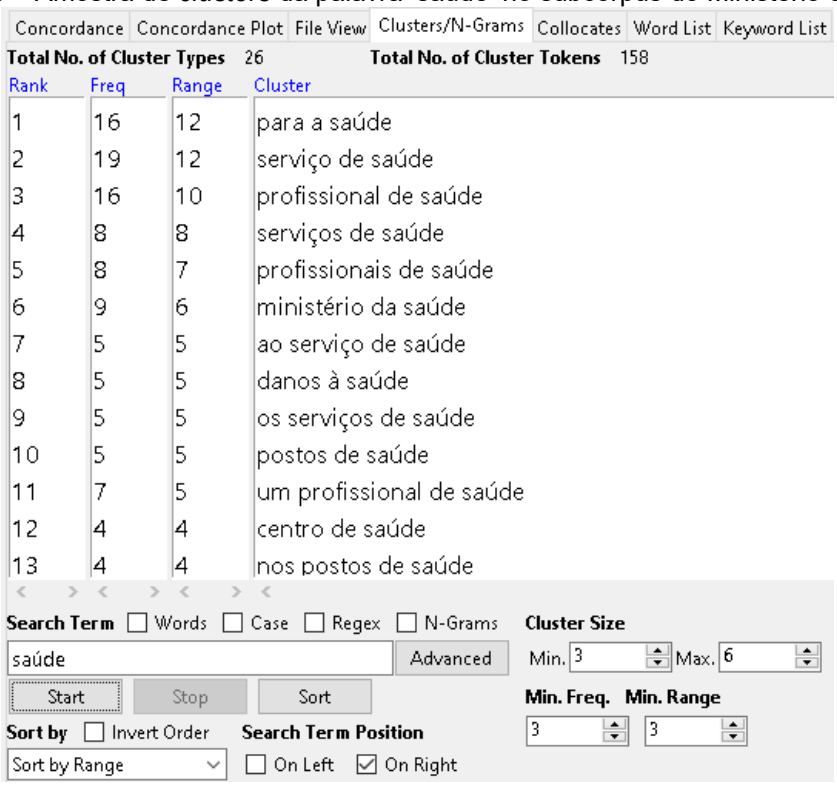
Há uma diferença de 33 textos entre o português (117 textos) e o inglês (84 textos). Como mencionado na seção 1.1.3, os textos em inglês de mesmo gênero costumam ser mais longos do que os em português (FUCHS, 2018), demandando, assim, que sejam compilados mais textos em português para obter número semelhante de palavras ao de inglês.

3.2.2.4 Clusters/N-grams

A ferramenta *Clusters/N-grams* permite que o usuário faça uma busca de sequências de n palavras no corpus. Pode-se escolher qualquer valor de n, fazendo buscas de bigramas ($n = 2$), trigramas ($n = 3$), quadrigramas ($n = 4$), e assim por diante. De acordo com Anthony (2019), essa ferramenta pode ajudar o usuário a encontrar expressões recorrentes em um corpus de estudo.

Além da pesquisa por sequências de palavras, por meio dessa ferramenta também é possível inserir uma palavra de busca. Com base nisso, os resultados mostrados apresentarão a palavra de busca colocada à esquerda ou à direita das sequências. Na Figura 13, a seguir, é possível observar exemplos de *clusters* extraído do subcorpus de textos do Ministério da Saúde, utilizando o termo de busca ‘saúde’.

Figura 13 – Amostra de *clusters* da palavra ‘saúde’ no subcorpus do Ministério da Saúde



Fonte: AntConc (ANTHONY, 2019).

Essa busca foi feita utilizando-se os parâmetros de uma janela de três (mínimo) a seis (máximo) palavras, com o termo de busca posicionado à direita. É possível, também, estabelecer uma frequência mínima, estabelecida em 3 no exemplo, e um número mínimo de textos (*range*) em que o *cluster* ocorre, que também foi estabelecido em 3.

Quando o *cluster* é formado por um número alto de palavras, a ferramenta mostra a sequência de palavras começando pelo menor número de palavras, até o número máximo. Por exemplo, o *cluster* ‘profissional de saúde’ está contido em ‘um profissional de saúde’, que, por sua vez, está contido em ‘por um profissional de saúde’.

3.2.3 Corpus paralelo: alinhamento e análise

A fim de fazer o contraste entre textos originais e textos traduzidos do corpus paralelo do MedlinePlus, foi utilizado o *software* AntPConc (ANTHONY, 2017). A partir

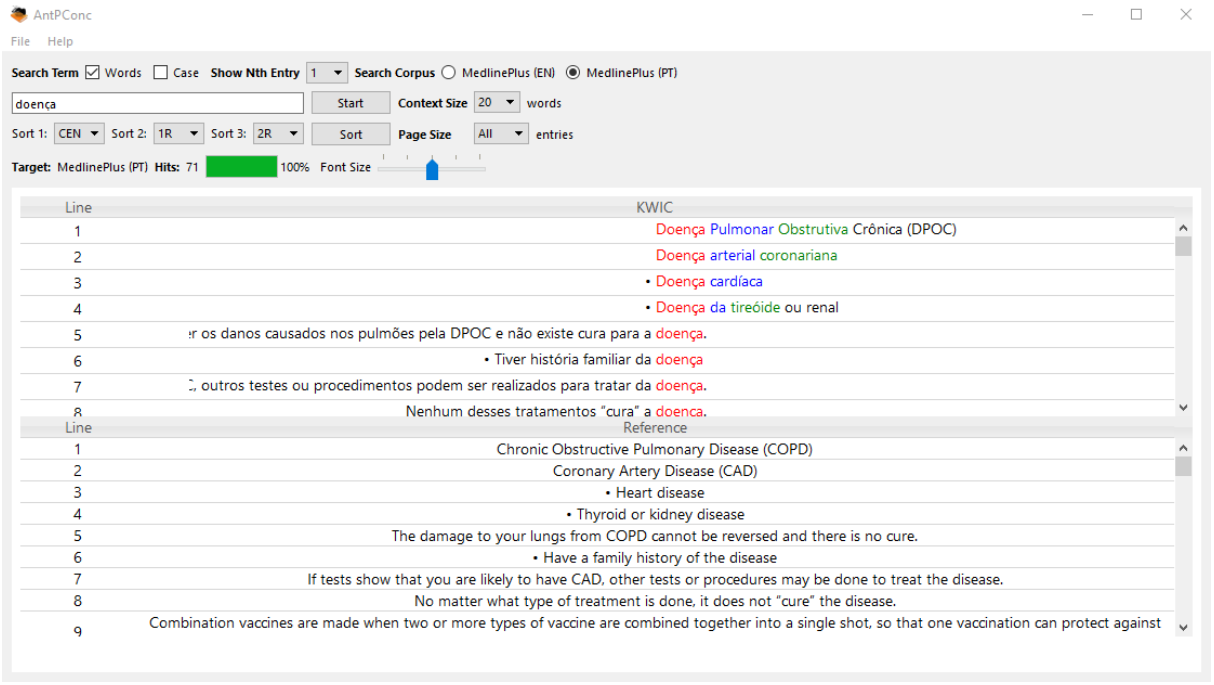
dele, é possível buscar linhas de concordância de uma palavra de busca do texto original alinhadas a seus trechos correspondentes no texto traduzido e vice-versa.

Para utilizar o *software*, antes de tudo, foi necessário fazer o alinhamento dos segmentos originais com os das traduções. Isso porque o tradutor pode optar por dividir uma frase em duas ou mais, bem como o contrário, unindo diferentes frases em uma só – ainda, o tradutor pode optar por não traduzir algum segmento do texto, ou por incluir informações ausentes no original. Para fazer o alinhamento do texto, foi utilizado o *software* gratuito LF Aligner (FARKAS, 2018), desenvolvido com o propósito de ajudar tradutores a criarem memórias de tradução com seus textos traduzidos, alinhando-os aos originais, para utilizá-los em *computer-assisted translation tools* (*CAT Tools*). O LF Aligner utiliza algoritmo para selecionar o segmento do original correspondente ao traduzido, com base em seu comprimento e no uso de dicionários (FARKAS, 2018).

O alinhamento precisa ser feito, necessariamente (até então), com base em um único texto-fonte com um único texto-alvo. Por isso, os 66 textos de cada língua foram colocados em arquivo único. Para que o *software* faça o alinhamento dos dois arquivos de texto, é necessário apenas escolher: I) o formato de arquivo que será anexado – nesse caso, txt, como já mencionado; II) as línguas dos textos desses dois arquivos – inglês e português; e III) selecionar os arquivos dos textos. É possível escolher entre a segmentação por frases ou por parágrafos e, ao final do processo, revisar e corrigir problemas de alinhamento.

Após o alinhamento, foi possível carregar os arquivos dos textos adequadamente segmentados no AntPConc. Como o *software* não faz nenhum tipo de etiquetagem ou lematização, é possível buscar por uma palavra (em qualquer um dos subcorpora, basta selecionar qual) e a ferramenta trará o seguimento que foi alinhado àquele em que se encontra a palavra de busca. Na Figura 14, a seguir, trazemos um exemplo a partir da palavra de busca ‘doença’.

Figura 14 – Alinhamento a partir da palavra ‘doença’



Fonte: AntPConc (ANTHONY, 2017).

No subcorpus selecionado para buscar a palavra, a ferramenta mostra *Keyword in Context* em vermelho, esquematizando visualmente as palavras que a seguem ou precedem por meio de cores. Já na outra língua, ela mostra o segmento em que a palavra deve estar contida, mas sem selecionar nenhuma palavra como central, pois o *software* não possui as ferramentas necessárias para fazer essa identificação. Assim, cabe ao usuário encontrar no segmento a palavra correspondente àquela buscada no texto apresentado.

No próximo capítulo, serão apontados os resultados dos levantamentos quantitativo e qualitativo.

4 RESULTADOS

Neste capítulo serão apresentados os resultados das análises quantitativa e qualitativa. Os resultados servirão para refletir sobre a adequação dos textos para o público-alvo, ou seja, a população geral. Primeiro, serão mostrados os dados levantados utilizando-se o Coh-Metrix e o Coh-Metrix-Port, apontando as especificidades encontradas em cada subcorpus de estudo. Posteriormente, serão expostos dados referentes ao levantamento feito utilizando-se o AntConc.

4.1 ANÁLISE QUANTITATIVA UTILIZANDO COH-METRIX E COH-METRIX-PORT

Os resultados do cálculo do Índice Flesch podem ir de 0 a 100, variando entre 'muito difícil' e 'muito fácil'. A Tabela 5 apresenta a escala de valores, além de uma estimativa do nível de escolaridade que compreende cada um desses níveis de inteligibilidade. Essa distribuição dos índices com base na escolaridade da população estadunidense foi apontada junto à fórmula do Índice Flesch. A adaptação para o cenário de educação brasileiro foi proposta por Martins *et al.* (1996).

Tabela 5 – Interpretação do Índice Flesch

Valor do Índice	**Descrição de Inteligibilidade**	**Escolaridade Estimada (EUA)**	**Escolaridade Estimada (BR)**
0 a 29	Muito difícil	*College graduate*	Universitários*
30 a 49	Difícil	*13th to 16th grade*	EM ou universitários
50 a 59	Razoavelmente difícil	*10th to 12th grade*	EM
60 a 69	Padrão	*8 th to 9 th grade*	Até 8ª série do EF
70 a 79	Razoavelmente fácil	*7 th grade*	Até 8ª série do EF
80 a 89	Fácil	*6 th grade*	Até 8ª série do EF
90 a 100	Muito fácil	*5 th grade*	Até 4ª série do EF

* Apenas para áreas acadêmicas específicas.
EM Ensino Médio
EF Ensino Fundamental
Fonte: elaborada pela autora com base em Flesch (1949 apud DUBAY, 2004) e Martins *et al.* (1996).

Para o levantamento do Índice Flesch, o corpus paralelo foi utilizado na íntegra, sendo feitos os levantamentos dos 66 textos (numerados na primeira coluna da Tabela 6) em língua inglesa (segunda coluna) e dos 66 textos em língua portuguesa (terceira coluna). Já para o levantamento do Índice Flesch dos textos escritos originalmente em língua portuguesa, foi utilizada uma amostra de 66 textos dos 191 textos (quarta

coluna), a fim de equiparar ao número de textos do corpus do MedlinePlus. A Tabela 6, a seguir, apresenta, na primeira coluna, o número correspondente ao texto nos subcorpora, sendo que os textos do MedlinePlus de mesmo número correspondem a original e tradução (conforme detalhado no Apêndice A).

Tabela 6 – Resultados referentes ao levantamento do Índice Flesch

	MedlinePlus (EN)	**MedlinePlus (PT)**	**Ministério da Saúde**
Texto 1	76,873		25,888
Texto 2	77,378	53,837	53,542
Texto 3	81,346	50,519	45,811
Texto 4	71,832	58,939	53,319
Texto 5	78,245	58,847	46,761
Texto 6	73,896	49,836	32,701
Texto 7	78,601	54,766	56,114
Texto 8	57,434	54,892	14,794
Texto 9	66,370	49,450	38,345
Texto 10	74,610	48,958	34,654
Texto 11	87,678	72,550	50,020
Texto 12	85,045	57,377	33,201
Texto 13	79,228	54,387	29,899
Texto 14	66,487		19,314
Texto 15	52,050	69,485	49,229
Texto 16	53,491	69,353	58,226
Texto 17	61,372	61,832	39,052
Texto 18	67,885	51,729	23,829
Texto 19	81,927	58,472	38,302
Texto 20	58,455		25,647
Texto 21	54,586		31,113
Texto 22	67,703	55,250	25,156
Texto 23	81,509	55,045	31,495
Texto 24	76,540	68,098	27,124
Texto 25	81,079	62,481	46,055
Texto 26	78,708	52,573	31,495
Texto 27	70,518	44,786	32,500
Texto 28	70,580	65,792	40,790
Texto 29	72,861	53,336	53,733
Texto 30	79,356	54,996	18,274
Texto 31	75,851	47,181	38,021
Texto 32	82,767	61,935	52,926
Texto 33	58,966	57,499	59,482
Texto 34	68,461	49,363	50,894
Texto 35	71,948	47,099	53,502
Texto 36	85,185	61,562	53,712
Texto 37	68,510	67,122	61,585
Texto 38	56,334		61,137
Texto 39	52,413		55,800
Texto 40	78,021	57,256	23,852

	MedlinePlus (EN)	MedlinePlus (PT)	Ministério da Saúde
Texto 41	76,492	53,693	44,308
Texto 42	66,521		11,234
Texto 43	71,913	49,737	33,862
Texto 44	84,907	54,203	23,197
Texto 45	80,842	52,888	23,952
Texto 46	85,023	58,821	36,721
Texto 47	72,383	45,878	32,131
Texto 48	83,922	72,667	32,382
Texto 49	85,921	77,367	40,446
Texto 50	74,856	66,228	37,132
Texto 51	77,741	54,021	22,267
Texto 52	92,773	55,990	63,272
Texto 53	84,310	63,377	50,474
Texto 54	68,780	38,654	47,961
Texto 55	99,821	87,276	20,675
Texto 56	89,292	60,599	43,357
Texto 57	73,945	57,789	61,074
Texto 58	88,529	59,726	48,732
Texto 59	81,362	60,505	40,675
Texto 60	73,463	63,403	28,724
Texto 61	79,402	61,690	40,082
Texto 62	72,973	51,793	47,892
Texto 63	83,278	51,696	40,640
Texto 64	77,447	54,598	21,896
Texto 65	87,497	63,592	42,615
Texto 66	64,270	49,103	28,584

Fonte: elaborada pela autora com base em dados do Coh-Metrix (GRAESSER *et al*., 2017) e do Coh-Metrix-Port (NILC, 2020).

Como se pode depreender a partir do número de células vazias, a ferramenta desenvolvida para a língua portuguesa possui algumas limitações em comparação com a da língua inglesa. Todos os textos em inglês foram processados pelo Coh-Metrix. Já em português, o Coh-Metrix-Port conseguiu processar 59 dos textos do MedlinePlus (PT). Os textos 1, 14, 20, 21, 38, 39 e 42 não foram processados. Esses textos eram mais longos que os demais que compunham o subcorpus e ultrapassaram o limite de palavras da ferramenta. O limite do Coh-Metrix é de 15 mil caracteres, enquanto o do Coh-Metrix-Port é de mil palavras. Portanto, as duas ferramentas utilizam unidades diferentes para limitar o levantamento. Com base nos valores de Índice Flesch levantados, foram feitos os cálculos estatísticos de média, mediana, variância e desvio padrão, utilizando a ferramenta de equações do programa Excel.

A média e a mediana são medidas de tendência central. A média consiste na soma de todos os elementos divididos pelo número total de elementos. Já a mediana

é determinada ao se ordenarem os dados de forma crescente ou decrescente, determinando o valor central da série (MORATO, 2011).

A variância e o desvio padrão são medidas de dispersão, servindo para apontar a variabilidade dos dados em torno da média. A variância mostra o quão distantes os valores estão da média. Ela é calculada a partir da soma dos quadrados da diferença entre cada valor e a média, dividida pelo número total de elementos. O desvio padrão é a medida do grau de dispersão em relação à média. Para calculá-lo, basta extrair a raiz quadrada da variância (MORATO, 2011).

A Tabela 7, a seguir, apresenta os resultados referentes às medidas de tendência central e às medidas de dispersão. As medidas de dispersão do subcorpus do MedlinePlus (EN) foram calculadas utilizando-se a equação de população, pois todos os textos foram avaliados pelo Coh-Metrix. Já do corpus comparável em português, foram calculadas as medidas de dispersão de amostra, pois não foram avaliados todos os textos que compõem seus subcorpora.

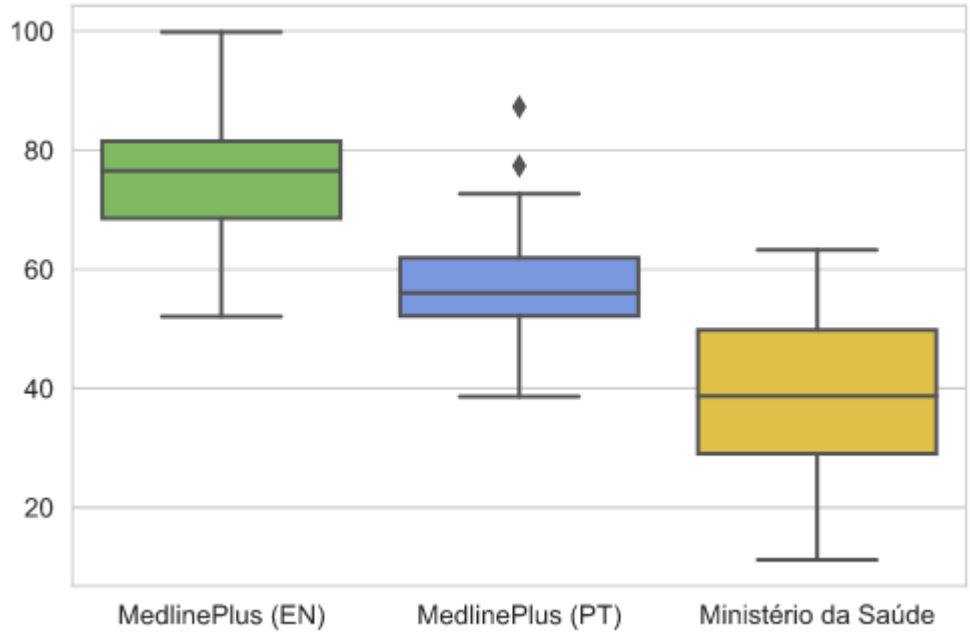
Tabela 7 – Resultado de cálculos estatísticos descritivos

	MedlinePlus (EN)	**MedlinePlus (PT)**	**Ministério da Saúde**
Média	74,845	57,659	39,115
Mediana	76,516	55,99	38,699
Variância	102,501	72,0444	170,942
Desvio padrão	10,124	8,488	13,074

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

Com base nos dados estatísticos levantados, foram construídos gráficos, a fim de facilitar a visualização e comparação entre os levantamentos dos diferentes subcorpora. O Gráfico 6, do tipo *boxplot*, apresenta visualmente os dados de mediana, que corresponde à linha horizontal no centro dos blocos, e de desvio padrão, que corresponde aos blocos coloridos. As linhas horizontais fora dos blocos representam o valor máximo e o valor mínimo. Por fim, os pequenos losangos representam os valores discrepantes.

Gráfico 6 – *Boxplot* dos dados do Índice Flesch



Fonte: elaborado pela autora.

A partir dos dados apresentados, é possível observar que as médias calculadas para o Índice Flesch apontam os textos do MedlinePlus (EN) com índices de inteligibilidade mais altos (média de 74,845), o subcorpus traduzido para o português, intermediários (57,659), e o subcorpus dos textos do Ministério da Saúde, originalmente escritos em português, com índices de inteligibilidade mais baixos (39,115). Além disso, os valores de desvio padrão mostram que os índices do subcorpus do Ministério da Saúde são os que contam com maior dispersão – de 13,074 –, seguidos pelos do MedlinePlus (EN) – de 10,124 –, por último, vem os do MedlinePlus (PT), com menor dispersão – de 8,488. O valor mínimo dos índices do MedlinePlus (EN) é de 52,05, o central (mediana) é de 76,516, e o máximo é de 99,821; no MedlinePlus (PT), o valor mínimo é de 38,654, o central é de 55,99, e o máximo, 72,367, com 77,367 e 87,276 como valores discrepantes; para o subcorpus do Ministério da Saúde, observamos valor mínimo de 11,234, valor central de 38,699, e máximo de 63,272.

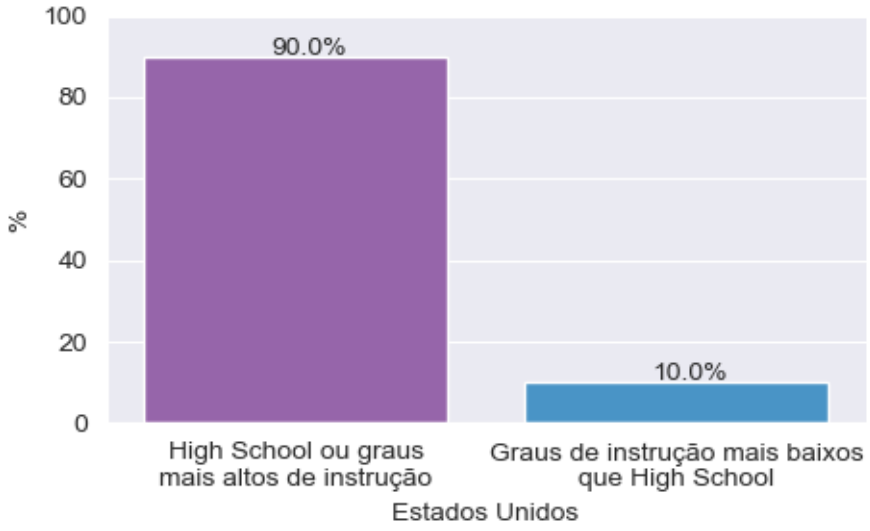
Esses dados contrariam, em parte, a hipótese que se tinha originalmente, baseada nos resultados da pesquisa de Pasqualini (2012). Apesar de aquela pesquisa ter sido realizada com um gênero diferente (contos literários), partimos do pressuposto que os resultados obtidos nos levantamentos do Coh-Metrix e Coh-Metrix-Port apresentassem médias similares dos índices. Os resultados de Pasqualini (2012)

apontam maior complexidade das traduções em português em relação a seus textos-fonte em inglês; também, maior complexidade dos textos traduzidos em comparação com os textos originalmente escritos em português.

Conforme esperado, o subcorpus composto por textos em língua inglesa representa os textos de mais fácil leitura, contando com uma média de Índice Flesch de 74,845. Porém, entre traduções e textos originalmente escritos em português, acreditava-se que o resultado se daria de forma contrária: que os textos originalmente escritos em português apresentariam Índice Flesch mais elevados que os textos traduzidos para o português – ou seja, com mais alto nível de inteligibilidade. Contrariando essa hipótese, os textos traduzidos para a língua portuguesa apresentaram a média do índice de 57,659. Já os textos escritos originalmente em língua portuguesa obtiveram a média mais baixa, de 39,115, apontando para um nível de complexidade mais elevado.

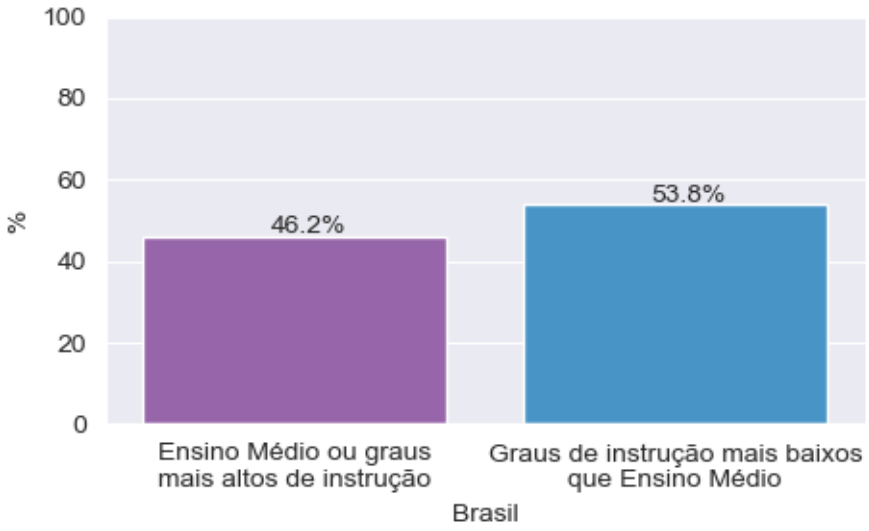
A seguir, os Gráficos 7 e 8 retomam as informações referentes ao nível de escolaridade no Brasil e nos Estados Unidos. Essa informação deve ser levada em consideração para determinar a adequação necessária dos textos pensando nos públicos-alvo (escolaridade da maior parcela da população) de cada país.

