Mandinga ISSN:25263455

ACESSIBILIDADE TEXTUAL E TERMINOLÓGICA: NOVOS GLOSSÁRIOS SOBRE ONCOLOGIA PARA A FERRAMENTA MEDSIMPLES

Guillermo Silva VILLAR (UFRGS, PROBIT-FAPERGS)[1]
Maria José Bocorny FINATTO (UFRGS, PPG-LETRAS, CNPq)[2]

Resumo: Este artigo relata as etapas da construção do novo módulo de Oncologia para a Ferramenta MedSimples (https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/), um recurso para auxiliar a simplificação de textos sobre temas de Saúde para leitores de escolaridade limitada. Destacam-se a construção de definições acessíveis para um conjunto de terminologias médicas e a proposição de conjuntos sinonímicos facilitados para itens considerados difíceis que não sejam terminologias. A base teórico-metodológica do trabalho inclui princípios da Acessibilidade Textual e Terminológica, da Terminografia Didático-Pedagógica e da Linguística de *Corpus*. Depois de reunir um *corpus* sobre Oncologia, destacando-se textos para leigos, processou-se o material com a ferramenta AntConc para levantamento de terminologias e de itens lexicais mais empregados. Em seguida, esses elementos foram comparados com a base do CorPop, um *corpus* de referência para situar possíveis níveis de simplicidade vocabular. Após a obtenção de listagens de terminologias (A) e de palavras potencialmente difíceis (B), inicia-se o estudo de definições acessíveis e sinônimos/alternativas mais simples. Para (A), um exemplo é “angiogênese”, explicado, inicialmente, como "formação de novos vasos sanguíneos”. Para (B), um exemplo é “erradicado", com as alternativas “acabado, eliminado, exterminado”. Assim, situam-se dois produtos e processos distintos, produzidos em uma mesma lógica, salientando-se que o CorPop também baliza a escolha das palavras para o enunciado da definição e para um *ranking* dos possíveis sinônimos a sugerir. Além disso, todas as definições terão revisão conceitual por profissionais da saúde que atuam em Oncologia. Os resultados apontam o bom potencial dos dados produzidos e a necessidade de uma série de decisões a serem tomadas quanto à compatibilidade, sobreposições e compartilhamento das informações do módulo de Oncologia com as informações dos demais módulos temáticos já existentes nas bases de dados da ferramenta.

Palavras-chave: acessibilidade textual e terminológica; linguagem simples; lexicometria; oncologia

Abstract: This work reports the basic steps in constructing the new Oncology module for the MedSimples Tool (https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/), a tool designed to simplify healthcare texts for readers with limited education. The focus here is on creating accessible definitions for a set of medical terminologies and proposing simplified synonym sets for challenging items that are not terminologies. The theoretical and methodological basis of this work includes principles of Textual and Terminological Accessibility, Didactic-Pedagogical Terminography, and Corpus Linguistics. After gathering a corpus on Oncology, with a specific

[1] Acadêmico de Letras da Universidade Federal do Rio Grande do Sul (UFRGS), Bacharelado em Letras, hab. Tradutor Português-Francês. Bolsista de Iniciação Científica Tecnológica do Programa PROBIT-FAPERGS-UFRGS-CAPES. E-mail: gsilvavillar@gmail.com ORCID: https://orcid.org/0009-0009-5199-5689

[2] Doutora em Letras/Estudos da Linguagem pela UFRGS, Docente do PPG-Letras-UFRGS, pesquisadora PQ do CNPq. E-mail: maria.finatto@gmail.com ORCID: https://orcid.org/0000-0002-6022-8408

emphasis on texts for laypeople, the material was processed using the AntConc tool to identify terminologies and frequently used lexical items. These elements were then compared with the CorPop reference corpus to assess possible levels of lexical simplicity. Following the compilation of terminology lists (A) and potentially difficult words (B), the study of accessible definitions and simpler synonyms/alternatives began. For (A), an example is "angiogenesis," initially explained as "the formation of new blood vessels." For (B), an example is "eradicated," with alternative options such as "finished, eliminated, exterminated." Thus, two distinct products and processes are demonstrated, both produced under the same logic, emphasizing that the CorPop helps the selection of words for definition statements and for ranking suggested synonyms. In addition, all definitions will still have a conceptual review by Health professionals who work in Oncology. The results point to the good potential of the data produced, and the need for a series of decisions to be taken regarding the compatibility, overlapping and sharing of information from the Oncology module with other thematic modules that already exist in the tool's databases.

Keywords: textual and terminological accessibility; plain language; lexicometry; oncology

Introdução