Gráfico 7 – Grau de instrução dos estadunidenses por parcelas da população



Fonte: elaborado pela autora com base em U.S. Census Bureau (2017).

Gráfico 8 – Grau de instrução dos brasileiros por parcelas da população



Fonte: elaborada pela autora com base em IBGE (2019).

Desde 2017, a grande maioria da população estadunidense possui no mínimo *High School* completo, sendo essa parcela da população 90% do total. Já no Brasil, comparando ao mesmo período, a parcela da população que tem grau equivalente de ensino não chega aos 50%. É uma diferença bastante significativa, principalmente se considerarmos que, ao passo que 46,2% da população brasileira possui Ensino Médio ou graus mais altos de instrução, há uma parcela de 33,7% que não completou ao menos o Ensino Fundamental. É importante que, ao produzir conteúdo escrito para o público geral, as informações sejam passadas mantendo essa população em mente, pois dentre os 53,8% da população que possui graus de instrução mais baixos que o Ensino Médio, a maioria se concentra no Ensino Fundamental incompleto. A Tabela 8 esquematiza as informações sobre escolaridade e a relação com os índices de inteligibilidade.

Tabela 8 – Índices conforme porcentagem mais significativa da população

País	**Índice**	**Descrição**	**Grau de instrução**	**% da população (2017)**
Estados Unidos	30 a 49	Difícil	*High School*	90%
Brasil	80 a 89	Fácil	EF Incompleto	33,7%

EF Ensino Fundamental

Fonte: elaborada pela autora com base em U.S. Census Bureau (2017) e IBGE (2019).

Considerando as informações esquematizadas na Tabela 8 e os dados de alfabetização referentes a cada um dos países apresentados no capítulo 2, se

fôssemos estabelecer um intervalo mais adequado de índices de inteligibilidade, os textos brasileiros deveriam se encaixar na inteligibilidade fácil, ou seja, nos índices entre 70 e 100. Já os textos estadunidenses, para serem acessíveis à maior parcela da população, poderiam apresentar inteligibilidade difícil, com índices entre 30 e 60. Portanto, com base nos levantamentos apresentados e nos dados de escolaridade, a população que necessita que as informações sejam fornecidas de maneira mais facilitada seria, justamente, a população que a receberia de maneira mais complexa.

Nesta seção, foram apresentados os dados de inteligibilidade levantados automaticamente com o Coh-Metrix e o Coh-Metrix-Port. Como explicado anteriormente, apesar de se tratar de uma análise estatística da superfície das sentenças e do texto como um todo, esses dados são importantes para apontar diferenças na distribuição das frases e parágrafos entre os textos originalmente escritos em português e os textos traduzidos para o português. Fundamentando-se nos resultados do levantamento do Índice Flesch, pôde-se notar que os textos em português, tanto traduzidos quando originais, foram enquadrados como complexos para o leitor médio brasileiro. Por outro lado, os textos em inglês obtiveram uma média de Índice Flesch que indica que os textos são fáceis para o leitor médio estadunidense.

Com base nesses dados, procederemos à análise textual utilizando o *software* AntConc. A análise textual, apesar de também partir de dados estatísticos, é feita de forma mais aprofundada. Pode-se olhar para palavras dentro dos contextos, considerando frequências de uso, chavicidade e colocação. A partir disso, serão feitas comparações entre o corpus comparável, retomando o corpus paralelo quando for necessário comparar traduções com seus originais.

4.2 ANÁLISE DOS DADOS TEXTUAIS LEVANTADOS NO ANTCONC

Para iniciar a análise dos dados textuais, o primeiro levantamento a ser feito nesta seção será de *type-token ratio* (TTR) dos corpora de estudo. O cálculo de *type-token ratio* indica a porcentagem da riqueza lexical do corpus, sendo demonstrado pelo cálculo $types \div tokens \times 100$. De acordo com Berber Sardinha (2004, p. 94), “Quanto maior o seu valor, mais palavras diferentes o texto conterá. Em contraposição, um valor baixo indicará um número alto de repetições, o que pode

indicar um texto menos rico, ou variado, do ponto de vista de seu vocabulário". A Tabela 9 sintetiza os números de textos, *types*, *tokens* e de *type-token ratio* dos corpora utilizados nesta pesquisa.

Tabela 9 – Números de *types* e *tokens* dos corpora de estudo

Subcorpus	**MedlinePlus (EN)**	**MedlinePlus (PT)**	**Ministério da Saúde (PT)**
Textos	66	66	191
Tokens	34.765	39.476	84.085
Types	3.088	4.554	9.666
TTR	8,88%	11,53%	11,49%

TTR *Type-Token Ratio*

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

O valor obtido no cálculo da riqueza lexical no subcorpus do MedlinePlus (EN) foi de 8,88%, o que significa que, de todo o conteúdo do subcorpus, 91,12% indica o universo das palavras que são repetidas. Na língua portuguesa, para os textos traduzidos, o percentual de riqueza lexical foi de 11,53%, ou seja, 88,47% do subcorpus é marcado por palavras que se repetem. O valor do *type-token ratio* do subcorpus do Ministério da Saúde é de 11,49%. Assim, de todo o conteúdo do subcorpus, 88,51% é formado por palavras que se repetem, e 11,49% representa o percentual de variedade vocabular, ou riqueza lexical. Por estarmos lidando com línguas diferentes, o índice calculado para o inglês não pode ser comparado com o do português.

O português é uma língua muito flexional quando comparada ao inglês. Cada verbo da língua portuguesa conta com, em geral, seis conjugações para cada modo-tempo, além de poder flexionar diversos substantivos e adjetivos em gênero, grau e número. Na língua inglesa, além de os substantivos em geral não indicarem gênero, os verbos têm menos flexões, fazendo uso de auxiliares para compor modo-tempo.

A partir desse levantamento, pode-se concluir que os índices de riqueza lexical dos textos de divulgação em português são muito próximos, com uma diferença de apenas 0,04%.

4.2.1 Levantamento de palavras-chave

Como mencionado no capítulo de metodologia, para o levantamento de palavras-chave dos três subcorpora de estudo, os índices estatísticos utilizados foram

o *log-likelihood* para a significância e o *odds ratio* para o tamanho de efeito. O ponto de corte para o *log-likelihood* foi de $p < 0{,}0001$, ou seja, a probabilidade de que um item seja selecionado como palavra-chave sem que seja estatisticamente relevante é de 0,01%. Além disso, foi estabelecido como ponto de corte para o *odds ratio* índice de no mínimo 10. Isso quer dizer que a diferença entre as frequências do item no corpus de estudo é dez vezes maior do que no corpus de referência.

A análise quantitativa do subcorpus do MedlinePlus (PT) resultou em 269 palavras-chave, a do subcorpus do Ministério da Saúde, 195, e a do MedlinePlus (EN), 280. A partir da lista de palavras-chave organizada, fazendo uso de uma ferramenta do próprio Excel[26], levantaram-se as palavras que recorrem nas duas listas de palavras-chave em português. Isso configura o que Berber Sardinha (2004) aponta como palavras-chave-chave. O levantamento de palavras-chave-chave foi feito para que estas fossem excluídas; restando, então, para a análise, somente palavras que fossem características de um dos subcorpora.

Por não se tratar de um corpus etiquetado morfossintaticamente e devido ao número de flexões que são feitas na língua portuguesa (em substantivos, adjetivos, artigos e verbos), optou-se por fazer a lematização manual das palavras do corpus comparável. As palavras foram agrupadas sob aquela com o valor de *log-likelihood* mais alto. Por exemplo, as palavras-chave do subcorpus do MedlinePlus (PT) 'precisa' (com *log-likelihood* de 95,08) e 'precisará' (45,34) foram agrupadas sob 'precisar' (97,83).

As palavras homógrafas foram diferenciadas, também, de forma manual. Por exemplo, a forma verbal 'sente' pode ser conjugação dos verbos 'sentir' e 'sentar'[27]. Por meio da análise dessas palavras em contexto, observou-se que de suas 13 ocorrências, 4 são do verbo 'sentar' e 9 são do verbo 'sentir'. Para separar as palavras, foi utilizada uma calculadora[28] de *effect size*, a fim de checar se, ao distinguir entre as diferentes acepções, essas continuariam sendo palavras-chave. O cálculo de *odds ratio* para a palavra 'sente' como conjugação de 'sentir' resultou no valor de 6,59, que não chegou ao ponto de corte de 10 estabelecido para o levantamento. Da mesma

[26] Em Formatação Condicional, selecionam-se Regras de Realce das Células, sendo aplicado o filtro 'Valores Duplicados'.
[27] 'Sentir' na 3ª pessoa do singular do presente do indicativo ou na 2ª pessoa do singular do imperativo afirmativo. 'Sentar' na 1ª pessoa do singular do presente do subjuntivo, na 3ª pessoa do singular do presente do subjuntivo ou na 3ª pessoa do singular do imperativo.
[28] Disponível em: http://ucrel.lancs.ac.uk/llwizard.html. Acesso em: 15 maio 2020.

forma, o índice de ‘sente’ (‘sentar’) é de 5,01, também não alcançando o ponto de corte. Por esse motivo, a palavra ‘sente’ foi excluída da lista de palavras-chave.

O mesmo processo foi repetido para o subcorpus do MedlinePlus (PT). A palavra ‘sente’ tem 16 ocorrências, sendo metade delas do verbo ‘sentir’ e a outra metade do verbo ‘sentar’. O valor do cálculo de *odds ratio* dos itens resultou em um índice de 7,51, que também não atingiu o ponto de corte, sendo a palavra excluída da lista de palavras-chave.

Após a lematização, a lista de palavras-chave exclusivas do MedlinePlus (PT) conta com 167 itens; na lista exclusiva do Ministério da Saúde há 115. Uma amostra de palavras-chave exclusivas do corpus comparável, organizada de forma decrescente por *log-likelihood*, e suas frequências está na Tabela 10. A lista de palavras-chave na íntegra consta nos Apêndices C, D e E.

Tabela 10 – Amostra de palavras-chave exclusivas do corpus comparável

MedlinePlus (PT)		**Ministério da Saúde**	
Frequência	**Palavras-chave**	**Frequência**	**Palavras-chave**
393	seu	153 [+25]	alimentos; alimento
251	médico	135	evitar
154	dor	48	camisinha
134	cirurgia	38	lixo
86 [+35]	tomar; tome	45	dentes
64	ligue	58	mulher
93	poderá	49	pé
100	fazer	35 [+16]	picada; picadas
116	pessoas	38	roupas
92	depois	43 [+21]	acidentes; acidente
52 [+20]	incisão; incisões	39	coluna
44	catapora	38	objetos
83	sinais	40	provocar
56	reação	25	manchas
35	vaginal	37	passo
55	use	24	hpv
35	alérgica	27	nariz
62	semanas	25	chão
48 [+13; +12; +36; +11; +12]	ajuda; ajudá[-lo/-la]; ajudam; ajudar; ajudará; ajude	36	veias
36	converse	22 [+18]	joelhos; joelho
36	enfermeira	28	articulações
50	exercícios	28	casas
44 [+14]	causar; causadas	32	fezes
27	hib	21	varizes
26 [+7]	respirar; respire	25	adolescente
26	termômetro	33	limpeza

MedlinePlus (PT)		Ministério da Saúde	
Frequência	**Palavras-chave**	**Frequência**	**Palavras-chave**
47	casa	35	sol
25	vaers	19	luvas
36	imediatamente	21	dengue

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

As escolhas lexicais feitas para determinado gênero de textos não ocorrem de forma aleatória. O falante faz uso de unidades semi-pré-construídas que, apesar de serem analisáveis em segmentos, constituem escolhas únicas. Isso quer dizer que os falantes dispõem de unidades linguísticas previamente armazenadas em blocos na memória, de modo que reiteram e repetem combinações anteriormente usadas para construir o discurso, causando a cristalização de certos padrões linguísticos em detrimento de outros (SINCLAIR, 1991; TAGNIN, 2013).

O objetivo deste trabalho é investigar se a tradução pode ser um mecanismo que acaba por dificultar o entendimento do leitor, principalmente por meio da quebra da convencionalidade. A fim de analisar a quebra de padrões linguísticos, o foco da análise serão as palavras-chaves características do subcorpus de textos traduzidos. Quando necessário, será feito um paralelo com ocorrências comparáveis no subcorpus de textos escritos originalmente em português, seja por meio de possíveis sinônimos[29], de comparação entre os contextos de uso de determinada palavra ou da apresentação dos colocados da palavra em corpus de língua geral.

Para fazer a análise manual das palavras-chave, os itens escolhidos foram aqueles característicos do gênero dos textos que compõem os corpora – textos de divulgação –, independentemente dos temas tratados. Ou seja, palavras-chave que denominam doenças, como ‘catapora’ e ‘glaucoma’, e sintomas de doenças, como ‘tontura’ e ‘erupção’, não serão analisadas. As palavras ‘bebê’, ‘vacina’ e ‘cirurgia’, por exemplo, são chave no MedlinePlus (PT) porque o corpus paralelo conta com um número maior de textos abordando esses assuntos, como é possível constatar por meio do Apêndice A, mas não recorrem em textos que tratam outros temas.

[29] O conceito de ‘sinonímia’ é, ainda hoje, muito discutido. Para os fins deste trabalho, o emprego se refere à relação de sentido entre duas unidades que pareçam estar sendo utilizadas de maneira intercambiável. Entende-se que não há sinonímia perfeita ou absoluta, pois, se existe a necessidade de empregar uma palavra diferente para um significado já existente, algum traço distintivo deve existir entre essas duas palavras ou entre seus respectivos contextos de uso (LIU, 2010).

A partir da lista de palavras-chave, algumas características do corpus comparável saltam aos olhos. Por exemplo, o subcorpus do MedlinePlus (PT) conta com uma lista extensa de verbos no imperativo. A lista dos 28 verbos no imperativo característicos desse subcorpus está presente na Tabela 11, a seguir.

Tabela 11 – Verbos no imperativo na lista de palavras-chave do subcorpus do MedlinePlus (PT)

tome; ligue; fume; use; ajude; converse; fique; respire; pare; lave; pergunte; fale; verifique; limpe; peça; coma; informe; consulte; descanse; comece; dirija; visite; tente; levante; seque; siga; consuma; troque.

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

Em contraste, o número de verbos no imperativo característicos do subcorpus do Ministério da Saúde é bem baixo, contando com apenas 2 verbos: ‘utilize’ e ‘retire’.

Como apresentado anteriormente a respeito do emprego de verbos no imperativo, no subcorpus do MedlinePlus (PT) ocorre emprego estatisticamente relevante da forma verbal ‘use’, enquanto no do Ministério da Saúde há o de ‘utilize’. Portanto, a análise começará pelas distinções ou semelhanças que se dão entre o emprego desses dois verbos. Posteriormente, será detalhado o emprego do pronome possessivo ‘seu’. Esse pronome chama a atenção por estar em primeiro lugar entre as palavras-chave exclusivas do subcorpus do MedlinePlus (PT). Após, serão apresentadas as semelhanças entre os empregos das palavras ‘sinais’ (no subcorpus do MedlinePlus) e ‘sintomas’ (no subcorpus do Ministério da Saúde). Por fim, serão apresentados os usos da forma verbal ‘poderá’, conjugação de ‘poder’ na 3ª pessoa do singular do futuro do presente do indicativo.

4.2.1.1 O caso das formas verbais ‘use’ e ‘utilize’

Para analisar os contextos de uso dos verbos ‘use’, frequentemente utilizado no subcorpus do MedlinePlus (PT), e de ‘utilize’, com alta frequência no subcorpus do Ministério da Saúde, foi necessário que se analisassem as linhas de concordância, pois o levantamento de colocados foi inconclusivo. Isso porque os subcorpora não são suficientemente extensos para revelarem alta recorrência de colocação, fazendo com que as palavras que atendem aos requisitos sejam majoritariamente gramaticais

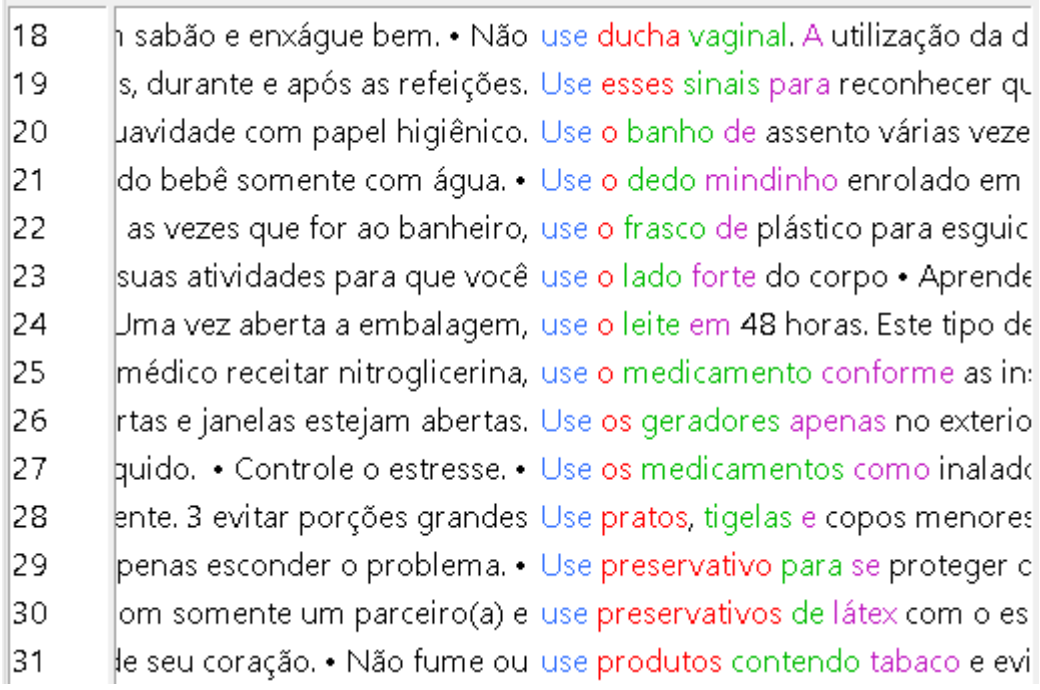
(como ‘um’, ‘para’, ‘de’ e ‘o’). A seguir, serão descritos os contextos de uso e as diferenças na aplicabilidade dos verbos conforme o subcorpus de estudo.

4.2.1.1.1 ‘Use’ e ‘utilize’ no subcorpus do MedlinePlus (PT)

A forma verbal ‘use’ tem 55 ocorrências no subcorpus do MedlinePlus (PT) – 13,93 ocorrências a cada 10.000 palavras. As linhas de concordância evidenciaram construções de ‘use’ com diversos substantivos. Por exemplo, observou-se a utilização de sequências como ‘use absorventes’ (4 ocorrências), ‘use tampões’ (4 ocorrências) e ‘use preservativo(s)’ (2 ocorrências). Para traçar um paralelo, no subcorpus do Ministério da Saúde, a palavra ‘tampões’ tem ocorrência única, estando relacionada à aplicação nos olhos: ‘assim como o uso de tampões para manter o olho fechado’. A respeito do colocado ‘preservativo(s)’, a palavra ocorre 6 vezes no subcorpus do MedlinePlus (PT). Enquanto isso, no subcorpus do Ministério da Saúde, apesar de haver 10 ocorrências de ‘preservativo(s)’, utiliza-se muito mais a palavra ‘camisinha’ (que aparece também como colocado de ‘use’), com 48 ocorrências.

Os contextos de ‘use’ parecem estar associados a orientações médicas do que fazer em determinadas situações. Algumas linhas de concordância que exemplificam o tipo de orientações dadas podem ser observadas na Figura 15.

Figura 15 – Linhas de concordância de ‘use’ no subcorpus do MedlinePlus (PT)



Fonte: AntConc (ANTHONY, 2019).

Já a forma verbal ‘utilize’ ocorre apenas uma vez no subcorpus do MedlinePlus (PT), na frase ‘Nunca utilize este gerador dentro de casa ou na garagem’. Vale ressaltar que, dentre as linhas apresentadas na figura, há uma instância da forma verbal ‘use’ aparecendo também associada a geradores.

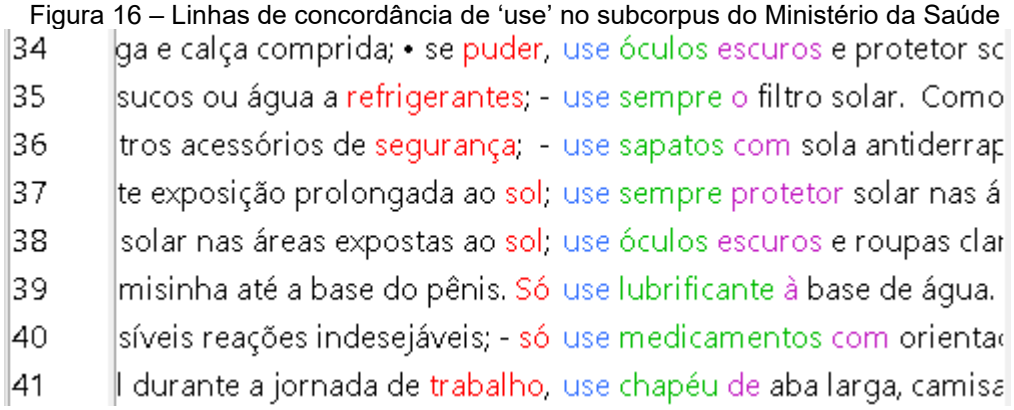
Para investigar a origem de ‘use’ no português, convém analisar o corpus paralelo de originais e traduções. Diferentemente do português, as palavras não contam com muitas flexões no inglês. Há algumas flexões que ocorrem no verbo, como, por exemplo, a forma ‘*use*’ é utilizada na terceira pessoa do plural e na primeira e terceira pessoas do singular (*I* e *you*), enquanto a forma ‘*uses*’ é empregada na terceira pessoa do singular; há também ‘*using*’, que é a forma no gerúndio; e ‘*used*’, que é a flexão do verbo no passado. Além disso, há, também, as formas ‘*use*’ e ‘*uses*’ como substantivo.

Dessa forma, apesar de ocorrer 90 vezes em inglês, ‘*use*’ pode ser utilizado de forma equivalente ao verbo no infinitivo (‘usar’) no português e ao substantivo (‘uso’). Mais da metade das ocorrências de ‘*use*’ do inglês foram traduzidas para ‘use’ em português (55 ocorrências). A partir disso, podemos supor que a utilização desse verbo esteja frequentemente relacionada às ocorrências do verbo cognato ‘*use*’ em língua inglesa.

4.2.1.1.2 ‘Use’ e ‘utilize’ no subcorpus do Ministério da Saúde

Diferentemente do subcorpus traduzido, há recorrência tanto de ‘use’ quanto de ‘utilize’ nos textos originalmente escritos em português. A palavra ‘use’ ocorre 41 vezes (4,88 a cada 10.000), e ‘utilize’, 17 (2,02 a cada 10.000). Serão analisados os contextos de uso para verificar se essas palavras estabelecem relação de sinonímia ou não. A fim de determinar como ocorre a diferenciação entre esses dois verbos, apresentaremos, na Figura 16, suas linhas de concordância no subcorpus do Ministério da Saúde.

Figura 16 – Linhas de concordância de ‘use’ no subcorpus do Ministério da Saúde

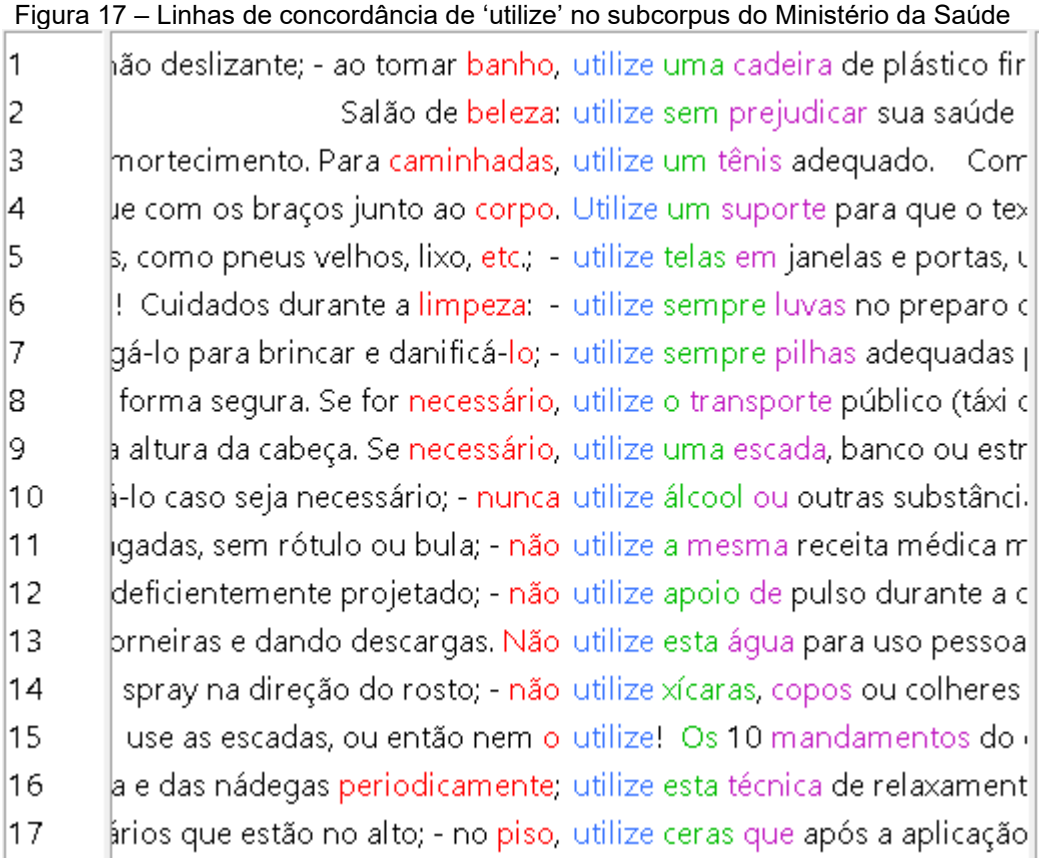


Fonte: AntConc (ANTHONY, 2019).

As ocorrências de ‘use’ no subcorpus do Ministério da Saúde estão bastante ligadas a objetos que são portados (‘chapéu’, ‘cintos’, ‘roupas’, ‘sapatos‘ e ‘óculos escuros’), medicamentos e, como já apontado, a forma verbal ‘use’ aparece próxima à palavra ‘camisinha’. Ou seja, nesse subcorpus, as utilizações da forma verbal ‘use’ parecem estar restritas a vestimentas e medicamentos.

Já a palavra ‘utilize’, no subcorpus do Ministério da Saúde, ocorre em contextos semelhantes aos de ‘use’ no subcorpus do MedlinePlus (PT). Aparecem formulações como ‘não utilize’ ou ‘nunca utilize’, ‘utilize sempre’ seguidos de alguns substantivos, porém sem recorrências. Assim como as ocorrências de ‘use’ no subcorpus do MedlinePlus (PT), as ocorrências de ‘utilize’ estão relacionadas a contextos em que se orienta o leitor ao que ele deve fazer ao se deparar com determinada situação. Os exemplos podem ser observados na Figura 17.

Figura 17 – Linhas de concordância de ‘utilize’ no subcorpus do Ministério da Saúde



Fonte: AntConc (ANTHONY, 2019).

Nas linhas 3 e 6, é possível notar que ‘utilize’ está associado a itens de vestuário, como ‘tênis’ e ‘luvas’. Exceto por essas linhas, outras ocorrências estão seguidas por objetos que servirão para ajudar a cumprir determinada ação. Por exemplo, na linha 1, orienta-se utilizar uma cadeira de plástico como auxílio para tomar banho. Esse tipo de contexto parece se repetir em todas as outras linhas.

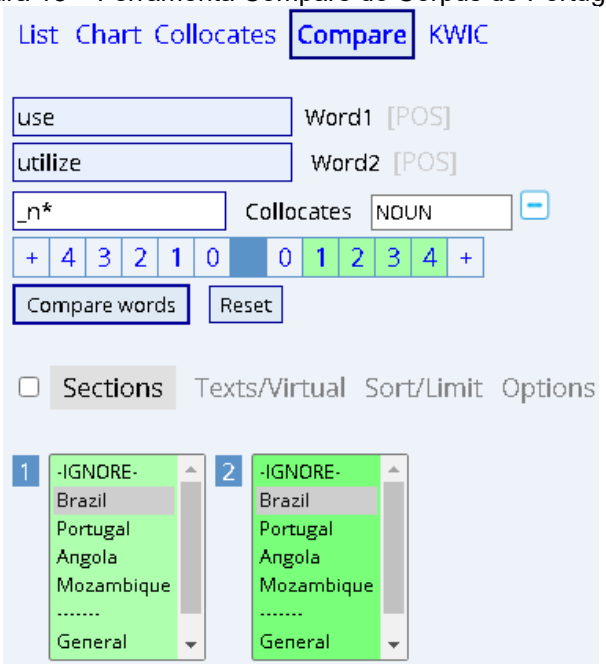
4.2.1.1.3 ‘Use’ e ‘utilize’ em corpus de língua geral

Para ir além dos colocados e das linhas de concordância do corpus comparável e investigar os contextos em que ‘use’ e ‘utilize’ são aplicados na língua geral, foi utilizado como base o Corpus do Português (DAVIES, 2015). Esse corpus, que está disponível gratuitamente on-line, permite que seja feita a comparação entre os colocados de duas palavras por meio da ferramenta *Compare*.

Apesar de se tratar de um corpus geral de língua portuguesa, que abrange textos de diferentes países falantes da língua, o subcorpus Web/Dialects permite que a busca seja feita apenas em textos em português brasileiro, conforme a Figura 18.

Para traçar um paralelo com os levantamentos do AntConc, optou-se por levantar somente os substantivos que aparecem à direita dos verbos. A janela escolhida foi, também, de até 4 palavras.

Figura 18 – Ferramenta Compare do Corpus do Português



Fonte: Corpus do Português (DAVIES, 2015).

A Tabela 12 apresenta os 25 primeiros substantivos que ocorrem à direita dos verbos ‘use’ e ‘utilize’, organizados pelo valor do índice (*score*) de relação entre o colocado e a palavra de busca. Nesta tabela, a frequência do colocado associado à palavra ‘use’ está na coluna P1; e a frequência de associação entre ‘utilize’ e o colocado está na coluna P2.

Tabela 12 – Colocados de ‘use’ e ‘utilize’ no Corpus do Português

Colocado de ‘use’	**P1**	**P2**	**Score**	**Colocado de ‘utilize’**	**P2**	**P1**	**Score**
curador	357	0	216,5	selo	101	1	333,1
bálsamo	211	0	128	créditos	136	2	224,3
camisinha	206	0	124,9	fins	575	19	99,8
balsamo	152	0	92,2	login	106	4	87,4
comentários	88	0	53,4	polegares	26	1	85,8
autoridade	67	0	40,6	palhas	11	0	72,6
box	119	1	36,1	pagamento	20	1	66
mente	44	0	26,7	responsabilidade	114	6	62,7
blusas	40	0	24,3	fórum	26	2	42,9
maquiagem	66	1	20	x	141	11	42,3
vestido	33	0	20	email	123	29	14

Colocado de ‘use’	P1	P2	Score	Colocado de ‘utilize’	P2	P1	Score
mouse	237	4	18	sistemas	12	3	13,2
perfume	29	0	17,6	formulário	139	45	10,2
criatividade	342	6	17,3	formas	33	13	8,4
kit	114	2	17,3	rolagem	24	10	7,9
instrumentos	104	2	15,8	senha	118	51	7,6
chicote	25	0	15,2	aparelhos	29	13	7,4
calça	48	1	14,6	botão	321	153	6,9
jeans	24	0	14,6	dispositivo	10	5	6,6
saias	24	0	14,6	seção	13	7	6,1
branco	23	0	13,9	forma	149	81	6,1
talento	22	0	13,3	letras	54	30	5,9
vestidos	22	0	13,3	código	117	67	5,8
chapéu	21	0	12,7	tag	12	7	5,7
salto	21	0	12,7	aço	13	8	5,4

Fonte: elaborada pela autora com base em dados do Corpus do Português (DAVIES, 2015).

O que se pode constatar a partir dos colocados de ‘use’ é que suas ocorrências estão mais associadas a objetos. Por exemplo, observa-se o uso de ‘use camisinha’, consoante à recorrência no subcorpus do Ministério da Saúde. Além disso, ocorrem, também, as palavras ‘blusas’, ‘maquiagem’, ‘vestido’, ‘instrumentos’, ‘calça’, ‘jeans’, ‘saias’, ‘vestidos’, ‘chapéu’ e ‘salto’. As palavras citadas, com exceção de ‘instrumentos’, se referem a vestuário. Ou seja, além de a forma verbal ‘use’ aparecer associada a objetos, frequentemente estes objetos são itens de vestuário.

Por outro lado, a forma verbal ‘utilize’ é empregado com referência a palavras como ‘créditos’, ‘pagamento’, ‘responsabilidade’, ‘sistemas’, ‘aparelhos’, ‘dispositivo’, ‘forma’, ‘letras’, dentre outras.

4.2.1.2 O caso do pronome possessivo ‘seu’

Como é possível constatar por meio da amostra de palavras-chave (Tabela 10), ‘seu’ é a primeira palavra estatisticamente relevante característica do subcorpus do MedlinePlus (PT). Para entender o seu papel nesse subcorpus, partiu-se para a análise dos colocados da palavra.

O levantamento de colocados de ‘seu’ foi feito na ferramenta *Collocate* do AntConc, utilizando uma janela de 4 palavras, sendo feito primeiro o levantamento para a direita e posteriormente para a esquerda. A frequência mínima estabelecida para uma palavra ser considerada colocado foi de 6 ocorrências.

A partir da lista de colocados, foi possível depreender que as palavras com maior frequência relevante de proximidade com ‘seu’ foram alguns substantivos (à direita) e alguns verbos (à esquerda). ‘Filho’, ‘parceiro’ e ‘médico’ foram os três principais substantivos. Vale ressaltar que, das 393 ocorrências (99,55 a cada 10.000) de ‘seu’ no subcorpus do MedlinePlus (PT), 168 (42,56 a cada 10.000) são na sequência de palavras ‘seu médico’. ‘Converse’, ‘pergunte’ e ‘ligue’ são as formas verbais que mais recorrem – que estão no imperativo, vale ressaltar – antecedendo a palavra ‘médico’.

No subcorpus dos textos originais em inglês, a palavra ‘*your*’ [seu, sua, seus, suas] é a segunda palavra mais frequente, com 1.141 ocorrências (328,2 a cada 10.000) – sendo que, em primeiro lugar, está a palavra ‘*the*’ [o, a, os, as], com 1.241 ocorrências (356,97 a cada 10.000). Apesar da recorrência do pronome possessivo, muitos ‘*your*’ do MedlinePlus (EN) não foram traduzidos como ‘seu’. Com isso em mente, ao considerar-se que as palavras do português variam em número e gênero, foi necessário atentar para as flexões do pronome possessivo. Por meio do levantamento, observou-se que ‘seus’ aparece 28 vezes, ‘sua’ aparece 118 vezes, e ‘suas’ aparece 45 vezes. Somando esses números, há um total de 584 ocorrências (147,94 a cada 10.000) do lema ‘seu’ no MedlinePlus (PT).

Para fazer os levantamentos com todas as variantes, foi possível pesquisar os colocados utilizando ‘seu*’ como busca para o masculino e o plural. Entretanto, para pesquisar os colocados de ‘sua’ e ‘suas’ foi necessário que fossem feitas duas buscas, pois a partir da busca ‘sua*’, foram levantadas palavras como ‘suavemente’, ‘suavidade’, ‘suar’, dentre outras. A configuração para levantar os colocados foi a mesma mencionada anteriormente.

Alguns substantivos foram levantados como colocados de ‘seu’, como ‘parceiro’, ‘filho’, ‘filha’, ‘médico’, ‘coração’, ‘profissional’, ‘quarto’, ‘enfermeira’, ‘bebê’, ‘corpo’ e ‘caso’. Para ‘sua(s)’, os principais substantivos colocados que apareceram foram ‘filha’, ‘família’, ‘incisão’, ‘casa’ e ‘pernas’.

Para fins de comparação, foi feito o mesmo levantamento de colocados no subcorpus de originais português. Os substantivos colocados em comum de ‘seu(s)’ e ‘sua(s)’ que apareceram na lista do Ministério da Saúde foram ‘filho’, ‘médico’, ‘corpo’ e ‘casa’.

A frequência de associação entre as palavras ‘seu’ e ‘médico’ no subcorpus do MedlinePlus (PT) chamou a atenção, principalmente considerando que o subcorpus dos textos traduzidos é quase duas vezes menor que o dos produzidos em português. A palavra ‘médico’ é colocado de ‘seu’ 176 vezes (44,58 a cada 10.000) no subcorpus do MedlinePlus (PT), já no do Ministério da Saúde, apenas 25 vezes (2,97 a cada 10.000). Por esse motivo, foi feita a pesquisa partindo do movimento contrário, a fim de descobrir se há outras palavras que aparecem imediatamente à esquerda de ‘médico’ e de ‘*doctor*’ nesses corpora. O resultado do levantamento está esquematizado na Tabela 13.

Tabela 13 – Colocados imediatamente à esquerda de ‘médico’ e ‘*doctor*’ nos corpora de estudo

MedlinePlus (EN)			**MedlinePlus (PT)**			**Ministério da Saúde**		
Freq.	**Colocado**	**Item de busca**	**Freq.**	**Colocado**	**Item de busca**	**Freq.**	**Colocado**	**Item de busca**
204	*your*	*doctor*	168	seu	médico	27	o	médico
21	*the*		53	o		25	um	
8	*baby's*		8	ao		23	seu	
			6	um		17	pelo	
						7	ao	
						6	atendimento	
						6	do	

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

A partir desse levantamento, foi possível observar que, apesar de ‘médico’ ser colocado de ‘seu’ no subcorpus do Ministério da Saúde, na verdade, ‘o médico’ e ‘um médico’ são mais utilizadas do que ‘seu médico’. Já no subcorpus de textos traduzidos, a sequência mais utilizada realmente é ‘seu médico’, enquanto ‘o médico’ ocorre apenas 53 vezes, quase $1/3$ das ocorrências daquela. Essas duas sequências mais utilizadas vão ao encontro do que ocorre no subcorpus dos textos originais, sendo ‘*your doctor*’ a sequência mais utilizada, com 204 ocorrências, e ‘*the doctor*’ a segunda mais utilizada, com 21 ocorrências. Também, no MedlinePlus (PT), ‘o médico’ ocorre mais que o dobro de vezes de ‘*the doctor*’.

Ainda, vale ressaltar que o corpus paralelo foi analisado no *software* AntPConc. Ao olhar para as linhas de concordância de forma alinhada, foi possível encontrar alguns usos que chamaram a atenção. Por exemplo, em inglês, havia ocorrências de ‘*your child*’, que, no português, foram traduzidas como ‘seu adolescente’.

4.2.1.3 O caso dos substantivos ‘sinais’ e ‘sintomas’

Observou-se que a palavra ‘sinais’ tem alta chavicidade na lista exclusiva do MedlinePlus (PT). Por outro lado, na lista de palavras-chave do subcorpus do Ministério da Saúde não foi encontrada ‘sinais’ ou alguma palavra que denote o mesmo sentido. A partir das buscas no corpus paralelo do MedlinePlus (EN-PT), então, foi possível encontrar ocorrências da palavra ‘sintomas’, como será detalhado a seguir.

Assim, como mencionado anteriormente, por vezes foi necessário analisar palavras que poderiam ser consideradas sinônimas (como ‘use’ e ‘utilize’). Nesta seção, serão analisadas as palavras ‘sinais’, chave no subcorpus do MedlinePlus (PT), e seu possível sinônimo ‘sintomas’. A palavra ‘sinais’ aparece 83 vezes (21,03 a cada 10.000) nesse subcorpus, enquanto a palavra ‘sintomas’ aparece 52 (13,17 a cada 10.000).

No subcorpus do Ministério da Saúde, ‘sintomas’ tem 262 ocorrências (31,16 a cada 10.000); a palavra ‘sinais’ tem 42 ocorrências (4,99 a cada 10.000). Nenhuma delas está presente na lista de palavras-chave, fato que será detalhado no decorrer da seção.

4.2.1.3.1 ‘Sinais’ e ‘sintomas’ no corpus paralelo do MedlinePlus (EN-PT)

Observou-se que a palavra ‘sinais’ tem alta chavicidade no corpus MedlinePlus (PT), ao passo que não figura entre as palavras-chave do subcorpus do Ministério da Saúde. Portanto, a partir do alinhamento dos textos em inglês e suas traduções, buscou-se identificar a(s) palavra(s) que tivesse(m) originado o equivalente ‘sinais’. Então, ‘sinais’ foi o equivalente tradutório de ‘*signs*’ em 83 (21,03 a cada 10.000) de um total de 117 ocorrências da palavra nos textos em inglês. A partir disso, realizou-se nova procura, dessa vez partindo da palavra ‘*signs*’, a fim de verificar outro(s) equivalente(s) utilizado(s), além de ‘sinais’, para recuperar o termo nas traduções. Uma amostra do alinhamento das linhas de concordância pode ser observada na Figura 19.

Figura 19 – Linhas de concordância de '*signs*' no corpus paralelo do MedlinePlus (PT-EN)

Line	KWIC
17	Signs of Chronic Kidney Failure include:
18	Signs of Coronary Artery Disease
19	Signs of Labor
20	Signs of Retinal Tears and Detachment
21	Signs of STDs
22	Signs of Substance Abuse or Dependency
23	Signs of a severe allergic reaction can include hive

Line	Reference
17	Os sintomas da falência crônica dos rins incluem:
18	Sinais de doença arterial coronariana
19	Sinais de trabalho de parto
20	Sinais de ruptura ou descolamento de retina
21	Sinais de DSTs
22	Sintomas de abuso ou dependência de substâncias
23	Os sinais de reação alérgica grave podem incluir urticária, inchaço da face e garganta, dificuldade para respirar, batimentos cardíacos acelerados, tontura e fraqueza.

Fonte: AntPConc (ANTHONY, 2017).

Por meio desse levantamento, verificou-se que, das 117 ocorrências (29,64 a cada 10.000) de '*signs*', 31 foram traduzidas por 'sintomas'. Ou seja, por meio desse levantamento, poder-se-ia concluir que 'sinais' e 'sintomas' seriam sinônimos, escolhidos a critério do tradutor.

Observou-se, posteriormente, que a palavra 'sintomas' ocorre nos textos traduzidos 52 vezes no total (13,17 a cada 10.000) – sendo 3 vezes na forma singular 'sintoma'. Por observar que havia ainda 21 ocorrências da palavra com origens desconhecidas, optou-se por buscá-la no corpus paralelo. Dessas 21 ocorrências, uma parte tem origem em frases que não se utiliza a palavra '*signs*', ou seja, Ø (17 ocorrências), conforme a Figura 20.

Figura 20 – Linhas de concordância de ‘sintomas’ no corpus paralelo do MedlinePlus (PT-EN)

Line	KWIC
37	▪ Se achar que os sintomas indicam uma reação alérgica grave ou o
38	os cardíacos acelerados, tontura e fraqueza. Esses sintomas normalmente começariam alguns minut
39	os cardíacos acelerados, tontura e fraqueza. Esses sintomas normalmente começariam alguns minut
40	▪ sintomas parecidos com os da a gripe, mas que n.
41	usadas por um vírus não afetado pela vacina ou ▪ sintomas parecidos com a gripe, mas que não são
42	corro ou entre em contato com seu médico se os sintomas piorarem ou se você tiver febre de mais
43	Contate seu médico caso estes sintomas piorem ou não desapareçam em poucas
44	Sarampo O vírus do sarampo causa sintomas que podem incluir febre, tosse, nariz esc
45	Estes sintomas são normais e devem passar no dia segu
Line	**Reference**
37	▪ If you think it is a severe allergic reaction or other emergency that can’t wait, call 9-1-1 or get the person to the nearest hospital.
38	
39	
40	
41	
42	▪ Return to the Emergency Department or call your doctor if your signs get worse or you have a fever of more than 100.5 degrees F or 38 degrees C.
43	Call your doctor if this gets worse or does not go away in a few weeks.
44	MEASLES (M) can cause fever, cough, runny nose, and red, watery eyes, commonly followed by a rash that covers the whole body.

Fonte: AntPConc (ANTHONY, 2017).

Por fim, descobriu-se que há outra parte que foi traduzida da palavra ‘*symptoms*’ (4 ocorrências). Na Figura 21, podem ser observadas as linhas de concordância que ilustram os contextos de uso da palavra ‘sintomas’ e ‘*symptoms*’.

Figura 21 – Linhas de concordância de ‘sintomas’ como tradução de ‘*symptoms*’ no corpus paralelo do MedlinePlus (EN-PT)

Line	KWIC
10	ão utilizados antes de a pessoa apresentar algum sintoma.
20	semanas seguintes você poderá ter que lidar com sintomas de falta do cigarro e compulsões.
26	icos e o relaxamento podem ajudar a melhorar os sintomas de raiva, nervosismo e irritabilidade.
27	as as mulheres devem ser informadas dos riscos e sintomas do câncer do endométrio.
Line	**Reference**
10	Screening tests are used to find cancer before a person has any symptoms.
20	Over the next days and weeks you may be coping with withdrawal symptoms and cravings.
26	Exercise and relaxation can help with withdrawal symptoms of anger, edginess or irritability.
27	The American Cancer Society recommends that at the time of menopause, all women should be told about the risks and symptoms of endometrial cancer.

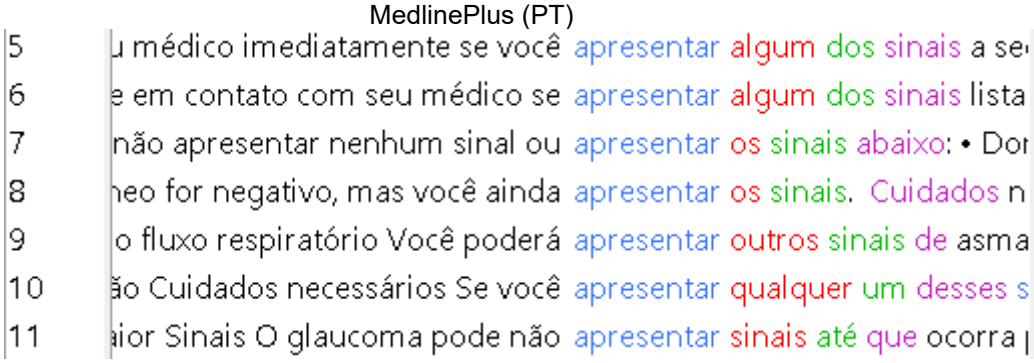
Fonte: AntPConc (ANTHONY, 2017), com marcações da autora.

A fim de tentar mapear se ocorre alguma distinção entre os contextos de ‘sinais’ e de ‘sintomas’ entre traduções e originais em português, foi feita a pesquisa por colocados. O levantamento de colocados de ‘sinais’ foi feito respeitando-se a janela

de até 4 itens à direita e à esquerda, estabelecendo a frequência mínima de 6 ocorrências. A pesquisa por colocados de ‘sintomas’ no subcorpus de originais em português será relatada na seção a seguir.

No subcorpus do MedlinePlus (PT), como principais colocados de ‘sinais’ à esquerda, ocorrem as formas verbais ‘apresentar’ e ‘pode’, o adjetivo ‘preocupante’ e algumas palavras gramaticais, como ‘algum’, ‘os’, ‘você’, ‘um’, dentre outras. A Figura 22 mostra linhas de concordância de ‘apresentar’ como colocado de ‘sinais’.

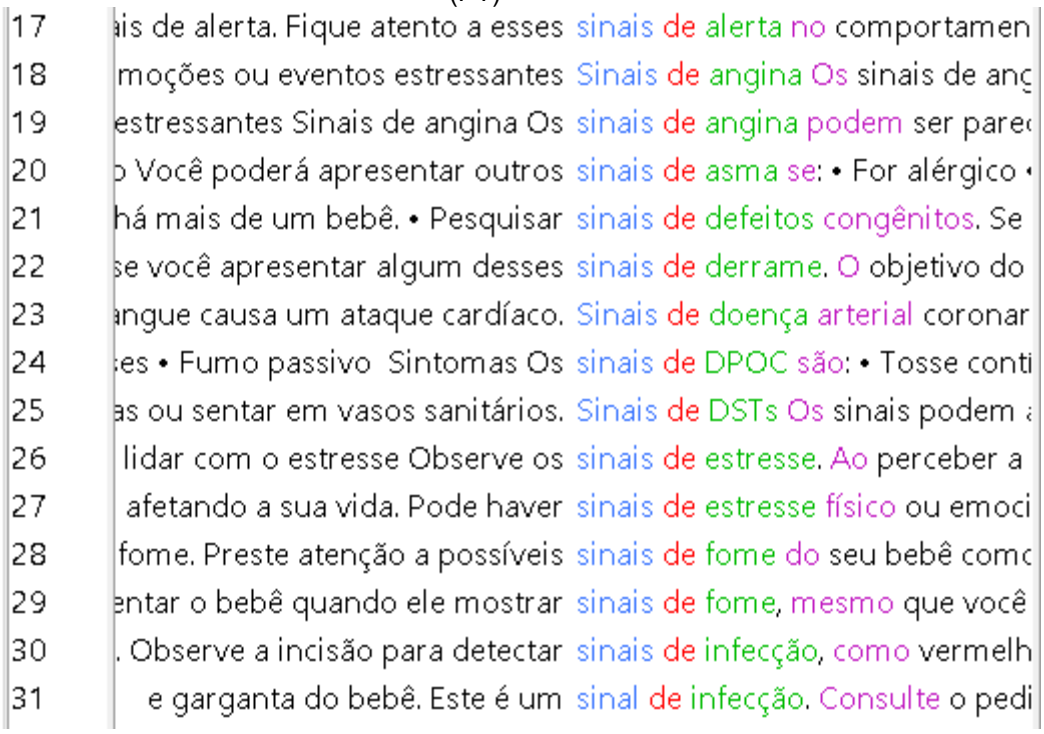
Figura 22 – Linhas de concordância de ‘apresentar’ como colocado de ‘sinais’ no subcorpus do MedlinePlus (PT)



Fonte: AntConc (ANTHONY, 2019).

À direita de ‘sinais’, o colocado mais frequente é a preposição ‘de’, que conta com 62 ocorrências nas proximidades de ‘sinais’ (Figura 23). Seu principal uso é em construções que seguem o padrão de ‘sinais de $[substantivo]$’, com 41 ocorrências.

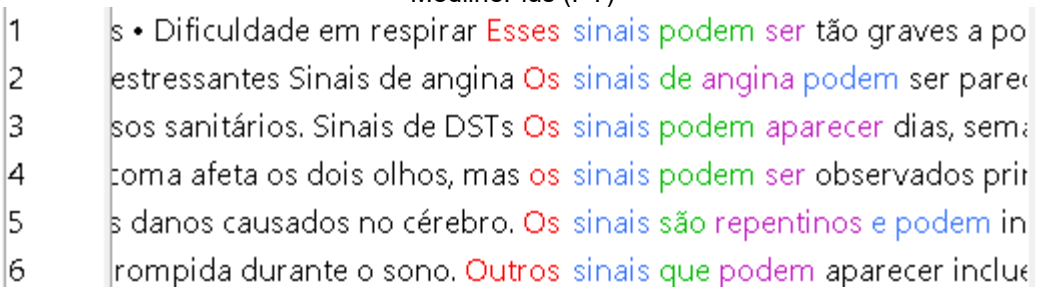
Figura 23 – Linhas de concordância de ‘de’ como colocado de ‘sinais’ no subcorpus do MedlinePlus (PT)



Fonte: AntConc (ANTHONY, 2019).

Além disso, outro colocado à direita com relevância estatística é a forma verbal ‘podem’. Ela ocorre 6 vezes na janela de 4 palavras após a palavra ‘sinais’. Alguns exemplos estão apresentados na Figura 24.

Figura 24 – Linhas de concordância de ‘podem’ como colocado de ‘sinais’ no subcorpus do MedlinePlus (PT)



Fonte: AntConc (ANTHONY, 2019).

A partir das linhas de concordância mostradas pelo AntPConc e AntConc, é possível verificar que os usos de ‘sinais’ e de ‘sintomas’ parecem estar indiscriminados. Ou seja, não é possível distinguir os contextos em que os substantivos aparecem. Assim como ‘sinais’ aparece perto de doenças (como

‘catarata’, ‘asma’ e ‘ataque cardíaco’), também aparece perto de indícios de problemas de saúde (‘estresse’, ‘fome’ e ‘ruptura ou descolamento de retina’).

Na seção a seguir, apontaremos que muitas das linhas de concordância de ‘sinais’ no subcorpus do MedlinePlus (PT) têm contextos de uso similares aos de ‘sintomas’ no subcorpus do Ministério da Saúde. Além disso, as ocorrências aparecem cercadas pelos mesmos verbos – ‘apresentar’ e ‘poder’.

4.2.1.3.2 ‘Sinais’ e ‘sintomas’ no subcorpus do Ministério da Saúde

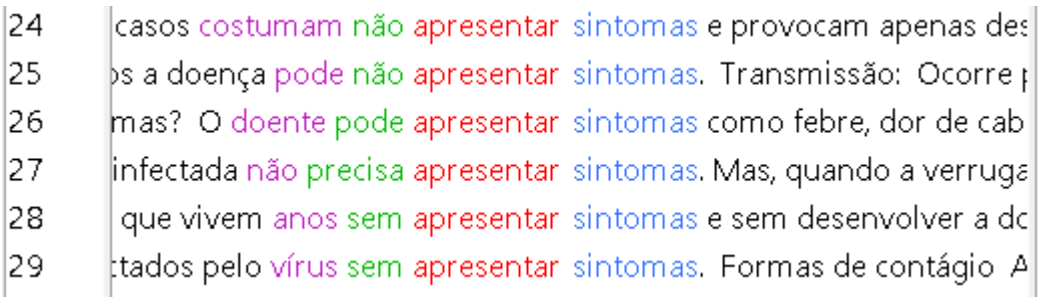
Enquanto ‘sintomas’ ocorre 262 vezes (31,16 a cada 10.000), ‘sinais’ ocorre 42 vezes (4,99 a cada 10.000). No subcorpus do Ministério da Saúde, a palavra ‘sintomas’ aparece com frequência como um subtítulo dos textos, fazendo com que sua frequência aumente significativamente.

Em razão do número elevado de ocorrências, investigou-se o motivo pelo qual a palavra não consta na lista de palavras-chave desse subcorpus. A partir da sua frequência, os índices da palavra foram determinados a partir da calculadora de chavicidade e efeito mencionada anteriormente. O *log-likelihood* da palavra é de 354,16 e o *odds ratio* é de 5,8163, que não alcança o ponto de corte para figurar a lista de palavras-chave. No corpus de referência de artigos científicos, a palavra ‘sintomas’ tem 242 ocorrências (5,37 a cada 10.000). Portanto, devido à recorrência também no corpus de referência, o *odds ratio* da palavra acabou sendo neutralizado, fazendo com que ela não seja considerada chave.

Foram levantados alguns colocados em comum entre as palavras ‘sinal(is)’ (pesquisando por ‘sina*’) no subcorpus de textos do MedlinePlus (PT) e ‘sintoma(s)’ (‘sintoma*’) no subcorpus do Ministério da Saúde. As palavras lexicais que ocorrem à esquerda foram ‘apresentar’ e ‘pode’. Além disso, aparecem ocorrências de palavras gramaticais, como ‘ou’ e ‘se’.

O verbo no infinitivo ‘apresentar’ ocorre 13 vezes como colocado do substantivo ‘sintomas’. Algumas linhas de concordância podem ser observadas na Figura 25.

Figura 25 – Linhas de concordância de ‘apresentar’ como colocado de ‘sintomas’ no subcorpus do Ministério da Saúde

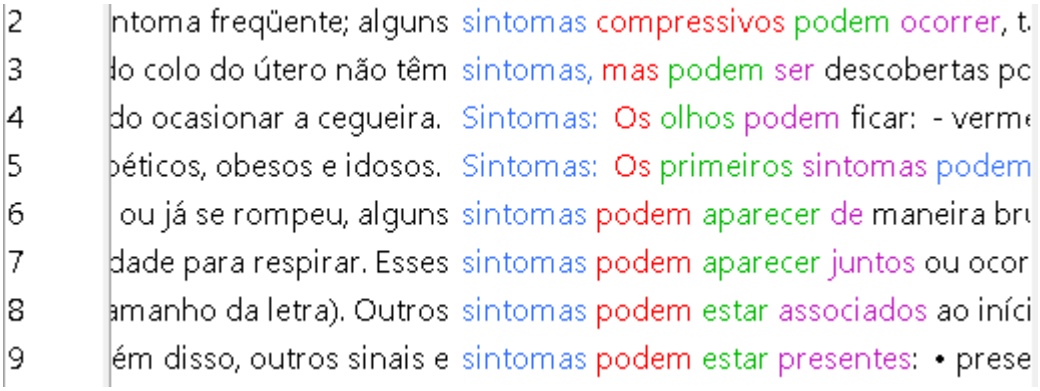


Fonte: AntConc (ANTHONY, 2019).

Já colocadas à esquerda, em comum com o subcorpus do MedlinePlus (PT), estão a forma verbal ‘podem’ e as palavras gramaticais ‘de’, ‘não’, ‘ou’ e ‘que’. A palavra ‘de’ ocorre nas proximidades de ‘sintomas’ 60 vezes. Entretanto, diferentemente do subcorpus do MedlinePlus (PT) – em que essa construção com ‘sinais’ conta com 41 ocorrências –, o padrão ‘sintomas de [$substantivo$]’ tem apenas 11 ocorrências.

Em contraste a isso, semelhante ao uso de ‘sinais’ no subcorpus traduzido, a forma verbal ‘podem’ aparece próxima à palavra ‘sintomas’ 15 vezes no subcorpus do Ministério da Saúde. Na Figura 26, é possível observar algumas linhas de concordância desse colocado.

Figura 26 – Linhas de concordância de ‘podem’ como colocado de ‘sintomas’ no subcorpus do Ministério da Saúde

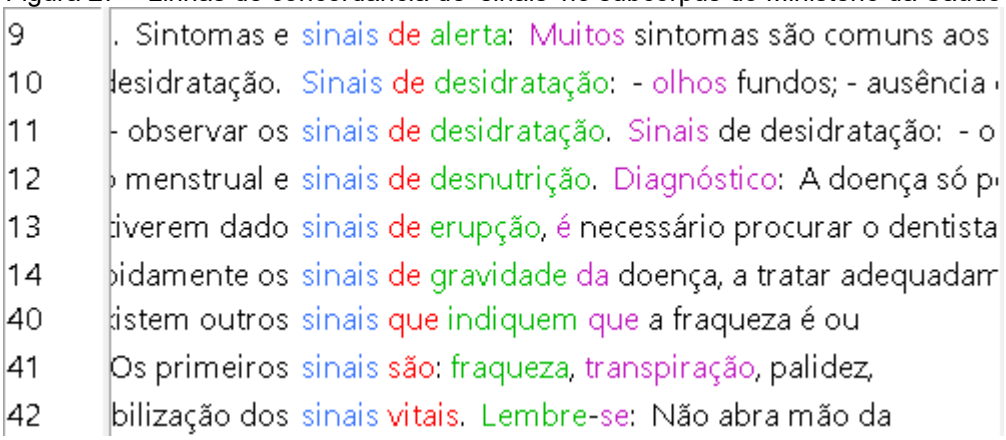


Fonte: AntConc (ANTHONY, 2019).

Há ocorrências da palavra ‘sinais’ que se dão na sequência ‘sinais e sintomas’, com 16 ocorrências. Essa sequência é, muitas vezes, utilizada como um subtítulo para organizar as partes do texto, como mencionado anteriormente sobre a palavra ‘sintomas’. As linhas de concordância de ‘sinais’ no subcorpus do Ministério da Saúde

parecem indicar distinção entre os limites das palavras ‘sintomas’ e ‘sinais’ (Figura 27).

Figura 27 – Linhas de concordância de ‘sinais’ no subcorpus do Ministério da Saúde



Fonte: AntConc (ANTHONY, 2019).

Pode-se notar que, por exemplo, em ‘sinais de desidratação’, a primeira característica é ‘olhos fundos’. Em outro momento, é possível ver a frase ‘Os primeiros sinais são: fraqueza, transpiração, palidez [...]’. Aqui, podemos observar que se usa ‘sinais’ para aspectos visíveis (‘fraqueza’, ‘transpiração’ e ‘palidez’).

Outro aspecto é que, em geral, a palavra ‘sinais’ não aparece associada diretamente às doenças, como no subcorpus do MedlinePlus (‘sinais de asma’; ‘sinais de doença arterial coronariana’; ‘sinais de derrame’; ‘sinais de glaucoma’; dentre outros). A palavra ‘sinais’ é mais associada a manifestações de problemas de saúde, por exemplo, ‘sinais de desidratação’, ‘sinais de desnutrição’, ‘sinais de erupção’, dentre outros.

Em um primeiro momento, apenas pelas linhas de concordância, depreende-se que, de fato, há uma distinção entre ‘sinais’ e ‘sintomas’. A fim de comprovar a existência dessa distinção, optou-se por fazer este mesmo levantamento em um corpus de língua geral.

4.2.1.3.3 ‘Sinais’ e ‘sintomas’ em corpus de língua geral

A fim de buscar mais embasamento sobre as distinções entre os traços que abrangem ‘sinais’ e ‘sintomas’, procedeu-se para a pesquisa em um corpus geral de língua portuguesa. Utilizou-se, novamente, o Corpus do Português, cujo papel, nesse

caso, é de extrema relevância para comprovar os contextos convencionais de uso dessas palavras que são familiares ao público geral.

Como no levantamento de colocados feito nos subcorpora de estudo, manteve-se a janela de até quatro palavras. Esse levantamento foi feito apenas para colocados à direita das palavras ‘sinais’ e ‘sintomas’.

Os primeiros 25 colocados de ‘sinais’ e de ‘sintomas’ que ocorrem nessa janela estão listados na Tabela 14. Pode-se observar suas frequências de co-ocorrência e o resultado estatístico da associação entre as palavras.

Tabela 14 – Colocados de ‘sinais’ e ‘sintomas’ no Corpus do Português

Colocado de ‘sinais’	**P1**	**P2**	**Score**	**Colocado de ‘sintomas’**	**P2**	**P1**	**Score**
prodígios	491	0	983,9	TPM	159	0	317,4
tempos	354	0	709,4	psicóticos	157	0	313,4
maravilhas	277	0	555,1	hemolítico	112	0	223,6
vitais	243	0	486,9	urêmico	110	0	219,6
trânsito	233	0	466,9	sujeitos	93	0	185,6
elétricos	181	0	362,7	poliúria	84	0	167,7
libras	152	0	304,6	relatadas	76	0	151,7
rádio	149	0	298,6	artrite	150	1	149,7
distintivos	133	0	266,5	TDAH	75	0	149,7
pontuação	133	0	266,5	mosquito	62	0	123,8
emitidos	109	0	218,4	tratamentos	57	0	113,8
aparições	98	0	196,4	fibromialgia	51	0	101,8
gráficos	86	0	172,3	neuróticos	51	0	101,8
céu	81	0	162,3	queixas	87	1	86,8
seguirão	81	0	162,3	taquicardia	43	0	85,8
digitais	80	0	160,3	menopausa	167	2	83,3
sol	77	0	154,3	gastrite	39	0	77,8
vinda	74	0	148,3	cabeça	72	1	71,9
luminosos	147	1	147,3	ascensão	34	0	67,9
enviados	70	0	140,3	histéricos	34	0	67,9
sonoros	70	0	140,3	sinto	66	1	65,9
arrombamento	68	0	136,3	psiquiátricos	33	0	65,9
milagres	131	1	131,3	duram	32	0	63,9
espécie	112	1	112,2	alérgicos	31	0	61,9
terra	56	0	112,2	pré-menstrual	30	0	59,9

Fonte: elaborada pela autora com base em dados do Corpus do Português (DAVIES, 2015).

Por meio do levantamento nesse corpus, foi possível confirmar que há distinção entre as manifestações de ‘sinais’ e de ‘sintomas’. A palavra ‘sinais’ aparece associada a ‘trânsito’, ‘aparições’, ‘gráficos’ e ‘arrombamento’. Isso dá indícios de que a maneira como os ‘sinais’ se manifestam é de forma visual. Ou seja, são traços que você pode observar e enxergar. Por exemplo, ‘sinais de arrombamento’ em uma casa

podem ser portas quebradas, com vestígios de terem sido forçadas por alguém, bagunça nos cômodos, demonstrando que alguém esteve por ali procurando algo. Logo, de acordo com os colocados, ‘sinais’ são traços que podem ser identificados por uma terceira pessoa observadora.

Além disso, podem ser observados colocados de ‘sinais’ mais relacionados à noção física de um conjunto que carrega informações ou dados. Aparecem colocados como ‘elétricos’, ‘rádio’, ‘emitidos’, ‘digitais’, ‘luminosos’, ‘enviados’ e ‘sonoros’, que indicam haver uma influência de contextos mais especializados.

Já associadas à palavra ‘sintomas’, há ocorrências como ‘psicótico’, ‘poliúria’, ‘neuróticos’, ‘taquicardia’, ‘histéricos’, ‘psiquiátricos’, ‘alérgicos’ e ‘pré-menstrual’. Ou seja, essas são manifestações que as pessoas sentem mais do que visualizam. Assim, em oposição às manifestações de sinais, que podem ser observadas por uma terceira pessoa, as manifestações de sintomas parecem ser mais facilmente identificadas pela própria pessoa. Até porque, no geral, antes de procurar atendimento que confirme a doença, é necessário que o paciente reconheça os sintomas, para então partir para a análise de uma terceira pessoa (médico), para quem serão relatadas as manifestações. Além dessas manifestações, há algumas doenças e distúrbios que estão na lista de colocados, como ‘[síndrome] hemolítico urêmica’, ‘artrite’, ‘TDAH’ (Transtorno do Déficit de Atenção com Hiperatividade), ‘fibromialgia’ e ‘gastrite’.

Apesar dessa constatação, no subcorpus traduzido do MedlinePlus, a palavra ‘sinais’ parece ser utilizada de forma indiscriminada, aparecendo diversas vezes como um falso sinônimo de ‘sintomas’ e como equivalente de ‘*signs*’. Já a palavra ‘sintomas’, por se tratar de uma palavra motivada pelo uso de ‘*symptoms*’ ou por Ø, está sendo empregada de forma semelhante aos aspectos aqui apresentados.

4.2.1.4 O caso da forma verbal ‘poderá’

‘Poderá’ também é palavra-chave exclusiva do subcorpus de textos traduzidos, ocorrendo 93 vezes (23,56 a cada 10.000); portanto, foi feita a análise de seus colocados. Como colocados à esquerda, apareceram o pronome ‘você’, a palavra ‘também’ e os substantivos ‘médico’ e ‘cirurgia’. À direita, apareceram as palavras ‘lhe’, ‘banho’ e os verbos infinitivos ‘ter’, ‘sentir’, ‘tomar’ e ‘ser’.

O levantamento de colocados dessa palavra no subcorpus do Ministério da Saúde não levou a resultados conclusivos, pois a forma verbal ‘poderá’ tem apenas 25 ocorrências (2,97 a cada 10.000). Pelas linhas de concordância, foi possível observar que à esquerda aparece, também, a palavra ‘médico’. À direita também se observou a ocorrência de alguns verbos, como ‘ser’, ‘indicar’, ‘solicitar’, dentre outros.

A fim de investigar a origem dessa diferença tão alta de ocorrências de ‘poder’ na 3ª pessoa do singular no futuro do presente do indicativo, foi utilizado o *software* AntPConc para procurar o que havia por trás dessa palavra nos textos originais (Figura 28).

Figura 28 – Linhas de concordância de ‘poderá’ no corpus paralelo do MedlinePlus (PT-EN)

Line	KWIC
2	Você poderá apresentar outros sinais de asma se:
3	No dia da cirurgia você poderá chupar cubos de gelo ou tomar líquidos transp
4	Você poderá comer e beber normalmente.
5	Você poderá cortar as pontas quando começarem a enrolar
6	Controle sua dor, assim você poderá cuidar de si mesma, de seu bebê e ser ativa.
7	Você poderá eliminar alguns coágulos de sangue.
8	▪ Alguém da equipe médica poderá ensiná-la a tossir, respirar fundo ou usar um e
9	Agindo rapidamente em relação ao choro, você poderá evitar que o bebê fique muito angustiado.
Line	**Reference**
2	You may have more signs of asthma if you:
3	The day of surgery, you will be able to have ice chips or clear fluids.
4	You can eat and drink your normal diet.
5	You may trim the edges as they curl.
6	Manage your pain so you can care for yourself, your baby and be active.
7	You may pass small blood clots.
8	▪ The staff may teach you how to cough, deep breathe and use an incentive spirometer.
9	By responding to the crying quickly, you may keep him or her from getting too upset.

Fonte: AntPConc (ANTHONY, 2017).

Observou-se, a partir do *software*, que a origem de ‘poderá’ estava nos verbos modais (*modal verbs*) do inglês ‘*may*’, ‘*can*’, ‘*will*’ e ‘*could*’. Os verbos modais são auxiliares que acompanham os verbos principais para expressar algum sentido. Como mencionado anteriormente, os verbos no inglês não sofrem muitas flexões, como ocorre no português. Portanto, é recorrente o emprego desses auxiliares para complementar o sentido da frase, podendo, por exemplo, indicar se algo é uma possibilidade, necessidade, permissão ou sugestão.

O maior número de ocorrências se refere ao modal ‘*may*’, com 66 ocorrências. O modal ‘*may*’ tem 310 ocorrências (89,17 a cada 10.000), sendo que usos de ‘*may be* $[verbo\ no\ particípio]$’ foram majoritariamente traduzidos como

‘[$conjugação\ de$ poder] ser [$verbo\ no\ particípio$]’ e usos de ‘*may* [$verbo$]’ foram traduzidos como ‘[$conjugação\ de\ poder$] [$verbo\ no\ infinitivo$]’, como é possível observar na Figura 28. No sentido em que o verbo está empregado nos textos, seus usos exprimem permissão.

Em segundo lugar, ‘poderá’ aparece como tradução de ‘*can*’, com 17 ocorrências. ‘*Can*’ tem 240 ocorrências (69,03 a cada 10.000) no subcorpus de textos originais. Outras formas utilizadas para traduzir ‘*can*’ foram algumas conjugações do verbo ‘poder’ – como ‘posso’, ‘pode’, ‘podem’ etc. – seguido de verbos no infinitivo. Além disso, em ocorrências de ‘*can be* [$verbo\ no\ particípio$]’, repete-se o mesmo padrão utilizado nesse tipo de construção com ‘*may*’. O verbo modal ‘*can*’, nos contextos em que está sendo empregado, é utilizado para apontar possibilidade.

O terceiro modal mais traduzido como ‘poderá’ é ‘*will*’, com 7 ocorrências. O verbo modal ‘*will*’ conta com 236 ocorrências (67,88 a cada 10.000) no subcorpus em inglês. Em comparação com os números entre ocorrências no original e em traduções, ‘poderá’ é o que tem menos proporção de emprego. Isso porque seus sentidos expressam mais certeza, tratando de uma ação futura que irá, de fato, ocorrer. Portanto, a maioria das ocorrências de ‘*will* [$verbo$]’ foram traduzidas para ‘[$verbo\ no\ futuro\ do\ presente\ do\ indicativo$]’ ou ‘irá [$verbo\ no\ infinitivo$]’.

Por fim, o quarto e menos traduzido como ‘poderá’ foi ‘*could*’, com apenas 3 ocorrências. Nos originais, ‘*could*’ tem poucas ocorrências, aparecendo apenas 13 vezes (3,74 a cada 10.000). Provavelmente, isso ocorre porque o modal é empregado para dar sentido de possibilidade, sendo que esse já é o papel de ‘*can*’. Entretanto, o verbo modal ‘*could*’ expressa possibilidades remotas, com menos probabilidade de acontecerem. A outra opção utilizada para traduzir ‘*could*’ foi a forma verbal ‘poderia’, que demonstra bem a ideia de possibilidade remota.

Como mencionado anteriormente, ‘poderá’ ocorre apenas 25 vezes no subcorpus de textos originalmente escritos em português. Já a construção ‘irá [$verbo\ no\ infinitivo$]’ ocorre apenas 4 vezes nos textos do Ministério da Saúde.

O uso mais recorrente no subcorpus do Ministério da Saúde é de ‘[$conjugação\ de\ 'poder'$] ser [$verbo\ no\ particípio$]’, sendo que o verbo ‘poder’ ocorre majoritariamente no presente, variando para plural (‘podem’) e singular (‘pode’). Esse uso tem 195 ocorrências. Além disso, há grande recorrência de verbos no futuro do presente do indicativo, com 147 ocorrências dos mais variados verbos. Novamente,

os resultados apontam forte indício de que ocorre uma influência do texto-fonte no produto final da tradução.

4.2.2 Levantamento de n-gramas

Para fazer o levantamento de n-gramas, as especificações utilizadas foram de sequências de no mínimo três e no máximo cinco palavras. Foi estabelecida como frequência mínima 10 ocorrências, sendo que a distribuição das ocorrências precisaria ser em no mínimo cinco textos diferentes.

A organização das sequências de palavras foi feita pela distribuição delas nos subcorpora. Ou seja, em quanto mais textos a sequência ocorrer, mais perto do topo ela estará na lista. Essa escolha se deu devido ao interesse de olhar para aquelas sequências mais frequentes nos subcorpora como um todo, e não para sequências que ocorrem frequentemente dentro de textos específicos.

Após feito o levantamento dos n-gramas, procedeu-se para a limpeza da lista. Diversos n-gramas acabaram se repetindo por estarem contidos uns dentro de outros. Por esse motivo, as sequências que estavam contidas dentro de outras foram excluídas se o seu número de ocorrências fora da sequência maior fosse menor que 10. Por exemplo, no subcorpus em inglês do MedlinePlus, o n-grama '*to your doctor*' tem 37 ocorrências, sendo que 30 delas são como parte do n-grama '*talk to your doctor*'. Fora desse n-grama maior, '*to your doctor*' ocorre apenas 7 vezes, portanto não alcança o ponto de corte para figurar na lista. Além disso, observou-se que algumas sequências de palavras eram complementares a outras. No subcorpus do MedlinePlus (PT), o n-grama 'em contato com seu médico' foi unido a 'entre em contato com seu', pois ambos têm 12 ocorrências, indicando que as sequências se complementam para formar 'entre em contato com seu médico'.

A Tabela 15 apresenta a lista completa de n-gramas que corresponderam aos critérios delimitados.

Tabela 15 – Lista de n-gramas dos corpora de estudo

Freq.	**MedlinePlus (EN)**	**Freq.**	**MedlinePlus (PT)**	**Freq.**	**Ministério da Saúde**
91	*if you have*	43	com seu médico	53	o que é
38	*call your doctor*	58	o seu médico	49	o uso de
33	*if you are*	32	se você tiver	46	é uma doença

Freq.	MedlinePlus (EN)	Freq.	MedlinePlus (PT)	Freq.	Ministério da Saúde
30	*talk to your doctor*	19	entre em contato com	26	de acordo com
34	*you may have*	28	para o seu	29	dor de cabeça
26	*you may be*	27	ligue para o	25	por meio de
21	*if you have any*	15	a maioria das	29	o risco de
20	*your doctor will*	20	converse com seu médico	20	anos de idade
19	*if you have a*	16	seu médico ou	22	fatores de risco
22	*signs of a*	18	médico se você	19	para evitar a
16	*you need to*	16	para o seu médico	21	qualidade de vida
30	*you will be*	17	ao seu médico	18	o tratamento é
15	*your doctor if you*	18	com o seu	14	o aparecimento de
12	*if you feel*	13	dificuldade para respirar	15	a quantidade de
14	*talk to your doctor about*	12	os sinais de	19	em caso de
13	*you do not*	11	se você estiver	20	o consumo de
17	*your doctor may*	13	seu médico se você	12	a maioria das
12	*call your doctor right away if you have*	11	é uma doença	12	a presença de
11	*any of these*	10	a maioria das pessoas	12	de bebidas alcoólicas
18	*by your doctor*	38	anos de idade	12	de uma pessoa
16	*may be given*	16	com o seu médico	14	em contato com
11	*not go away*	15	dor de cabeça	16	para a saúde
12	*see your doctor*	11	o que você	13	para que o
11	*such as a*	10	o risco de	13	perda de peso
11	*tell your doctor*	11	seu médico imediatamente	19	serviço de saúde
14	*your doctor or nurse*	16	todos os anos	11	alguns tipos de
10	*do not have*	40	após a cirurgia	11	mais comuns são
13	*while you are*	12	entre em contato com seu médico	12	que pode ser
10	*do not drive*	14	ligue para o seu	16	sinais e sintomas
10	*do not use*	10	médico imediatamente se	10	a maioria dos
11	*in the hospital*	11	é chamado de	10	de acordo com o
10	*may need to*	10	de anos de idade	10	de saúde mais
10	*take your medicines*	11	falta de ar	10	deve ser feito
23	*there is a*	11	horas após a	11	em alguns casos
10	*to see if*	11	o consumo de	11	entre e anos
14	*with soap and water*	13	todos os dias	11	maioria das vezes
10	*you can do*	10	com seu médico sobre	10	maioria dos casos
13	*ask your doctor*	16	de uma reação	10	no caso de
10	*be able to*	10	do seu médico	10	o uso de medicamentos
11	*high blood pressure*	11	nos estados unidos	16	profissional de saúde
10	*you may need*	10	problemas de saúde	12	quando a pessoa

Freq.	MedlinePlus (EN)	Freq.	MedlinePlus (PT)	Freq.	Ministério da Saúde
10	*you will be given*	13	a anos de idade	10	uma vez que
11	*you will need to*	15	a vacina contra	11	é importante que
10	*for at least*	13	as pessoas que	10	é mais comum
13	*swelling of the*	11	com água e sabão	11	a partir dos
16	*a severe allergic reaction*	11	de sangue para	10	com o uso de
12	*after your surgery*	10	fluxo de sangue	10	levar à morte
11	*blood flow to*	10	minutos a algumas horas	10	os sintomas são
10	*can give you*	16	tomar a vacina	12	uma alimentação saudável
10	*months of age*	18	uma reação alérgica grave	10	à base de
12	*of a vaccine*	12	www cdc gov	11	da pressão arterial
10	*people who are*	11	cirurgia se você	10	de saúde para
11	*website at www*	11	da vacina contra	10	que a pessoa
11	*an allergic reaction*	10	dose da vacina	15	o câncer de
10	*the national vaccine injury compensation program*	11	durante a cirurgia	23	colo do útero
10	*of the blood*	10	meses de idade	10	céu da boca
10	*surgery if you*	10	o seu bebê		
10	*through years of age*				
29	*your health care provider*				
11	*your nurse will*				

Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

Ademais, após a lista de n-gramas estar organizada, utilizou-se novamente a ferramenta do Excel que aponta células em comum entre colunas. Portanto, estão grifadas em amarelo as sequências de palavra que ocorrem nas duas listas levantadas do corpus comparável.

4.2.2.1 N-gramas no corpus do MedlinePlus (EN-PT)

No que diz respeito aos textos do MedlinePlus, é possível observar semelhanças entre os n-gramas dos originais e das traduções. Estes textos parecem mais focados em orientar o leitor a procurar ajuda de um médico, pois há vários n-gramas que contam com a palavra ‘médico’ ou ‘*doctor*’ e ‘enfermeira’ ou ‘*nurse*’. Há, por exemplo, no inglês, ‘*call your doctor*’ e ‘*talk to your doctor*’, enquanto no português observamos ‘com seu médico’ e ‘com o seu médico’, que vêm seguidos na lista por sequências como ‘entre em contato com’ e ‘ligue para o’, que apontam para

sequências que se complementam. Inclusive, existe no MedlinePlus (EN) uma sequência mais longa de palavras, que se repete em 11 dos 66 textos: '*call your doctor right away if you have*'.

No português há recorrência de alguns sintomas, como 'dificuldade para respirar', 'dor de cabeça', 'falta de ar' e 'uma reação alérgica grave' – que é o equivalente do inglês '*a severe allergic reaction*' ou '*an allergic reaction*'. Já no inglês, além de '*a severe allergic reaction*' e '*an allergic reaction*', podemos observar sequências de palavras que apontam sintomas como '*high blood pressure*' e '*swelling of the*'.

Ademais, ocorre a construção '*signs of a*', que no português ocorre como 'os sinais de'. Como mencionado anteriormente em relação à palavra 'sinais', o seu uso parece ocorrer de forma indiscriminada, abarcando contextos tanto de usos convencionais de 'sinais' quanto de 'sintomas'.

Na lista do subcorpus traduzido, há n-gramas que são traduções *prima facie* ou equivalentes de sequências do inglês. A Tabela 16 lista todos os equivalentes encontrados no corpus paralelo.

Tabela 16 – Lista de n-gramas equivalentes do corpus paralelo do MedlinePlus (EN-PT)

MedlinePlus (EN)	**MedlinePlus (PT)**
if you have	se você tiver
if you are	se você estiver
talk to your doctor	converse com seu médico
signs of a	os sinais de
your doctor if you	seu médico se você
talk to your doctor about	com seu médico sobre
with soap and water	com água e sabão
a severe allergic reaction	uma reação alérgica grave
blood flow [*to*]	fluxo de sangue
months of age	meses de idade
people who are	as pessoas que
surgery if you	cirurgia se você
[*through*] *years of age*	anos de idade

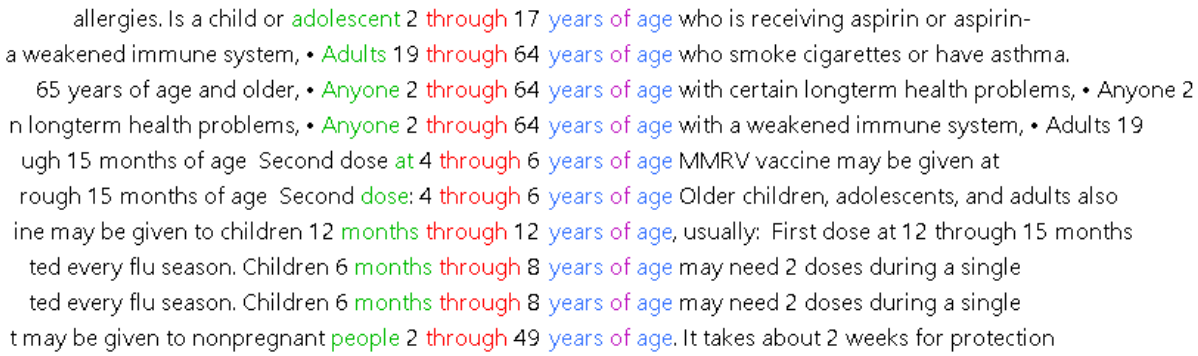
Fonte: elaborada pela autora com base em dados do AntConc (ANTHONY, 2019).

A partir da tabela, nota-se que, na tradução de '*talk to your doctor about*', está indicado 'com seu médico sobre'. Apesar de não aparecer nos n-gramas por não estar dentro do ponto de corte, as formas verbais que aparecem ao lado de 'com seu médico sobre' são 'fale' e 'converse', que são duas traduções possíveis para o verbo '*talk*'. A

tradução para ‘fluxo de sangue’ foi ‘*blood flow* [*to*]’, já que, das 10 ocorrências de ‘fluxo de sangue’, 6 delas eram seguidas de ‘para’ – novamente, não atingindo o ponto de corte para ser considerada um n-grama. Vale ressaltar, também, conforme observado por meio do AntPConc, que há 3 ocorrências de ‘*blood flow*’ que foram traduzidas para ‘fluxo sanguíneo’.

Por último, na tabela, constata-se a sequência ‘*through * years of age*’ cujo equivalente é ‘anos de idade’. Essa sequência de palavras, em inglês, é utilizada para fazer referência a intervalos de idade. Para ilustrar, na Figura 29, há linhas de concordância com aplicações da sequência.

Figura 29 – Linhas de concordância de ‘*through * years of age*’ no subcorpus do MedlinePlus (EN)



Fonte: AntConc (ANTHONY, 2019).

Como se pode observar, ela é construída da seguinte forma: ‘$[substantivo]$ x *through* y *years of age*’, sendo que x e y são números que representam um intervalo de idade. No português, há mais de uma tradução para essa construção. Essas traduções foram localizadas utilizando-se o *software* AntPConc. Os equivalentes encontrados foram: ‘$[substantivo]$ de x a y anos de idade’; ‘$[substantivo]$ entre x e y anos de idade’; ‘$[substantivo]$ com x a y anos de idade’. Ainda, em uma ocorrência, observou-se a alternativa de substituir a construção por: ‘$[substantivo]$ menores de $y + 1$ anos de idade’.

4.2.2.2 N-gramas no subcorpus do Ministério da Saúde

No que diz respeito ao subcorpus de textos escritos originalmente em língua portuguesa, publicados pelo Ministério da Saúde, já é possível constatar a ausência de sequências de palavras que abarquem a palavra ‘médico’. Nesse sentido, observa-

se a ocorrência de ‘serviço de saúde’ e ‘profissional de saúde’, que é o mais próximo de menções a atendimento médico.

Assim como no subcorpus do MedlinePlus (PT), também há algumas sequências de palavras que representam sintomas de doenças, por exemplo, ‘dor de cabeça’, ‘fatores de risco’ e ‘perda de peso’. Além disso, ocorre ‘levar à morte’, que está relacionado às consequências de algumas doenças.

Há, também, ocorrências ligadas a tratamentos das doenças. Alguns exemplos disso são ‘qualidade de vida’, ‘o uso de medicamentos’ e ‘uma alimentação saudável’. Ainda, encontram-se sequências que remetem a partes do corpo, como ‘colo do útero’ e ‘céu da boca’.

Por fim, dentre os n-gramas do Ministério da Saúde, aparece ‘sinais e sintomas’. A sequência corrobora que, de fato, essas duas palavras não são intercambiáveis, mas sim manifestações que se dão de forma diferente. Das 16 ocorrências da sequência, 7 são empregadas dentro do corpo do texto. Por outro lado, é utilizada 9 vezes como subtítulo que introduz esse assunto nos textos. Como mencionado na seção 3.1.2, os textos que compõem o subcorpus do Ministério da Saúde contam com estruturas bastante similares. Adicionalmente, a sequência ‘o que é’ também é utilizada como subtítulo, introduzindo a parte do texto em que a doença é explicada e descrita. Outros exemplos de subtítulos bastante empregados são: ‘diagnóstico’; ‘tratamento’; e ‘prevenção’.

A seguir, discutiremos os resultados apresentados neste capítulo. A discussão se dividirá entre aspectos macroestruturais e microestruturais dos corpora.

5 DISCUSSÃO

Para procedermos à discussão dos resultados, foi feita uma divisão entre aspectos macro e microestruturais dos corpora analisados. Nos aspectos macroestruturais serão abordados: o número médio de palavras por texto; a estrutura de subtítulos dos textos e de lista de itens; o Índice Flesch e a adequação dos textos ao público geral; e a riqueza lexical. Os aspectos da parte microestrutural a serem analisados são relacionados ao vocabulário empregado nos textos, que foram apontados na parte qualitativa da pesquisa.

5.1 ASPECTOS MACROESTRUTURAIS

Em relação à macroestrutura do trabalho, primordialmente serão feitos apontamentos sobre a estrutura de organização dos textos. Como mencionado no capítulo de metodologia, os corpora de estudo deste trabalho contam com estruturas bastante similares. Por exemplo, a grande maioria dos textos tem subtítulos dividindo os diferentes tópicos, como: ‘O que é’, ‘Sinais e sintomas’, ‘Diagnóstico’ (no Ministério da Saúde); e ‘Causas’, ‘Sinais’, ‘Tratamento’ (no subcorpus traduzido do MedlinePlus). Há, também, listas de itens introduzidas por elementos gráficos, como previsto nas regras de escrita acessível de DuBay (2004), que tornam as informações visualmente mais claras.

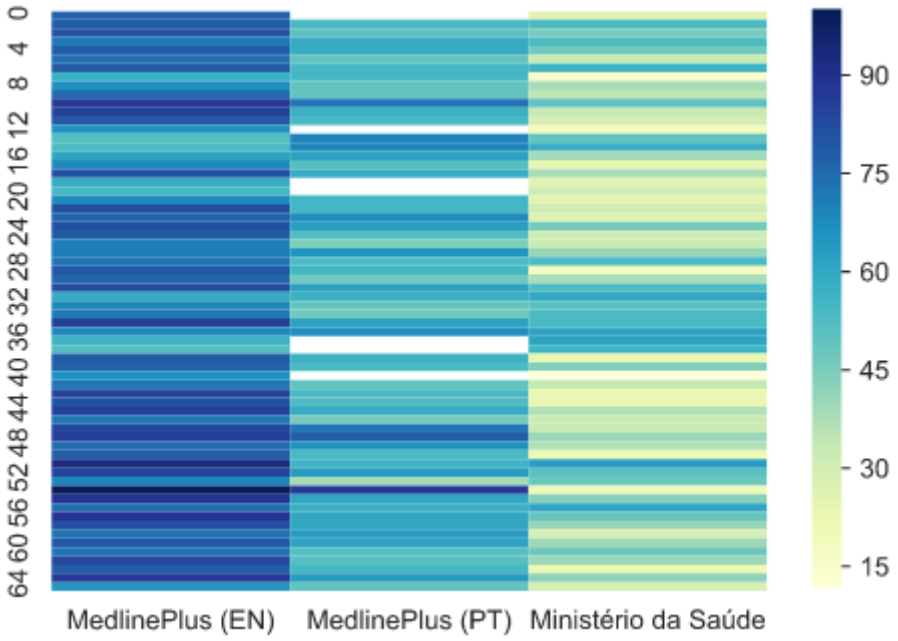
O segundo aspecto a ser analisado é o número médio de palavras por texto. A média de palavras por textos no subcorpus do MedlinePlus (EN) é de 526,76. Nos textos traduzidos, esta média aumenta para 598,12. Já no subcorpus de textos escritos em português, do Ministério da Saúde, a média de palavras por texto fica em 440,23.

Os números referentes aos subcorpora escritos originalmente nas línguas vão ao encontro do que foi observado por Fuchs (2018) em sua pesquisa, demonstrando que textos escritos em inglês têm a tendência de serem mais longos do que em português. Entretanto, as traduções para o português do MedlinePlus (PT) causam uma quebra dessa tendência, contando com uma média de palavras por texto maior do que tinham os originais em inglês. Portanto, esperar-se-ia que as traduções para o português apresentassem média de palavras mais baixa do que seus originais em

língua inglesa, ocorrendo a implicitação de informações que poderiam ser consideradas desnecessárias para os falantes de português.

O terceiro aspecto macroestrutural a ser apontado é referente aos índices que estimam a inteligibilidade dos textos. Retomando as médias de Índice Flesch, para a língua portuguesa, obtiveram-se índices de 57,659 para o subcorpus do MedlinePlus (PT) e 39,115 para o do Ministério da Saúde; para a língua inglesa, a média observada foi de 74,845 para o subcorpus do MedlinePlus (EN). Lembrando que índices mais próximos de 100 apontam para maior grau de facilidade, enquanto mais próximos de 0 demonstram maior grau de dificuldade. O Gráfico 9 apresenta os índices de todos os textos analisados pelas ferramentas Coh-Metrix e Coh-Metrix-Port classificando os índices por meio de cores.

Gráfico 9 – Dados do Índice Flesch



Fonte: elaborada pela autora com base em dados do Coh-Metrix (GRAESSER *et al*., 2017) e do Coh-Metrix-Port (NILC, 2020).

Com base no gráfico, depreende-se que praticamente todos os textos do subcorpus em inglês contam com índices entre 60 e 100, ficando predominantemente classificados em cor azul. Já os textos traduzidos para o português se situam entre as cores azul e verde, sendo que em branco estão os textos que a ferramenta não processou. Maior parte dos textos do Ministério da Saúde foram classificados em índices próximos a 30, representados na cor verde. Como mencionado na seção 4.1,

com base nos níveis de escolaridade das populações estadunidense e brasileira, o índice de inteligibilidade dos textos em português não estão adequados para o público geral. O intervalo mais adequado dos índices para os textos em português seria de classificação ‘fácil’, entre 70 e 100. Já os textos em inglês, para serem acessíveis à maior parcela da população estadunidense, poderiam apresentar inteligibilidade ‘difícil’, com índices entre 30 e 60.

Vale ressaltar que, de acordo com Graesser *et al*. (2004), as fórmulas matemáticas de inteligibilidade e de avaliação de complexidade, como a de Flesch, ignoram componentes linguísticos e discursivos que influenciam a compreensão textual. Dessa forma, apesar de os parâmetros de tamanho de sentenças e de palavras serem válidos, eles não conseguem revelar com precisão a complexidade de um texto. Por esse motivo, se viu a necessidade de aliar a pesquisa quantitativa à pesquisa qualitativa – que, aliás, conta com métricas quantitativas para ser feita. Para demonstrar exemplos de que a métrica está sujeita a falhas, pode-se citar um aspecto do texto *About Your Pain*, do subcorpus do MedlinePlus (EN):

> Pain is the body’s way of sending a message to your brain that help is needed. Tell your doctor or nurse about your pain so they can keep you comfortable. These are questions you may be asked about your pain:
> • Where is your pain? Point to the place on your body where it hurts.
> • Does the pain spread to other parts of your body?
> • When did the pain start?
> • How much does it hurt? Point to a number or face that shows us how much pain you are having.
> • What does it feel like? Does it burn, tingle or ache? Is it dull or sharp? Is it constant or does it come and go?
> • Is it worse at any time of the day? Morning? Evening?
> • What makes the pain feel better? What makes the pain feel worse? What have you done to try to relieve the pain? Does the medicine make it feel better? Does it hurt more when you are active or lying still?
> • Does the pain affect other parts of your life? Does it make it hard to sleep, eat, or care for yourself or others? Does it cause you to be upset, cry or to be less patient? (U.S. NATIONAL LIBRARY OF MEDICINE, 2020, on-line).

Esse é o texto que a ferramenta apontou o Índice Flesch mais alto, de 99,821. O índice está muito próximo de 100, se enquadrando como um texto ‘muito fácil’. Apesar de o texto contar com vocabulário repetido, vale ressaltar que o índice tão alto não foi devido a esse motivo. O texto apresenta vários questionamentos e, em algumas das vezes, estes são expressos por uma só palavra, como em ‘*Morning? Evening?*’. A ferramenta conta as exclamações, interrogações e pontos finais para apontar o fim de uma frase e o início de outra. É por isso que o índice apontado para

o texto foi ‘muito fácil’. Isso demonstra a necessidade de análises textuais qualitativas aliadas ao método quantitativo de análise de inteligibilidade. A discussão dos resultados qualitativos será apresentada na próxima seção.

O quarto e último aspecto da macroestrutura a ser descrito é a riqueza lexical (ou *type-token ratio*). O valor obtido no cálculo da riqueza lexical no subcorpus do MedlinePlus (EN) foi de 8,88%. Como já foi dito, os índices de *type-token ratio* de línguas diferentes não são comparáveis, pois as línguas se distanciam em vários níveis. Em vista disso, os índices comparáveis do português foram: para os textos traduzidos do MedlinePlus (PT), o valor de 11,53%; para o subcorpus de originais do Ministério da Saúde, 11,49%. Portanto, a riqueza lexical do corpus comparável é muito próxima, com uma diferença de apenas 0,03%. Isso significa dizer que a distribuição do universo de palavras repetidas e palavras únicas dos subcorpora é bem equilibrada, tendo nível semelhante de repetitividade.

Em relação aos aspectos macroestruturais, dois dos aspectos tiveram diferenças significativas – Índice Flesch e média de tamanho dos textos –, enquanto os outros dois se aproximaram – estrutura dos textos e riqueza lexical.

Na próxima seção, será apresentada a discussão em relação aos aspectos microestruturais, obtidos a partir do levantamento de palavras-chave e n-gramas, portanto, mais relacionados ao vocabulário dos textos.

5.2 ASPECTOS MICROESTRUTURAIS

Nesta seção, os aspectos microestruturais que serão discutidos referem-se aos levantamentos de palavras-chave e n-gramas, feitos por meio do AntConc. Para a análise de palavras-chave, foram escolhidas aquelas relacionadas ao gênero texto de divulgação da área médica, excluindo-se as que estivessem estritamente relacionadas a doenças e seus sintomas. Então, partiu-se para a observação de colocados ou linhas de concordância (quando os colocados eram inconclusivos) dos seguintes itens: as formas verbais ‘use’ e ‘utilize’; o pronome ‘seu’; os substantivos ‘sinais’ e ‘sintomas’; e a forma verbal ‘poderá’.

À direita da palavra ‘use’ no subcorpus do MedlinePlus (PT), ocorreram diversos substantivos. Com recorrência, observaram-se ‘tampões’ – que parece ser

uma tradução *prima facie* do inglês '*tampons*' –, 'absorventes' e 'preservativo(s)'. Viu-se que nos textos escritos originalmente em português, as ocorrências de 'camisinha' são bem mais altas do que de 'preservativo', apontando que o Ministério da Saúde faz uso da palavra de cunho mais popular.

Além disso, as ocorrências de 'use', no MedlinePlus (PT), e de 'utilize', no Ministério da Saúde, ocorrem em contextos similares, em que é dada alguma orientação para o leitor do que fazer quando se deparar com determinadas situações. Descobriu-se que há, de fato, distinção entre aplicações de 'use' e de 'utilize', sendo que o primeiro se refere, principalmente, a objetos (com frequência, itens de vestuário), e o segundo é empregado com noções mais abstratas (como 'créditos', 'pagamento', 'sistemas' etc.). Essa descoberta foi feita a partir das linhas de concordância do subcorpus de textos originalmente escritos em português e, posteriormente, foi validada por meio do Corpus do Português (DAVIES, 2015).

A partir dos colocados de 'seu', foi possível constatar nos textos traduzidos grande incidência da palavra 'médico'. Isso chamou a atenção porque, nos textos do Ministério da Saúde, os principais colocados imediatamente à esquerda de 'médico' são os artigos 'o' e 'um'. Vale ressaltar que das 251 ocorrências (63,58 a cada 10.000) de 'médico' no MedlinePlus (PT), 168 estão precedidas de 'seu'; enquanto no Ministério da Saúde, das 130 ocorrências (15,46 a cada 10.000) da palavra, apenas 23 ocorrem ao lado de 'seu'. Isso demonstra interferência do texto em inglês sobre o texto traduzido, visto que no inglês, das 247 ocorrências (71,05 a cada 10.000) de '*doctor*', 204 estão antecedidas por '*your*'.

Outra clara interferência dos textos-fonte é o uso de 'sinais' e 'sintomas' nos textos traduzidos. Enquanto há distinção entre os usos dessas palavras no subcorpus dos textos escritos originalmente em português, nas traduções há indícios de que a motivação se dê puramente pelo uso de palavras cognatas no texto-fonte. Prova disso é que a grande maioria das ocorrências de '*signs*' foram traduzidas como 'sinais' (83 de 117 ocorrências), enquanto todas as ocorrências de '*symptoms*' foram traduzidas para 'sintomas'. Adicionalmente, quando houve uso de Ø nos originais em inglês (ou seja, não havia emprego nem da palavra '*symptoms*' nem de '*signs*'), aparece o emprego de 'sintomas' nas traduções, fazendo com que a palavra seja utilizada em contextos mais similares entre si. Isso pode ser explicado pela influência do texto original sobre as decisões do tradutor, sendo que, quando há Ø, o tradutor consegue

se desprender do padrão de uso da língua inglesa. Por exemplo, ‘sintomas’ aparece associado a ‘raiva’, ‘nervosismo’ e ‘irritabilidade’ e a ‘febre’ e ‘tosse’, manifestações essas que podem ser sentidas pela pessoa doente.

Pôde ser estabelecida a diferença entre os contextos de uso de ‘sinais’ e ‘sintomas’ a partir do subcorpus do Ministério da Saúde. Posteriormente, buscaram-se colocados também no Corpus do Português (DAVIES, 2015), para fazer a comprovação dos resultados em um corpo de textos que abrange diversas facetas da língua. Depreendeu-se que as ocorrências de ‘sinais’ estão mais ligadas a demonstrações que podem ser vistas, enquanto ‘sintomas’ se refere a sensações.

A última palavra-chave analisada, ‘poderá’, é estatisticamente frequente somente no subcorpus MedlinePlus (PT). A palavra tem 93 ocorrências (23,56 a cada 10.000) no subcorpus traduzido, enquanto no originalmente escrito em português, ela ocorre apenas 25 vezes (2,97 a cada 10.000). Ao fazer o levantamento dos trechos originais em que ‘poderá’ ocorre no MedlinePlus (EN), por meio do AntPConc, foi possível atestar que sua origem está nos verbos modais do inglês. Do maior índice de recorrência para o menor, respectivamente, a forma verbal ‘poderá’ originou-se nos modais ‘*may*’, ‘*can*’, ‘*will*’ e ‘*could*’.

Em sentido análogo, no subcorpus do Ministério da Saúde, a construção mais utilizada para transmitir a mesma ideia de ‘poderá’ foi ‘$[conjugação\ de\ 'poder']$ ser $[verbo\ no\ particípio]$’, ocorrendo 195 vezes. Além disso, também há recorrência de construções empregando verbos no futuro do presente do indicativo, com 147 ocorrências dos mais variados verbos.

Nos levantamentos de n-gramas, também foram encontrados indícios de influências do inglês sobre o português. Por exemplo, nas traduções, há grande recorrência de n-gramas com a palavra ‘médico’ nos textos traduzidos. Em vez disso, nos textos originais em português há referência apenas a ‘profissional de saúde’ e ‘serviço de saúde’. Vale lembrar que, como apontado anteriormente, ‘médico’ ocorre 251 vezes (63,58 a cada 10.000) no MedlinePlus (PT), enquanto no Ministério da Saúde, somente 130 (15,46 a cada 10.000).

Na próxima seção, serão retomadas as hipóteses deste trabalho e algumas das teorias que foram fundamentais para o desenvolvimento da pesquisa.

5.3 RETOMADA DAS HIPÓTESES E DE TEORIAS

Neste momento, serão retomadas as hipóteses da pesquisa, a fim de verificar se os resultados corroboram o que, em um primeiro momento, acreditava-se que aconteceria.

1) Os textos escritos originalmente em português são mais inteligíveis, quando em comparação com os traduzidos.

A análise dos textos originalmente escritos em português no Coh-Metrix-Port apontou nível de complexidade mais elevada do que os traduzidos para o português. Em relação a todos os subcorpora desta pesquisa, respectivamente, o subcorpus de originais em inglês retornou índices de inteligibilidade mais altos, com uma média de 74,845, enquadrando-se na categoria ‘razoavelmente fácil’; o subcorpus de textos traduzidos obteve uma média de 57,659, ficando na categoria ‘razoavelmente difícil’; por último, o subcorpus de originais em português revelou a média de Índice Flesch de 39,115, categorizada como ‘difícil’.

2) Os textos traduzidos apresentam nível de complexidade textual mais alto do que os escritos originalmente em língua portuguesa, apresentando quebras de convencionalidade que dificultam o entendimento.

Entre o corpus comparável, no que diz respeito ao levantamento do Índice Flesch, o subcorpus de textos traduzidos se enquadrou na categoria ‘razoavelmente difícil’, enquanto o de textos originais foi categorizado como ‘difícil’. Por outro lado, em relação à convencionalidade, os textos traduzidos apresentam quebras do que seria esperado como padrão para a língua portuguesa. Um exemplo disso é o uso indiscriminado de ‘sinais’ e ‘sintomas’ nas traduções, que pôde ser atestado como influência dos textos-fonte sobre os textos-alvo.

3) Os textos traduzidos apresentam nível de complexidade textual mais alta do que seus originais em língua inglesa.

De fato, os textos traduzidos obtiveram nível de complexidade textual significativamente mais elevada que os textos originais. A diferença entre as médias dos índices de inteligibilidade foi de 17,186. Essa diferença é significativa, considerando que ela faz com que o subcorpus de originais seja categorizado como ‘razoavelmente fácil’ e o de traduções como ‘razoavelmente difícil’.

De acordo com a abordagem funcionalista da tradução de Nord (2006), a finalidade da tradução deve determinar a metodologia e as estratégias utilizadas para traduzir o texto, sempre mantendo em mente o seu público-alvo e a função que o texto traduzido irá desempenhar. Com isso em mente, os resultados relacionados ao índice de inteligibilidade apontam que a tradução falha em entregar textos que sejam adequados para o público geral brasileiro.

Foi possível apontar inadequações principalmente no que diz respeito à convencionalidade. Retomando o conceito, a noção de convencionalidade diz respeito ao que está consolidado pela prática e obedece a padrões aceitos pelos falantes de uma língua (TAGNIN, 2013). Portanto, é importante apontar que, apesar de ser uma palavra mais curta (para fins de análise de Índice Flesch), 'sinais' não é convencionalmente utilizada para se referir às manifestações de doenças, por exemplo. Escrever 'seu médico' também causa quebra de convencionalidade, sendo mais frequente o emprego de 'um médico' ou 'o médico'. É mais adequado escrever 'procure um médico' ou 'vá ao médico' (exemplos extraídos do subcorpus do Ministério da Saúde) do que 'converse com o seu médico' (exemplo do subcorpus traduzido do MedlinePlus).

Por fim, pode-se depreender dos resultados obtidos por meio da comparação dos colocados das palavras-chave que há grande influência do texto-fonte sobre o texto-alvo no português. Por estar com a mente e o olhar no texto original, o tradutor pode acabar não associando a ocorrência de uma palavra a opções de tradução que vão além do seu cognato, como se observa na tradução de '*use*' pelo cognato 'use'. Esse fenômeno configura o que Baker (1993) denomina "terceira língua" na tradução, que é quando o texto traduzido fica com características do texto-fonte, distanciando-se, assim, de padrões de convencionalidade da língua-alvo.

Com base no que foi dito neste capítulo, ressalta-se a importância do uso de corpora comparáveis para desenvolver a tradução de textos de gênero e área de especialidade específicos, para tornar as traduções mais convencionais. Como exposto na seção 1.1.2, fazer uso de corpora paralelos na tradução pode não ser o ideal, pois, por passar pela mediação de um tradutor, o uso apenas dessa metodologia pode apontar respostas inconclusivas ou até mesmo opções não convencionais.

Neste capítulo, foram promovidas discussões pertinentes aos resultados dos levantamentos feitos no trabalho. Na seção do âmbito macroestrutural, foram discutidos o número médio de palavras por texto, a estrutura dos textos, o Índice Flesch e a riqueza lexical. No âmbito microestrutural, foram abordadas questões referentes ao vocabulário empregado, apontando a falta de convencionalidade dos textos traduzidos para o português.

Em seguida, serão delineadas as considerações finais deste trabalho.

CONSIDERAÇÕES FINAIS

Neste momento, serão tecidas considerações finais do estudo, apontadas limitações enfrentadas no desenvolvimento do trabalho e apresentadas perspectivas futuras desta pesquisa.

Em primeiro lugar, salienta-se a importância do uso de metodologia que associe a análise quantitativa à análise qualitativa. A associação dessas investigações, possibilitada por ferramentas que fazem cálculos estatísticos somadas ao olhar do pesquisador, permite o enriquecimento dos resultados da pesquisa.

Em relação aos levantamentos puramente quantitativos, dois dos aspectos tiveram diferenças significativas: o Índice Flesch e a média de tamanho dos textos. Os resultados do Índice Flesch apontaram como ‘difíceis’ os textos do Ministério da Saúde, como ‘razoavelmente difíceis’ os textos traduzidos do MedlinePlus, e como ‘razoavelmente fáceis’ os originais em inglês do MedlinePlus. Esses resultados contrariam, em parte, as hipóteses estabelecidas para a análise quantitativa, pois esperava-se que os textos apontados como mais difíceis fossem as traduções. Em relação ao tamanho dos textos, a média de palavras por texto no subcorpus do MedlinePlus (EN) é de 526,76; nos textos traduzidos, essa média aumenta para 598,12; já no subcorpus do Ministério da Saúde, a média fica em 440,23. O comprimento médio dos textos traduzidos vai de encontro ao relatado no referencial teórico, onde se aponta que textos em português tendem a ser mais curtos, e textos em inglês, mais longos (FUCHS, 2018).

As traduções rompem com essa tendência, ficando mais longas do que seus textos-fonte. Isso pode decorrer do fato de que o tradutor, no processo de elaboração da tradução, tende a deixar as informações mais explícitas, tendo como resultado traduções mais longas do que seus textos-fonte (BAKER, 1993; FRANKENBERG-GARCIA, 2006).

No que diz respeito aos resultados quantitativos, foi possível traçar algumas conclusões em relação aos levantamentos de palavras-chave e n-gramas. A partir dos levantamentos de colocados das palavras-chave, concluiu-se que os textos traduzidos apresentavam, em diversos momentos, quebras de convencionalidade (SINCLAIR, 1991; TAGNIN, 2013), distanciando-se dos padrões utilizados nos textos escritos originalmente em português. Essas quebras de convencionalidade ocorrem devido ao

uso de palavras cognatas do inglês e de traduções *prima facie*, fugindo dos padrões esperados para o português. O levantamento de n-gramas também mostrou indícios de influências do inglês sobre o português. Isso configura o que Baker (1993) chama de “terceira língua” na tradução, resultando em um texto distante de textos originalmente escritos na língua, carregando consigo traços característicos do texto-fonte.

No que diz respeito às limitações, a maior delas foi o tamanho dos corpora. Ao utilizar corpus paralelo em análises, é comum que seu tamanho não seja tão expressivo quanto de um formado somente por escritos originalmente na língua. Isso ocorre porque o número de textos que se enquadram nos critérios de seleção acaba se reduzindo, pois necessita-se que estejam disponíveis tanto o texto-fonte quanto o texto-alvo. Devido a essa barreira, foi necessário recorrer diretamente à análise das linhas de concordância em momentos que o levantamento de colocados era inconclusivo. Também, foi necessário recorrer a um corpus maior, de língua geral, a fim de confirmar os padrões de colocação de determinadas palavras-chave.

Além disso, há outras limitações impostas ao estudo pelos *softwares* utilizados para analisar os corpora. Em primeiro lugar, é necessário lembrar que o Coh-Metrix-Port estabelece um número máximo de palavras por texto para a análise. O tamanho máximo para um texto ser processado pelo Coh-Metrix-Port é de mil palavras, enquanto o Coh-Metrix coloca como limite 15 mil caracteres. Apesar da diferença entre os parâmetros utilizados, o Coh-Metrix aceita textos mais longos do que o Coh-Metrix-Port. Devido a essa barreira estabelecida pelo Coh-Metrix-Port, sete dos textos do subcorpus do MedlinePlus (PT) não foram analisados pela ferramenta. Em segundo lugar, vale ressaltar que o *software* AntConc não faz nenhum tipo de etiquetagem ou lematização de corpus. Portanto, a fim de reduzir o número de palavras-chave dos subcorpora, foi necessário partir para a lematização manual das palavras.

Como perspectivas futuras, pode-se ampliar os corpora deste estudo. Além de aumentá-los em número de *types* e *tokens*, seria interessante reproduzir a pesquisa investigando se esses resultados também se comprovam em corpus paralelo com textos traduzidos do português para o inglês. Outra abordagem interessante seria

reproduzir a pesquisa com foco em outros gêneros de textos em tradução, principalmente naqueles voltados ao público geral.

Além disso, poder-se-ia analisar as palavras-chave partindo de uma perspectiva cultural. Isso ajudaria a entender, por exemplo, quais os motivos por trás do alto índice de colocação de 'seu' e 'médico' como traduções para '*your*' e '*doctor*', enquanto nos originais em português utilizam-se 'o' ou 'um' como colocados de 'médico'. Também ajudaria a compreender o porquê de 'vacina' ter chavicidade mais alta nos textos traduzidos – sendo que ocorrerem outras variantes, como 'vacinas', 'vacinação' e 'vacinadas' – do que nos originais.

Pode-se, também, apontar a simplificação textual como uma metodologia que teria efeitos positivos na adequação dos textos para o leitor médio. Esse método, que configura uma forma de tradução intralinguística, consiste em uma forma de derrubar barreiras ao entendimento impostas pela língua. A simplificação pode ser feita por intermédio da reformulação do repertório utilizando variedades da mesma língua. Assim, tratando-se de uma forma de tradução intralinguística, o texto traduzido seria visto desprendido do texto-fonte, possibilitando que o especialista aponte o vocabulário mais convencional para facilitar o entendimento do texto pelo leitor médio. Outro recurso pertinente seria o uso de ferramenta que auxilie na simplificação, como é a proposta do MedSimples (TEXTECC, 2020).

Por fim, espera-se que esta pesquisa tenha impactos positivos no que diz respeito às reflexões sobre o processo tradutório. Enfatiza-se, além de pensar a tradução na perspectiva funcionalista, a importância de levar em conta aspectos como convencionalidade, culturas de alto e baixo contexto, dentre outras questões que permeiam a prática tradutória e que foram abordadas neste trabalho.

REFERÊNCIAS

ANTHONY, Laurence. **AntConc**. Versão 3.5.8. Tokyo: Waseda University, 2019.

ANTHONY, Laurence. **AntCorGen**. Versão 1.1.2. Tokyo: Waseda University, 2019b.

ANTHONY, Laurence. **AntPConc**. Versão 1.2.1. Tokyo: Waseda University, 2017.

ASSOCIAÇÃO DOS MAGISTRADOS BRASILEIROS. **AMB lança campanha para simplificar linguagem jurídica**. Brasília, 2005. Disponível em: https://www.amb.com.br/amb-lanca-campanha-para-simplificar-linguagem-juridica/. Acesso em: 8 jul. 2020.

BAKER, Mona. Corpus Linguistics and Translation Studies: Implications and Applications. *In*: BAKER, Mona; FRANCIS, Gill; TOGNINI-BONELLI, Elena (ed.). **Text and Technology**: In Honour of John Sinclair. Philadelphia: John Benjamins, 1993.

BERBER SARDINHA, Tony. **Lingüística de Corpus**. São Paulo: Manole, 2004.

BIDERMAN, Maria Tereza Camargo. Estatística linguística. **Alfa**, São Paulo, v. 11, p. 117-128, 1967.

BRASIL. Ministério da Saúde. **Biblioteca Virtual em Saúde**. Brasília, DF, 2018.

BRASIL. **Lei nº 13.146, de 6 de julho de 2015**. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Brasília, DF: Secretaria-Geral da Presidência da República, 2015.

BRASIL. Ministério da Educação. **Referenciais de acessibilidade na educação superior e a avaliação in loco do sistema nacional de avaliação da educação superior (SINAES)**. Brasília, DF: Ministério da Educação, 2013.

BREZINA, Vaclav. **Statistics in Corpus Linguistics**: A Practical Guide. Cambridge: Cambridge University Press, 2018.

CIAPUSCIO, Guiomar Elena. La terminologia desde el punto de vista textual: selección, tratamento y variación. **Organon**, v. 12, n. 26, 1998.

DAVIES, Mark. **Corpus do Português**. Provo: Brigham Young University, 2015. Disponível em: https://www.corpusdoportugues.org/. Acesso em: 12 jun. 2020.

DUBAY, William H. **The Principles of Readability**. California: Impact Information, 2004.

EVEN-ZOHAR, Itamar. Polysystem Theory. *In*: Polysystem studies. **Poetics Today**, Durham, vol. 11, n. 1, p. 9-26, 1990.

FANTINUOLI, Claudio; ZANETTIN, Federico. Creating and using multilingual corpora in translation studies. *In*: FANTINUOLI, Claudio; ZANETTIN, Federico. (ed.). **New directions in corpus-based translation studies**. Berlim: Language Science Press, 2015. p. 1–11.

FARKAS, András. **LF Aligner**. Versão 4.2. 2018.

FINATTO, Maria José Bocorny. Acessibilidade textual e terminológica: promovendo a tradução intralinguística. **Revista Estudos Linguísticos**, v. 49, n. 1, p. 72-96, abr. 2020.

FINATTO, Maria José Bocorny. Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. **Organon**, Porto Alegre, v. 25, n. 50, 2011.

FRANKENBERG-GARCIA, Ana. Using a Parallel Corpus in Translation Practice and Research. **Actas da Contrapor**, Lisboa, p. 142-148, 2006.

FUCHS, Sandra Navarro. **Orientações culturais e suas implicações para a tradução funcionalista**: um estudo na área do turismo à luz da Linguística de Corpus. 2018. Tese (Doutorado em Estudos da Tradução) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2018.

FULGÊNCIO, Lúcia; LIBERATO, Yara. **Como facilitar a leitura**: como se processa a leitura. São Paulo: Contexto, 1992.

GABRIELATOS, Costas. Keyness Analysis: nature, metrics and techniques. *In*: MARCHI, Anna; TAYLOR, Charlotte. (ed.). **Corpus Approaches to Discourse**: A Critical Review. London: Routledge, 2018.

GRAESSER, Arthur C. *et al*. **Coh-Metrix**. Version 3.0. Tennessee: University of Memphis, 2017.

GRAESSER, Arthur C. *et al*. Coh-Metrix: Analysis of text on cohesion and language. **Behavioral Research Methods**, v. 36, n. 2, p. 193-202, 2004.

HUNSTON, Susan. **Corpora in Applied Linguistics**. Cambridge: Cambridge University Press, 2002.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Pesquisa Nacional por Amostra de Domicílios Contínua. **Educação 2018**. Rio de Janeiro: IBGE, 2019.

INSTITUTO PAULO MONTENEGRO. **Indicador de Analfabetismo Funcional**. São Paulo, 2017. Disponível em: https://ipm.org.br/inaf. Acesso em: 10 fev. 2020.

INSTITUTO PAULO MONTENEGRO. Indicador de Analfabetismo Funcional. **INAF Brasil 2018**: Resultados Preliminares. São Paulo, 2018.

KRIEGER, Maria da Graça. Divulgação científica e terminologia. *In*: SIMPÓSIO INTERNACIONAL DE ESTUDOS DE GÊNEROS TEXTUAIS, 5., 2009, Caxias do Sul. **Anais** [...]. Caxias do Sul: UCS, 2009.

LEFFA, Vilson J. Fatores da compreensão na leitura. **Cadernos do IL**, Porto Alegre, v.15, n.15, p.143-159, 1996.

LIMA, Kelen Cristina Sant'Anna. **Caracterização de registros orientada para a produção textual no ambiente multilíngue**: um estudo baseado em corpora comparáveis. 2013. 251f. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

LIU, Dilin. Is it a chief, main, major, primary, or principal concern? A corpus-based behavioral profile study of the near-synonyms. **International Journal of Corpus Linguistics**, v. 15, n. 1, p. 56–87, 2010.

MARTINS, Teresa B. F. *et al*. Readability formulas applied to textbooks in Brazilian Portuguese. **Notas do ICMSC**, São Paulo, n. 28, p. 1-11, 1996.

MASSARANI, Luísa; MOREIRA, Ildeu de Castro. A retórica e a ciência: dos artigos originais à divulgação científica. **MultiCiência**, Campinas, n. 4, mai. 2005, p. 01-12.

MAZUR, Beth. Revisiting Plain Language. **Technical Communication**, Virginia, v. 47, n. 2, maio 2000.

MORATO, Rúbia Gomes. **Conceitos Básicos de Estatística Descritiva**. Universidade de São Paulo, 2011.

NATIONAL CENTER FOR EDUCATION STATISTICS. **Adult Literacy in the United States**. Washington, D.C.: U.S. Department of Education, 2019.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. **Coh-Metrix-Port**. Versão 3.0. São Paulo: Universidade de São Paulo, NILC, 2020.

NORD, Christiane. Loyalty and fidelity in specialized translation. **Confluências**, n. 4, p. 29-42, maio 2006.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. **PorSimples**. São Paulo: Universidade de São Paulo, NILC, 2010.

PASQUALINI, Bianca Franco. **Leitura, tradução e medidas de complexidade textual para leitores com nível de letramento básico**. 2012. 141f. Dissertação (Mestrado em Estudos da Linguagem) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

PARAGUASSU, Liana Braga. **Tradução Especializada Acessível (TEA)**: revisão do tema e proposta de disciplina para cursos de graduação em tradução. 2018. Dissertação (Mestrado em Estudos da Linguagem) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

PARAGUASSU, Liana Braga *et al*. MedSimples: an automated simplification tool for promoting health literacy in Brazil. *In*: Digital Humanities and Natural Language Processing, 2020, Évora - Portugal. **Workshop**: Digital Humanities and Natural Language Processing. Aachen: CEUR- WS, 2020. v. 1. p. 78-80.

PLAIN ENGLISH CAMPAIGN. **About us**. 2020. Disponível em: http://plainenglish.co.uk/about-us.html. Acesso em: 3 jul. 2020.

POJANAPUNYA, Punjaporn; TODD, Richard Watson. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. **Corpus Linguistics and Linguistic Theory**, v. 14, n. 1, p. 133-167, 2018.

REBECHI, Rozane Rodrigues. **A Tradução da Culinária Típica Brasileira para o Inglês**: um Estudo sob o Enfoque da Linguística de Corpus. 2015. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2015.

REBECHI, Rozane Rodrigues. Fraseologias bilíngues português-inglês da culinária brasileira: estudo direcionado pelo corpus. *In*: RIBEIRO, Emílio Soares; TABOSA, Leila Maria Araújo; SILVA, Nilson Roberto Barros (org.). **Tradução em três vertentes**: teoria e prática, intersemiose e Linguística de Corpus. Mossoró: Queima-Bucha, 2017. p. 201-220.

SANTIAGO, Márcio Sales. **Redes de palavras-chave para artigos de divulgação científica da Medicina**: uma proposta à luz da Terminologia. 2007. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2007.

SCARTON, Carolina Evaristo; ALMEIDA, Daniel Machado; ALUÍSIO, Sandra Maria. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *In*: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 7., 2009, São Carlos. **Proceedings** […], v. 1, p. 1-10, 2009.

SILVA, Henrique César. O que é divulgação científica? **Ciência & Ensino**, v. 1, n. 1, p. 53-59, dez. 2006

SINCLAIR, John. **Corpus, concordance, collocation**. Oxford: Oxford University Press, 1991.

STEWART, Dominic. Conventionality, Creativity and Translated Text: the implications of electronic corpora in translation. *In*: OLOHAN, Maeve. (ed.) **Intercultural faultlines**: research models in translation studies in textual and cognitive aspects. Manchester: St. Jerome, 2000. p. 73-91.

STUBBS, Michael. Collocations and semantic profiles: on the cause of the trouble with quantitative studies. **Functions of Language**, v. 2, n.1, p. 1-24, 1995.

TAGNIN, Stella Esther Ortweiler. **O jeito que a gente diz**: combinações consagradas em inglês e português. Barueri: Disal, 2013.

TEXTECC. **Ferramenta MedSimples**. Universidade Federal do Rio Grande do Sul, 2020. Disponível em: http://www.ufrgs.br/textecc/acessibilidade/page/cartilha/. Acesso em: 9 jul. 2020.

TOURY, Gideon. **Descriptive Translation and Beyond**. Amsterdam: John Benjamins Publishing Company, 1995.

U.S. CENSUS BUREAU. **Current Population Survey**. Annual Social and Economic Supplement, 2017.

U.S. NATIONAL LIBRARY OF MEDICINE. **MedlinePlus**. Bethesda: U.S. Department of Health and Human Services, 2020.

APÊNDICES

APÊNDICE A – ARQUIVOS E TEXTOS DO CORPUS DO MEDLINEPLUS (EN-PT)

Nome do arquivo	Título do texto	Nome do arquivo	Título do texto
MLP_en1	Home Care after Total Joint Replacement	MLP_pt1	Cuidados em casa após uma substituição total de articulação
MLP_en2	Home Care Instructions after Surgery	MLP_pt2	Instruções sobre os cuidados em casa após a cirurgia
MLP_en3	Your Hospital Care after Surgery	MLP_pt3	Cuidados hospitalares após cirurgias
MLP_en4	Substance Abuse or Dependence	MLP_pt4	Abuso ou dependência de substâncias
MLP_en5	Angina	MLP_pt5	Angina
MLP_en6	Appendectomy	MLP_pt6	Apendicectomia
MLP_en7	Asthma	MLP_pt7	Asma
MLP_en8	Pneumococcal Polysaccharide Vaccine: What You Need to Know	MLP_pt8	Vacina pneumocócica polissacarídica: O que você precisa saber
MLP_en9	Starting an Exercise Program	MLP_pt9	Início de um programa de exercícios
MLP_en10	Prenatal Care	MLP_pt10	Cuidados no pré-natal
MLP_en11	Ultrasound in Pregnancy	MLP_pt11	Ultrassom na gravidez
MLP_en12	Fasting Blood Sugar Test	MLP_pt12	Exame de açúcar no sangue (glicemia) em jejum
MLP_en13	Receiving Blood Transfusions	MLP_pt13	Transfusões de sangue
MLP_en14	American Cancer Society Guidelines for the Early Detection of Cancer	MLP_pt14	Diretrizes da American Cancer Society para a Detecção Precoce de Câncer
MLP_en15	WHEN YOUR FURNACE KICKS ON, BE SURE POISON GAS ISN'T COMING OUT	MLP_pt15	Quando a sua caldeira estiver a arrancar, assegure-se de que não está a libertar gás tóxico
MLP_en16	WHEN THE POWER GOES OUT, KEEP YOUR GENERATOR OUTSIDE	MLP_pt16	Se falhar a energia eléctrica, mantenha o Seu gerador no exterior
MLP_en17	Prevention Guidelines: You Can Prevent Carbon Monoxide Exposure	MLP_pt17	Medidas de prevenção: Você pode evitar a exposição ao monóxido de carbono
MLP_en18	Cataract	MLP_pt18	Catarata
MLP_en19	Your Recovery After Cesarean Birth	MLP_pt19	Sua recuperação após uma cesariana
MLP_en20	MMRV Vaccine (Measles, Mumps, Rubella, and Varicella): What You Need to Know	MLP_pt20	Vacina MMRV (sarampo, caxumba, rubéola e catapora): o que você precisa saber
MLP_en21	Varicella (Chickenpox) Vaccine: What You Need to Know	MLP_pt21	Vacina contra catapora (varicela): o que você precisa saber

MLP_en22	Parents can help prevent suicide by watching for warning signs	MLP_pt22	Pais de adolescentes podem evitar o suicídio de seu filho(a) estando atentos alguns sinais de alerta.
MLP_en23	Epidural Pain Relief for Labor and Delivery	MLP_pt23	Anestesia peridural para alívio da dor no trabalho de parto e no parto
MLP_en24	Signs of Labor	MLP_pt24	Sinais de trabalho de parto
MLP_en25	Your Recovery After Vaginal Birth	MLP_pt25	Recuperação após parto normal
MLP_en26	Help your child feel safe	MLP_pt26	Ajude seu filho a sentir-se seguro
MLP_en27	Cholesterol	MLP_pt27	Colesterol
MLP_en28	Kidney Failure	MLP_pt28	Falência dos rins
MLP_en29	Cancer of the Colon and Rectum	MLP_pt29	Câncer do cólon e do reto
MLP_en30	Chronic Obstructive Pulmonary Disease (COPD)	MLP_pt30	Doença Pulmonar Obstrutiva Crônica (DPOC)
MLP_en31	Coronary Artery Disease (CAD)	MLP_pt31	Doença arterial coronariana (DAC)
MLP_en32	Feeling Sad	MLP_pt32	Sensação de tristeza
MLP_en33	About diabetes	MLP_pt33	Sobre a diabetes
MLP_en34	Glaucoma	MLP_pt34	Glaucoma
MLP_en35	Retinal Tears and Detachment	MLP_pt35	Ruptura e descolamento de retina
MLP_en36	Fetal Movement Count	MLP_pt36	Contagem de movimentos fetais
MLP_en37	Taking a Temperature	MLP_pt37	Como medir a temperatura
MLP_en38	Influenza (Flu) Vaccine (Inactivated or Recombinant): What you need to know	MLP_pt38	Vacina contra a Gripe (Influenza) (Inativada ou Recombinante): Tudo o que você precisa saber
MLP_en39	Influenza (Flu) Vaccine (Live, Intranasal): What You Need to Know	MLP_pt39	Vacina contra a Gripe (Influenza) (Viva, Intranasal): Tudo o que você precisa saber
MLP_en40	Pelvic Fracture	MLP_pt40	Fratura pélvica
MLP_en41	Gall Bladder Removal Surgery	MLP_pt41	Cirurgia para remoção da vesícula biliar
MLP_en42	Hib Vaccine: What You Need to Know	MLP_pt42	Vacina Hib: O que você precisa saber
MLP_en43	Common Sleep Problems	MLP_pt43	Problemas de sono comuns
MLP_en44	Hearing Test for Your Baby	MLP_pt44	Teste de audição para o seu bebê
MLP_en45	Heart Attack	MLP_pt45	Ataque cardíaco
MLP_en46	Heart Failure	MLP_pt46	Insuficiência Cardíaca
MLP_en47	Hysterectomy	MLP_pt47	Histerectomia
MLP_en48	Coping with Your Baby's Crying	MLP_pt48	Como lidar com o choro do bebê
MLP_en49	How to Bathe Your Newborn Baby	MLP_pt49	Como dar banho em seu recém nascido
MLP_en50	Bottle Feeding Your Baby	MLP_pt50	Alimentação do bebê com mamadeira
MLP_en51	Lung Cancer	MLP_pt51	Câncer de pulmão

MLP_en52	Mammogram	MLP_pt52	Mamografia
MLP_en53	Testicular Self Exam	MLP_pt53	Auto-exame dos testículos
MLP_en54	Choose MyPlate	MLP_pt54	Escolha MeuPrato
MLP_en55	About Your Pain	MLP_pt55	Informações sobre a dor
MLP_en56	Emotional Changes After Giving Birth	MLP_pt56	Alterações emocionais pós-parto
MLP_en57	A Healthy Pregnancy	MLP_pt57	Gravidez saudável
MLP_en58	Non-stress Test in Pregnancy	MLP_pt58	Teste sem estresse na gravidez
MLP_en59	How to Quit Smoking	MLP_pt59	Como parar de fumar
MLP_en60	STDs (Sexually Transmitted Diseases)	MLP_pt60	DSTs (Doenças Sexualmente Transmissíveis)
MLP_en61	CPAP (Continuous Positive Airway Pressure)	MLP_pt61	Pressão positiva contínua nas vias aéreas (PPCA)
MLP_en62	Coping with Stress	MLP_pt62	Como lidar com o estresse
MLP_en63	Stroke	MLP_pt63	Derrame
MLP_en64	Preparing for Your Surgery	MLP_pt64	Preparação para a cirurgia
MLP_en65	Tuberculosis (TB)	MLP_pt65	Tuberculose (TB)
MLP_en66	Vaginal Infection	MLP_pt66	Infecção vaginal

APÊNDICE B – ARQUIVOS E TEXTOS DO SUBCORPUS DO MINISTÉRIO DA SAÚDE

Nome do arquivo	Título do texto
BVS_dicsau1	Acidente vascular cerebral (AVC)
BVS_dicsau2	Acidentes com raios
BVS_dicsau3	Acidentes por afogamento
BVS_dicsau4	Acidentes por mergulho
BVS_dicsau5	Acne
BVS_dicsau6	Acolhimento
BVS_dicsau7	Adolescência saudável
BVS_dicsau8	Aftas
BVS_dicsau9	Albinismo
BVS_dicsau10	Alcoolismo
BVS_dicsau11	Alimentação da pessoa idosa
BVS_dicsau12	Alimentação de crianças
BVS_dicsau13	Alimentação saudável
BVS_dicsau14	Alimentos funcionais
BVS_dicsau15	Alongamento
BVS_dicsau16	Amamentação
BVS_dicsau17	Anabolizantes
BVS_dicsau18	Anemia
BVS_dicsau19	Anemia falciforme
BVS_dicsau20	Anestesia tem risco?
BVS_dicsau21	Aneurisma
BVS_dicsau22	Ansiedade
BVS_dicsau23	Aparelhos para audição
BVS_dicsau24	Artrite reumatoide e artrose (osteoartrite)
BVS_dicsau25	Artroplastia do joelho
BVS_dicsau26	Asma
BVS_dicsau27	Assédio moral
BVS_dicsau28	Ataque cardíaco (infarto)
BVS_dicsau29	Atividade física faz bem à saúde
BVS_dicsau30	Automedicação
BVS_dicsau31	Bicho de pé
BVS_dicsau32	Camisinha feminina
BVS_dicsau33	Camisinha masculina
BVS_dicsau34	Câncer de boca
BVS_dicsau35	Câncer de intestino
BVS_dicsau36	Câncer de mama
BVS_dicsau37	Câncer de pele
BVS_dicsau38	Câncer de próstata
BVS_dicsau39	Câncer do colo do útero

BVS_dicsau40	Câncer ocupacional
BVS_dicsau41	Câncer: dicas para se prevenir
BVS_dicsau42	Casas de apoio para pessoas vivendo com HIV/Aids
BVS_dicsau43	Catarata
BVS_dicsau44	Células-tronco
BVS_dicsau45	Centro de Referência em Saúde do Trabalhador - Cerest
BVS_dicsau46	Ceratocone
BVS_dicsau47	Choque anafilático
BVS_dicsau48	Cirurgia bariátrica
BVS_dicsau49	Cistite
BVS_dicsau50	Climatério
BVS_dicsau51	Cólicas menstruais
BVS_dicsau52	Como parar de fumar
BVS_dicsau53	Condiloma acuminado (HPV)
BVS_dicsau54	Conjuntivite
BVS_dicsau55	Constipação intestinal
BVS_dicsau56	Convulsão
BVS_dicsau57	Corrija sua postura
BVS_dicsau58	Cuidadores de idosos
BVS_dicsau59	Cuidados com a voz
BVS_dicsau60	Cuidados com o lixo
BVS_dicsau61	Cuidados com os saneantes (desinfetantes, detergentes, etc.)
BVS_dicsau62	Cuidados quando estamos gessados
BVS_dicsau63	Deficiência de iodo
BVS_dicsau64	Deficiência de vitamina A
BVS_dicsau65	Dengue
BVS_dicsau66	Dente do siso (dente serotino ou terceiro molar)
BVS_dicsau67	Depressão
BVS_dicsau68	Diabetes
BVS_dicsau69	Diarréia e desidratação
BVS_dicsau70	Disfunção da articulação temporomandibular (ATM)
BVS_dicsau71	Dislexia
BVS_dicsau72	Distrofia muscular
BVS_dicsau73	Distúrbios do sono
BVS_dicsau74	Doação de sangue
BVS_dicsau75	Doença celíaca
BVS_dicsau76	Doença de Alzheimer
BVS_dicsau77	Doença de Chagas
BVS_dicsau78	Doença de Crohn
BVS_dicsau79	Doença de Parkinson
BVS_dicsau80	Doença mão-pé-boca
BVS_dicsau81	Doenças sexualmente transmissíveis (DST)
BVS_dicsau82	Doenças transmitidas por alimentos e água
BVS_dicsau83	Endometriose

BVS_dicsau84	Engasgo
BVS_dicsau85	Envenenamento
BVS_dicsau86	Enxaqueca
BVS_dicsau87	Epilepsia
BVS_dicsau88	Erisipela
BVS_dicsau89	Esteatose hepática
BVS_dicsau90	Estrabismo (vesgueira, vesguice)
BVS_dicsau91	Estresse
BVS_dicsau92	Febre amarela
BVS_dicsau93	Febre Chikungunya
BVS_dicsau94	Febre maculosa brasileira
BVS_dicsau95	Febre reumática
BVS_dicsau96	Fibrose cística
BVS_dicsau97	Fissura lábio-palatal e lábio leporino
BVS_dicsau98	Glaucoma
BVS_dicsau99	Gravidez na adolescência
BVS_dicsau100	Gripes e resfriados
BVS_dicsau101	Halitose
BVS_dicsau102	Hanseníase
BVS_dicsau103	Hemofilia
BVS_dicsau104	Hemorroidas
BVS_dicsau105	Hepatite
BVS_dicsau106	Hérnia de disco
BVS_dicsau107	Herpes simples
BVS_dicsau108	Higiene para uma vida saudável
BVS_dicsau109	Higienização das mãos na assistência à saúde
BVS_dicsau110	Hipertensão arterial
BVS_dicsau111	Hipertireoidismo
BVS_dicsau112	Hipotireoidismo
BVS_dicsau113	HIV e aids
BVS_dicsau114	Implantes dentários
BVS_dicsau115	Importância do pré-natal
BVS_dicsau116	Incontinência urinária
BVS_dicsau117	Infecção pelo vírus Zika
BVS_dicsau118	Infertilidade feminina
BVS_dicsau119	Infertilidade masculina
BVS_dicsau120	Insolação
BVS_dicsau121	Insuficiência renal aguda
BVS_dicsau122	Insuficiência renal crônica
BVS_dicsau123	Intolerância à lactose
BVS_dicsau124	Intoxicação por agrotóxicos
BVS_dicsau125	Leishmaniose
BVS_dicsau126	Leptospirose
BVS_dicsau127	Lesões por esforços repetitivos (LER)

BVS_dicsau128	Limpeza de caixa d'água
BVS_dicsau129	Lombalgia (dor nas costas)
BVS_dicsau130	Lúpus
BVS_dicsau131	Malária
BVS_dicsau132	Medicamentos genéricos
BVS_dicsau133	Meningite
BVS_dicsau134	Micoses
BVS_dicsau135	Microcefalia
BVS_dicsau136	Miomas uterinos
BVS_dicsau137	Morte encefálica
BVS_dicsau138	Obesidade
BVS_dicsau139	Osteoporose
BVS_dicsau140	Papanicolau (exame preventivo de colo de útero)
BVS_dicsau141	Paralisia facial
BVS_dicsau142	Pé diabético
BVS_dicsau143	Pediculose da cabeça (piolhos)
BVS_dicsau144	Picadas de insetos e animais peçonhentos - parte 1
BVS_dicsau145	Picadas de insetos e animais peçonhentos - parte 2
BVS_dicsau146	Pneumonia
BVS_dicsau147	Poliomielite (paralisia infantil)
BVS_dicsau148	Pombos: riscos para a saúde humana
BVS_dicsau149	Psoríase
BVS_dicsau150	Qualidade de vida em cinco passos
BVS_dicsau151	Quedas de idosos
BVS_dicsau152	Queimaduras
BVS_dicsau153	Raiva
BVS_dicsau154	Refluxo gastroesofágico
BVS_dicsau155	Ronco
BVS_dicsau156	Rotaviroses
BVS_dicsau157	Rótulos de alimentos: orientações ao consumidor
BVS_dicsau158	Salão de beleza: utilize sem prejudicar sua saúde
BVS_dicsau159	Sarampo
BVS_dicsau160	Saúde bucal
BVS_dicsau161	Saúde da coluna
BVS_dicsau162	Saúde do coração
BVS_dicsau163	Saúde dos pés - sapatos
BVS_dicsau164	Saúde e segurança no trabalho
BVS_dicsau165	Saúde ocular
BVS_dicsau166	Sífilis
BVS_dicsau167	Síndrome metabólica
BVS_dicsau168	Síndrome vasovagal
BVS_dicsau169	Sinusite
BVS_dicsau170	Soluço
BVS_dicsau171	Surdez

BVS_dicsau172	Tabagismo
BVS_dicsau173	Teste da orelhinha
BVS_dicsau174	Teste do pezinho
BVS_dicsau175	Torção de tornozelo
BVS_dicsau176	Toxoplasmose
BVS_dicsau177	Tracoma
BVS_dicsau178	Trânsito saudável
BVS_dicsau179	Transplante de medula óssea
BVS_dicsau180	Transplante de órgãos e tecidos
BVS_dicsau181	Transtorno do déficit de atenção com hiperatividade - TDAH
BVS_dicsau182	Tuberculose
BVS_dicsau183	Uso de antibióticos - orientações
BVS_dicsau184	Uso de medicamentos - orientações
BVS_dicsau185	Vacinação
BVS_dicsau186	Varizes
BVS_dicsau187	Vasectomia
BVS_dicsau188	Verrugas
BVS_dicsau189	Violência contra crianças e adolescentes - parte 1
BVS_dicsau190	Violência contra crianças e adolescentes - parte 2
BVS_dicsau191	Vitiligo

APÊNDICE C – LISTA DE PALAVRAS-CHAVE DO SUBCORPUS DO MEDLINEPLUS (EN)

Frequência	Keyness	Effect	Keyword	Palavras lematizadas
1141	6043.13	3903.8292	your	
900	4464.93	291.1527	you	
247	1218.18	274.3955	doctor	
189 [+11]	982.93	1257.6896	baby	babies
344	933.76	13.5962	if	
236	809.35	25.3576	will	
129 [+15; +36]	559.79	74.5128	get	gets; getting
140 [+25]	550.84	43.2661	vaccine	vaccines
187 [+16; +37]	511.66	13.8203	do	doing; done
129	437.69	24.4803	help	
95 [+64]	431.48	105.0741	call	called
80	393.52	265.3401	feel	
70 [+8]	361.58	928.4286	talk	talking
117	354.53	17.6547	signs	
98	353.61	30.2496	take	
124	347.42	14.5735	pain	
77	344.44	92.8596	away	
61 [+16]	264.15	73.5302	check	checked
79	263.87	23.2882	water	
54 [+19]	244.45	102.2681	incision	incisions
50 [+13]	237.71	165.6942	sleep	sleeping
45	222.83	298.2079	tell	
43 [+20]	212.38	284.9378	eat	eating
45	211.95	149.1033	flu	
52	204.40	43.0819	go	
46	204.04	87.0972	ask	
48	193.71	48.9397	keep	
36	191.25	954.0223	yourself	
71	187.07	12.7239	day	
46	173.39	35.8628	nurse	
33	166.44	437.2215	clean	
67	161.83	10.7037	problems	
59	152.48	12.2215	weeks	
46	152.12	22.5798	reaction	
33 [+15]	150.55	109.3047	eye	eyes
37	149.37	49.0265	put	
28	148.75	741.8464	vaers	
39 [+16]	148.03	36.9136	breathing	breathe
52	147.48	14.9840	prevent	

29	145.44	384.1807	allergic	
29	145.44	384.1807	degrees	
27	143.43	715.3313	hib	
36	141.71	43.3636	swelling	
41	141.50	25.8722	give	
32	141.42	84.7913	abdomen	
49	139.44	15.1034	until	
47 [+12]	136.90	15.9720	hours	hour
33 [+9]	136.00	54.6519	drink	drinks
29	134.58	128.0597	foods	
25	132.81	662.3057	bath	
25	132.81	662.3057	vaccinated	
32	131.13	52.9942	vision	
46	130.04	14.8692	make	
28 [+22]	129.47	123.6403	feeling	feelings
27	129.17	178.8321	wash	
39 [+14]	128.77	22.4687	exercise	exercises
24	127.49	635.7952	varicella	
44	127.46	15.7595	back	
51	125.47	11.0814	home	
30	124.50	56.7764	formula	
49	119.07	10.8237	person	
28	117.93	61.8197	smoke	
38	117.57	18.6486	minutes	
21	111.56	556.2727	shower	
21	111.56	556.2727	thermometer	
25	110.12	82.7875	try	
29	108.61	34.9248	temperature	
32	105.41	22.3128	stress	
44	104.73	10.4121	right	
21 [+20]	103.57	278.1358	legs	leg
21	103.57	278.1358	soap	
22	103.39	145.6940	tube	
23	100.11	76.1601	trouble	
22	98.96	97.1291	learn	
27 [+25]	94.88	27.5120	removed	remove
25	94.82	36.7940	stop	
37	94.61	11.9569	serious	
17	90.31	450.2642	mmrv	
17	90.31	450.2642	warm	
21	90.15	69.5335	lungs	
20	88.87	88.2941	want	
35	88.12	11.5927	vaginal	
18	87.93	238.3815	tired	

22	85.83	41.6264	things	
16	84.99	423.7659	gently	
16	84.99	423.7659	wear	
29	83.14	15.3664	down	
17	82.73	225.1316	glaucoma	
17	82.73	225.1316	shot	
20	81.95	52.9763	night	
26	81.21	19.1331	influenza	
26	81.21	19.1331	limit	
22	80.83	32.3760	hard	
19 [+21]	80.26	62.9076	walk	walking
15	79.68	397.2691	dry	
31	78.53	11.7334	avoid	
18	75.34	59.5950	uterus	
14	74.37	370.7738	chickenpox	
14	74.37	370.7738	dizzy	
25	73.89	16.5569	too	
17	73.83	75.0435	bottle	
17	73.83	75.0435	sure	
16	72.73	105.9410	gas	
16	72.73	105.9410	slowly	
18	72.26	47.6759	happen	
19	71.66	35.9470	quit	
30	70.57	10.1899	occur	
17 [+18]	70.44	56.2825	fluids	fluid
29	69.95	10.6709	healthy	
13	69.06	344.2800	allergies	
13	69.06	344.2800	clots	
13	69.06	344.2800	feet	
13	69.06	344.2800	sad	
16	68.86	70.6272	move	
16	68.86	70.6272	someone	
16	68.86	70.6272	throat	
15	67.66	99.3168	burning	
22	67.47	18.2112	vaccination	
17	67.45	45.0259	hot	
14 [+9]	67.17	185.3865	crying	cry
23	64.33	14.5062	flow	
15	63.90	66.2111	bladder	
15	63.90	66.2111	compensation	
12	63.74	317.7878	redness	
18	62.56	26.4864	leave	
18 [+13]	62.56	26.4864	movement	movements
18	62.56	26.4864	rash	

24	62.04	12.2261	diet	
13	62.00	172.1396	milk	
18	60.58	23.8377	fat	
22	60.34	13.8750	immune	
17	60.13	28.1410	rest	
16	60.08	35.3134	drive	
25	59.31	10.3478	arm	
14 [+9]	58.96	61.7952	sit	sitting
11 [+15]	58.43	291.2971	friend	friends
15	57.92	39.7265	let	
20	57.67	15.5808	vagina	
13	57.57	86.0696	mouth	
13	57.57	86.0696	sugar	
12	56.84	158.8936	fruits	
12 [+16]	56.84	158.8936	muscles	muscle
21	56.39	13.2440	start	
14	55.88	46.3463	car	
14	55.88	46.3463	vegetables	
23	55.01	10.5043	pregnant	
13	54.05	57.3796	hands	
10	53.12	264.8079	bandage	
10	53.12	264.8079	detachment	
10	53.12	264.8079	faint	
10	53.12	264.8079	gall	
10	53.12	264.8079	germs	
10	53.12	264.8079	lose	
10	53.12	264.8079	staples	
10	53.12	264.8079	stitches	
10	53.12	264.8079	vicp	
12	52.55	79.4466	recover	
20	52.49	12.6130	clear	
11	51.69	145.6482	pounds	
11	51.69	145.6482	prepare	
11	51.69	145.6482	tub	
17	51.14	17.3174	cold	
14	50.81	30.8974	air	
12	49.17	52.9643	morning	
13	48.51	34.4276	cord	
9	47.81	238.3203	apnea	
9	47.81	238.3203	asleep	
9	47.81	238.3203	mucus	
9	47.81	238.3203	odor	
9	47.81	238.3203	steri-strips	
9	47.81	238.3203	tips	

11	47.56	72.8240	pneumococcal	
14	46.69	23.1730	exam	
10	46.56	132.4037	heartbeat	
10	46.56	132.4037	wake	
12	46.33	39.7231	nose	
18	45.96	11.9186	regular	
14	44.89	20.5981	choose	
11	44.32	48.5492	awake	
11	44.32	48.5492	weakened	
12	43.87	31.7784	dressing	
16	42.96	13.2422	sometimes	
10	42.60	66.2017	contractions	
8	42.49	211.8342	aches	
8	42.49	211.8342	cpap	
8	42.49	211.8342	exercising	
8	42.49	211.8342	lift	
8	42.49	211.8342	liquids	
8	42.49	211.8342	pat	
8	42.49	211.8342	perineal	
8	42.49	211.8342	ppsv	
8	42.49	211.8342	relax	
8	42.49	211.8342	season	
8	42.49	211.8342	washcloth	
8	42.49	211.8342	wheezing	
13	42.31	21.5171	ray	
9	41.45	119.1599	hurt	
9	41.45	119.1599	prenatal	
9	41.45	119.1599	retina	
9	41.45	119.1599	retinal	
14	40.23	15.4485	look	
12	39.75	22.6988	anything	
10	39.51	44.1344	protect	
10	39.51	44.1344	wait	
9	37.67	59.5798	hearing	
9	37.67	59.5798	kidneys	
9	37.67	59.5798	write	
14	37.59	13.2415	bacteria	
7	37.18	185.3496	angry	
7	37.18	185.3496	blanket	
7	37.18	185.3496	color	
7	37.18	185.3496	comfortable	
7	37.18	185.3496	disaster	
7	37.18	185.3496	ears	
7	37.18	185.3496	grains	

7	37.18	185.3496	hives	
7	37.18	185.3496	okay	
7	37.18	185.3496	pads	
7	37.18	185.3496	pillow	
7	37.18	185.3496	plenty	
7	37.18	185.3496	rubella	
7	37.18	185.3496	shoulder	
7	37.18	185.3496	stockings	
7	37.18	185.3496	tapes	
10	36.95	33.1007	instructions	
14	36.38	12.3587	ordered	
12	36.37	17.6545	parts	
8	36.36	105.9169	bones	
8	36.36	105.9169	coughing	
8	36.36	105.9169	sanitary	
8	36.36	105.9169	teach	
13	36.19	14.3446	oxygen	
13	36.19	14.3446	rectum	
11	35.39	20.8066	inside	
14	35.24	11.5863	stomach	
10	34.75	26.4805	appetite	
10	34.75	26.4805	quickly	
10	34.75	26.4805	stairs	
9	34.75	39.7198	watch	
14	34.16	10.9047	sudden	
11	33.73	18.2058	anyone	
12	33.50	14.4446	manage	
12	33.50	14.4446	told	
12	33.50	14.4446	ultrasound	
10	32.82	22.0670	claim	
10	32.82	22.0670	suicide	
8	32.78	52.9583	appointment	
8	32.78	52.9583	decide	
8	32.78	52.9583	itchy	
8	32.78	52.9583	meals	
8	32.78	52.9583	sore	
8	32.78	52.9583	vitamins	

APÊNDICE D – LISTA DE PALAVRAS-CHAVE DO SUBCORPUS DO MEDLINEPLUS (PT)

Frequência	Keyness	Effect	Keyword	Palavras lematizadas
521	2304.78	122.9480	você	
393	960.29	11.5460	seu	
175 [+44]	844.59	501.4886	vacina	vacinas
169 [+16]	814.60	484.2208	bebê	bebês
251	664.07	13.6557	médico	
119	599.77	2724.2407	tiver	
94	418.97	134.4073	estiver	
154 [+15]	414.91	14.2244	dor	dores
134	400.59	18.4834	cirurgia	
86 [+35]	354.74	70.2524	tomar	tome
70 [+16; +14]	328.20	266.7485	sentir	sentindo
64	322.48	1463.0933	ligue	
93	312.72	26.5929	poderá	
60	302.32	1371.5108	gripe	
100 [+57; +12]	297.03	18.1576	fazer	faça; fará
116 [+66]	296.41	12.6416	pessoas	pessoa
92	256.84	15.4733	depois	
80 [+43]	245.46	19.8850	medicamentos	medicamento
52 [+20]	239.22	198.0656	incisão	incisões
44	221.69	1005.3665	catapora	
83	196.10	10.7841	sinais	
56	188.45	26.6641	reação	
37	186.41	845.2718	inchaço	
40 [+8]	180.27	152.3118	fumar	fume
35	176.33	799.5408	vaginal	
55 [+24]	175.70	22.4461	use	usar
35	167.36	399.7695	alérgica	
62	166.81	14.1714	semanas	
48 [+13; +12; +36; +11; +12]	156.27	23.8439	ajuda	ajudá[-lo/-la]; ajudam; ajudar; ajudará; ajude
36	156.04	102.7998	converse	
44	152.35	29.5685	leite	
36	151.81	82.2397	enfermeira	
50	149.33	18.4283	exercícios	
31	147.45	354.0457	evite	
42	139.12	25.2521	febre	
44 [+14]	137.42	20.9440	causar	causadas
27	136.02	616.6636	hib	
37 [+17]	133.87	35.2187	ficar	fique
31	131.97	88.5108	mantenha	

26 [+7]	130.99	593.8091	respirar	respire
26	130.99	593.8091	termômetro	
47	126.30	14.1303	casa	
25	125.95	570.9558	vaers	
27	117.02	102.7766	peito	
23	115.87	525.2528	útero	
36	113.66	21.6413	imediatamente	
25	112.02	142.7383	dormir	
22 [+13]	110.83	502.4029	parar	pare
23 [+17]	107.72	262.6258	coloque	colocado
21	105.79	479.5543	procure	
21	105.79	479.5543	receberá	
21	105.79	479.5543	vagina	
22	102.77	251.2009	voltar	
24 [+15]	102.58	91.3500	pernas	perna
29	100.89	30.1073	vacinação	
20	100.76	456.7068	bebidas	
20 [+7]	100.76	456.7068	lave	lavar
20	100.76	456.7068	pergunte	
31	98.58	22.1271	coração	
21	97.83	239.7766	alguém	
21	97.83	239.7766	glaucoma	
21 [+34; +9]	97.83	239.7766	precisar	precisa; precisará
35	97.47	15.3749	sono	
29	96.31	25.4753	temperatura	
19	95.72	433.8605	fale	
19	95.72	433.8605	verifique	
34	95.08	15.5327	normalmente	
23	93.89	65.6560	instruções	
31	92.96	18.6332	parto	
20	92.88	228.3529	tontura	
18	90.68	411.0153	erupção	
18	90.68	411.0153	mmrv	
32	85.80	14.0560	lhe	
17	85.64	388.1713	banheira	
17	85.64	388.1713	limpe	
21 [+10]	84.48	59.9438	peça	pedir
16	80.60	365.3284	dsts	
26	79.65	19.7930	pulmões	
28	79.22	15.9873	cdc	
24	79.18	24.9132	incluir	
29	78.57	14.3987	cabeça	
33	76.78	10.4688	saber	
15	75.56	342.4867	chuveiro	

15	75.56	342.4867	coceira	
15	75.56	342.4867	tossir	
18	73.96	68.5021	retina	
17 [+17]	73.19	97.0424	coisas	coisa
16 [+14]	73.16	182.6638	comer	coma
16	73.16	182.6638	costas	
16	73.16	182.6638	garganta	
23	72.88	21.8849	site	
19	71.99	43.3856	cérebro	
19	71.99	43.3856	tubo	
14 [+12]	70.53	319.6461	beba	beber
14	70.53	319.6461	vermelhidão	
22	68.70	20.9329	informe	
30	68.38	10.0762	visão	
18	67.43	41.1011	curativo	
18	67.43	41.1011	injeção	
19	66.60	30.9896	noite	
13	65.49	296.8068	cansaço	
13	65.49	296.8068	deixe	
13	65.49	296.8068	mamadeira	
13	65.49	296.8068	odor	
13	65.49	296.8068	sabonete	
24	65.07	14.4232	sensação	
25	64.68	12.9756	cardíaco	
15	63.59	85.6213	achar	
15	63.59	85.6213	açúcar	
12	60.45	273.9685	consulte	
12	60.45	273.9685	devo	
12	60.45	273.9685	grávida	
12	60.45	273.9685	sabão	
20	60.43	19.0289	tosse	
15	59.85	57.0807	dê	
14	58.81	79.9112	compensação	
13	58.45	148.4031	limpa	
16	58.39	36.5324	mãos	
16 [+12]	58.39	36.5324	sanguíneo	sanguíneos
20	57.04	16.3105	sangramento	
18	55.86	20.5503	danos	
11	55.41	251.1315	alcoólicas	
11	55.41	251.1315	cama	
11	55.41	251.1315	descanse	
11	55.41	251.1315	steri-strips	
11	55.41	251.1315	toalha	
14 [+11]	55.19	53.2740	começar	comece

13	54.04	74.2014	cancerígenas	
13	54.04	74.2014	coágulos	
13	54.04	74.2014	sarampo	
13	54.04	74.2014	urinar	
20	53.98	14.2716	passar	
15	53.91	34.2483	amigos	
15 [+13]	53.91	34.2483	olhos	olho
22	53.53	11.4177	prevenir	
14	52.13	39.9554	gás	
10	50.38	228.2955	adesivas	
10	50.38	228.2955	carro	
10	50.38	228.2955	desaparecem	
10	50.38	228.2955	descolamento	
10	50.38	228.2955	dicas	
10 [+8]	50.38	228.2955	dirija	dirigir
10	50.38	228.2955	histerectomia	
10	50.38	228.2955	pescoço	
10 [+10]	50.38	228.2955	remédios	remédio
10	50.38	228.2955	tristeza	
10	50.38	228.2955	vicp	
10	50.38	228.2955	visite	
12	49.30	68.4918	quente	
12	49.30	68.4918	tente	
11	48.70	125.5654	reto	
13	47.61	37.1005	esperar	
17	46.77	14.9292	leves	
9	45.34	205.4608	choro	
9	45.34	205.4608	escadas	
9	45.34	205.4608	frio	
9	45.34	205.4608	grampos	
9	45.34	205.4608	levante	
9	45.34	205.4608	seque	
13	45.06	29.6803	ataque	
18 [+15]	45.04	12.0882	abdômen	abdome
14	44.97	22.8315	derrame	
14	44.97	22.8315	sexuais	
11	44.59	62.7826	sozinho	
10	43.84	114.1475	intravenoso	
10	43.84	114.1475	ombro	
10	43.84	114.1475	siga	
10	43.84	114.1475	tomado	
15	43.59	17.1240	quarto	
12	43.13	34.2458	contrações	
11	41.37	41.8550	acontecer	

14	41.25	17.7578	fraqueza	
13	40.77	21.2002	andar	
12	40.69	27.3965	alergia	
12	40.69	27.3965	preparar	
12	40.69	27.3965	sofrido	
15	40.48	14.2699	respiração	
8 [+9]	40.30	182.6272	acordar	acordado
8	40.30	182.6272	banheiro	
8 [+9]	40.30	182.6272	consuma	consumir
8	40.30	182.6272	debaixo	
8	40.30	182.6272	ervas	
8	40.30	182.6272	laiv	
8	40.30	182.6272	libras	
8	40.30	182.6272	monitorada	
8	40.30	182.6272	perineal	
8	40.30	182.6272	ppsv	
8	40.30	182.6272	ventre	
16	40.18	12.1772	sentimentos	
10	39.90	57.0736	angina	
10	39.90	57.0736	enfraquecido	
10	39.90	57.0736	pediatra	
10	39.90	57.0736	provedor	
9	39.00	102.7302	aliviar	
9	39.00	102.7302	rubéola	
9	39.00	102.7302	testículos	
10	36.83	38.0490	sair	
12	36.62	19.5689	pneumonia	
15	36.50	11.4158	houver	
11	36.38	25.1129	músculos	
7	35.26	159.7947	absorventes	
7	35.26	159.7947	batimentos	
7	35.26	159.7947	deitar	
7	35.26	159.7947	dura	
7	35.26	159.7947	fralda	
7	35.26	159.7947	meias	
7	35.26	159.7947	morna	
7	35.26	159.7947	ovários	
7	35.26	159.7947	rosto	
7	35.26	159.7947	subir	
7	35.26	159.7947	tiras	
7	35.26	159.7947	travesseiro	
7	35.26	159.7947	troque	
7	35.26	159.7947	vacinada	
9	35.24	51.3650	calafrios	

9	35.24	51.3650	muco	
9	35.24	51.3650	parceiro	
12	34.87	17.1227	quaisquer	
11	34.35	20.9273	prescritos	
10	34.30	28.5367	ir	
10	34.30	28.5367	removido	
13	34.23	13.4909	boca	
8	34.19	91.3134	caxumba	
8	34.19	91.3134	desastre	
8	34.19	91.3134	prisão	
8	34.19	91.3134	queimação	
8	34.19	91.3134	ultrassom	
8	34.19	91.3134	vesícula	
12	33.28	15.2202	pés	

Palavras-chave-chave

Palavras-chave que agrupam palavras-chave-chave

Palavras-chave-chave após a lematização manual

APÊNDICE E – LISTA DE PALAVRAS-CHAVE DO SUBCORPUS DO MINISTÉRIO DA SAÚDE

Frequência	Keyness	Effect	Keyword	Palavras lematizadas
153 [+25]	340.18	14.4076	alimentos	alimento
144	321.63	14.5821	pessoa	
103	312.58	50.2350	boca	
135 [+71]	305.42	15.0933	evitar	evite
123	273.05	14.3476	medicamentos	
79	256.13	84.7423	mãos	
111	230.38	12.1533	você	
67 [+22]	213.31	71.8598	olhos	olho
79	202.43	22.2999	febre	
67	195.70	39.9218	pés	
71	191.45	27.1972	usar	
50 [+39]	185.00	536.1647	procure	procurar
48	177.60	514.7058	camisinha	
50 [+19]	162.97	89.3602	pernas	perna
51	161.40	68.3612	bebê	
43	159.10	461.0632	bebidas	
49	154.31	65.6788	vacina	
67	153.48	15.6211	cabeça	
48	146.19	51.4700	cérebro	
38	140.60	407.4270	inchaço	
36	133.20	385.9742	útero	
38	125.40	101.8563	lixo	
36 [+19]	124.35	192.9867	coloque	colocar
33 [+11]	122.10	353.7971	alcoólicas	alcoólica
45	120.92	26.8061	dentes	
39	119.05	52.2687	mantenha	
58	117.58	11.5180	mulher	
49	116.74	17.5139	pé	
35 [+16]	114.62	93.8117	picada	picadas
49	114.23	16.4193	coração	
38	111.40	40.7422	roupas	
43 [+21]	105.50	19.2105	acidentes	acidente
36	100.80	32.1641	faça	
32	98.95	57.1783	fumar	
42	97.21	16.0829	dores	
26	96.20	278.7260	cama	
39	95.37	19.0065	coluna	
38	94.89	20.3709	objetos	
33	94.21	35.3793	músculos	

40	93.39	16.4949	provocar	
25	92.50	268.0026	manchas	
37	91.66	19.8346	passo	
24	88.80	257.2794	hpv	
41	87.35	12.9292	leite	
27	86.02	72.3621	nariz	
23	85.10	246.5565	lavar	
25	84.36	134.0010	chão	
36	83.32	16.0818	veias	
22 [+18]	81.40	235.8339	joelhos	joelho
28	80.85	37.5214	articulações	
28	80.85	37.5214	casas	
24	80.74	128.6394	costas	
32	78.32	19.0592	fezes	
21	77.70	225.1115	rosto	
21	77.70	225.1115	varizes	
34 [+18]	77.10	15.1880	ficar	ficam
24	75.38	64.3196	dormir	
25	74.45	44.6668	adolescente	
33	74.01	14.7412	limpeza	
20	74.00	214.3893	coceira	
35	71.52	11.7261	sol	
29	71.40	19.4308	estiver	
19	70.30	203.6674	luvas	
21	69.90	112.5555	dengue	
21	69.90	112.5555	garganta	
21 [+10]	69.90	112.5555	sapatos	sapato
21	69.90	112.5555	verduras	
33	67.60	11.7928	refeições	
23	67.53	41.0925	fumo	
26 [+12]	67.36	23.2269	provocada	provocadas
18	66.60	192.9458	insetos	
18	66.60	192.9458	pênis	
20	66.30	107.1944	braços	
21	64.79	56.2776	acne	
22	64.08	39.3054	cima	
28	63.22	15.0083	estômago	
17	62.90	182.2244	cansaço	
17 [+20]	62.90	182.2244	deixe	deixar
17	62.90	182.2244	halitose	
17	62.90	182.2244	lados	
17	62.90	182.2244	querido	
17	62.90	182.2244	tiver	
17	62.90	182.2244	vagina	

21	60.64	37.5183	dê	
21 [+13]	60.64	37.5183	sentir	
22	60.44	29.4790	calor	
16	59.20	171.5033	banheiro	
16 [+15]	59.20	171.5033	beber	beba
16	59.20	171.5033	caminhar	
16	59.20	171.5033	pescoço	
16	59.20	171.5033	unhas	
18	59.10	96.4727	comer	
19	57.77	50.9166	açúcar	
27	55.86	12.0601	respiração	
17	55.51	91.1120	cadeira	
17	55.51	91.1120	ente	
17 [+16]	55.51	91.1120	feridas	ferida
17 [+13]	55.51	91.1120	lábio	lábios
15	55.50	160.7824	gripe	
15	55.50	160.7824	vermelhidão	
18	54.28	48.2362	próstata	
21	53.95	22.5109	preventivo	
19	53.80	33.9443	dedos	
19	53.80	33.9443	transmitida	
20	53.74	26.7984	repouso	
16	51.93	85.7515	caixa	
14	51.80	150.0618	conjuntivite	
14	51.80	150.0618	dicas	
14	51.80	150.0618	remédios	
14	51.80	150.0618	sífilis	
22	51.74	16.8450	aparecem	
21	51.14	18.7590	diarréia	
17	50.79	45.5559	cabelos	
17	50.79	45.5559	utilize	
23	50.15	13.6973	fraqueza	
23	50.15	13.6973	orientações	
21	48.60	16.0791	iodo	
13	48.10	139.3415	barriga	
13	48.10	139.3415	raiva	
13	48.10	139.3415	vermelhas	
16	47.32	42.8756	amarela	
16	47.32	42.8756	anestesia	
17	47.01	30.3705	dst	
12	44.40	128.6214	deitar	
12	44.40	128.6214	parar	
12	44.40	128.6214	sabão	
12	44.40	128.6214	saneantes	

12	44.40	128.6214	transmissor	
15	43.87	40.1954	café	
15	43.87	40.1954	ler	
15	43.87	40.1954	sinusite	
16	43.65	28.5837	leishmaniose	
16	43.65	28.5837	óculos	
18	41.66	16.0786	ambientes	
13	41.23	69.6706	genital	
13	41.23	69.6706	levantar	
17	41.02	18.2222	legumes	
11	40.70	117.9015	embaixo	
11	40.70	117.9015	espinhas	
11	40.70	117.9015	estrabismo	
11	40.70	117.9015	iluminação	
11	40.70	117.9015	infectada	
11	40.70	117.9015	lacraias	
11	40.70	117.9015	portas	
11	40.70	117.9015	solo	
11	40.70	117.9015	suor	
11	40.70	117.9015	vitiligo	
16	40.55	21.4377	contaminados	
14	40.42	37.5153	córnea	
14	40.42	37.5153	falciforme	
14	40.42	37.5153	queimaduras	
14	40.42	37.5153	secreções	
15	40.30	26.7969	insônia	
15	40.30	26.7969	mucosas	
15	40.30	26.7969	peito	
18	39.34	13.7816	sexuais	
16	37.85	17.1501	bactéria	
12	37.69	64.3105	formigamento	
12	37.69	64.3105	lembre	
12	37.69	64.3105	psoríase	
10	37.00	107.1819	aranhas	
10	37.00	107.1819	avermelhadas	
10	37.00	107.1819	couro	
10	37.00	107.1819	escadas	
10	37.00	107.1819	escuros	
13	37.00	34.8352	fissura	
10	37.00	107.1819	limpo	
13	37.00	34.8352	mesa	
13	37.00	34.8352	mosquito	
10	37.00	107.1819	relaxar	
10	37.00	107.1819	retire	

10	37.00	107.1819	veneno	
14	36.98	25.0101	feminina	
14	36.98	25.0101	gorduras	
14	36.98	25.0101	herpes	
14	36.98	25.0101	ponta	
14	36.98	25.0101	praticar	

Palavras-chave-chave

Palavras-chave que agrupam palavras-chave-chave

Palavras-chave-chave após a lematização manual

ANEXOS

ANEXO A – ÍNDICES DO COH-METRIX 3.0

	Label	Description
Descriptive		
1	DESPC	Paragraph count, number of paragraphs
2	DESSC	Sentence count, number of sentences
3	DESWC	Word count, number of words
4	DESPL	Paragraph length, number of sentences, mean
5	DESPLd	Paragraph length, number of sentences, standard deviation
6	DESSL	Sentence length, number of words, mean
7	DESSLd	Sentence length, number of words, standard deviation
8	DESWLsy	Word length, number of syllables, mean
9	DESWLsyd	Word length, number of syllables, standard deviation
10	DESWLlt	Word length, number of letters, mean
11	DESWLltd	Word length, number of letters, standard deviation
Text Easability Principal Component Scores		
12	PCNARz	Text Easability PC Narrativity, z score
13	PCNARp	Text Easability PC Narrativity, percentile
14	PCSYNz	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	Text Easability PC Word concreteness, z score
17	PCCNCp	Text Easability PC Word concreteness, percentile
18	PCREFz	Text Easability PC Referential cohesion, z score
19	PCREFp	Text Easability PC Referential cohesion, percentile
20	PCDCz	Text Easability PC Deep cohesion, z score
21	PCDCp	Text Easability PC Deep cohesion, percentile
22	PCVERBz	Text Easability PC Verb cohesion, z score
23	PCVERBp	Text Easability PC Verb cohesion, percentile
24	PCCONNz	Text Easability PC Connectivity, z score
25	PCCONNp	Text Easability PC Connectivity, percentile
26	PCTEMPz	Text Easability PC Temporality, z score
27	PCTEMPp	Text Easability PC Temporality, percentile
Referential Cohesion		
28	CRFNO1	Noun overlap, adjacent sentences, binary, mean
29	CRFAO1	Argument overlap, adjacent sentences, binary, mean
30	CRFSO1	Stem overlap, adjacent sentences, binary, mean
31	CRFNOa	Noun overlap, all sentences, binary, mean
32	CRFAOa	Argument overlap, all sentences, binary, mean
33	CRFSOa	Stem overlap, all sentences, binary, mean
34	CRFCWO1	Content word overlap, adjacent sentences, proportional, mean

35	CRFCWO1d	Content word overlap, adjacent sentences, proportional, standard deviation
36	CRFCWOa	Content word overlap, all sentences, proportional, mean
37	CRFCWOad	Content word overlap, all sentences, proportional, standard deviation
38	CRFANP1	Anaphor overlap, adjacent sentences
39	CRFANPa	Anaphor overlap, all sentences
LSA		
40	LSASS1	LSA overlap, adjacent sentences, mean
41	LSASS1d	LSA overlap, adjacent sentences, standard deviation
42	LSASSp	LSA overlap, all sentences in paragraph, mean
43	LSASSpd	LSA overlap, all sentences in paragraph, standard deviation
44	LSAPP1	LSA overlap, adjacent paragraphs, mean
45	LSAPP1d	LSA overlap, adjacent paragraphs, standard deviation
46	LSAGN	LSA given/new, sentences, mean
47	LSAGNd	LSA given/new, sentences, standard deviation
Lexical Diversity		
48	LDTTRc	Lexical diversity, type-token ratio, content word lemmas
49	LDTTRa	Lexical diversity, type-token ratio, all words
50	LDMTLDa	Lexical diversity, MTLD, all words
51	LDVOCDa	Lexical diversity, VOCD, all words
Connectives		
52	CNCAll	All connectives incidence
53	CNCCaus	Causal connectives incidence
54	CNCLogic	Logical connectives incidence
55	CNCADC	Adversative and contrastive connectives incidence
56	CNCTemp	Temporal connectives incidence
57	CNCTempx	Expanded temporal connectives incidence
58	CNCAdd	Additive connectives incidence
59	CNCPos	Positive connectives incidence
60	CNCNeg	Negative connectives incidence
Situation Model		
61	SMCAUSv	Causal verb incidence
62	SMCAUSvp	Causal verbs and causal particles incidence
63	SMINTEp	Intentional verbs incidence
64	SMCAUSr	Ratio of casual particles to causal verbs
65	SMINTEr	Ratio of intentional particles to intentional verbs
66	SMCAUSlsa	LSA verb overlap
67	SMCAUSwn	WordNet verb overlap
68	SMTEMP	Temporal cohesion, tense and aspect repetition, mean
Syntactic Complexity		
69	SYNLE	Left embeddedness, words before main verb, mean
70	SYNNP	Number of modifiers per noun phrase, mean
71	SYNMEDpos	Minimal Edit Distance, part of speech

72	SYNMEDwrd	Minimal Edit Distance, all words
73	SYNMEDlem	Minimal Edit Distance, lemmas
74	SYNSTRUTa	Sentence syntax similarity, adjacent sentences, mean.
75	SYNSTRUTt	Sentence syntax similarity, all combinations, across paragraphs, mean
Syntactic Pattern Density		
76	DRNP	Noun phrase density, incidence
77	DRVP	Verb phrase density, incidence
78	DRAP	Adverbial phrase density, incidence
79	DRPP	Preposition phrase density, incidence
80	DRPVAL	Agentless passive voice density, incidence
81	DRNEG	Negation density, incidence
82	DRGERUND	Gerund density, incidence
83	DRINF	Infinitive density, incidence
Word Information		
84	WRDNOUN	Noun incidence
85	WRDVERB	Verb incidence
86	WRDADJ	Adjective incidence
87	WRDADV	Adverb incidence
88	WRDPRO	Pronoun incidence
89	WRDPRP1s	First person singular pronoun incidence
90	WRDPRP1p	First person plural pronoun incidence
91	WRDPRP2	Second person pronoun incidence
92	WRDPRP3s	Third person singular pronoun incidence
93	WRDPRP3p	Third person plural pronoun incidence
94	WRDFRQc	CELEX word frequency for content words, mean
95	WRDFRQa	CELEX Log frequency for all words, mean
96	WRDFRQmc	CELEX Log minimum frequency for content words, mean
97	WRDAOAc	Age of acquisition for content words, mean
98	WRDFAMc	Familiarity for content words, mean
99	WRDCNCc	Concreteness for content words, mean
100	WRDIMGc	Imagability for content words, mean
101	WRDMEAc	Meaningfulness, Colorado norms, content words, mean
102	WRDPOLc	Polysemy for content words, mean
103	WRDHYPn	Hypernymy for nouns, mean
104	WRDHYPv	Hypernymy for verbs, mean
105	WRDHYPnv	Hypernymy for nouns and verbs, mean
Readability		
106	RDFRE	Flesch Reading Ease
107	RDFKGL	Flesch-Kincaid Grade Level
108	RDL2	Coh-Metrix L2 Readability

ANEXO B – ÍNDICES DO COH-METRIX-PORT 3.0

Basic Counts
Adjective incidence
Adverb incidence
Content word incidence
Flesch index
Function word incidence
Mean sentences per paragraph
Mean syllables per content word
Mean words per sentence
Noun incidence
Number of Paragraphs
Number of Sentences
Number of Words
Pronoun incidence
Verb incidence
Logic operators
Incidence of ANDs.
Incidence of IFs.
Incidence of ORs.
Incidence of negations
Logic operators incidence
Content word frequencies
Content words frequency
Minimum among content words frequencies
Hypernyms
Mean hypernyms per verb
Tokens
Personal pronouns incidence
Type to token ratio
Constituents
Noun Phrase Incidence

Words before Main Verb
Connectives
Connectives incidence
Incidence of additive negative connectives
Incidence of additive positive connectives
Incidence of causal negative connectives
Incidence of causal positive connectives
Incidence of logical negative connectives
Incidence of logical positive connectives
Incidence of temporal negative connectives
Incidence of temporal positive connectives
Ambiguity
Ambiguity of adjectives
Ambiguity of adverbs
Ambiguity of nouns
Ambiguity of verbs
Coreference
Adjacent argument overlap
Argument overlap
Adjacent stem overlap
Stem overlap
Adjacent content word overlap
Anaphoras
Adjacent anaphoric references
Anaphoric references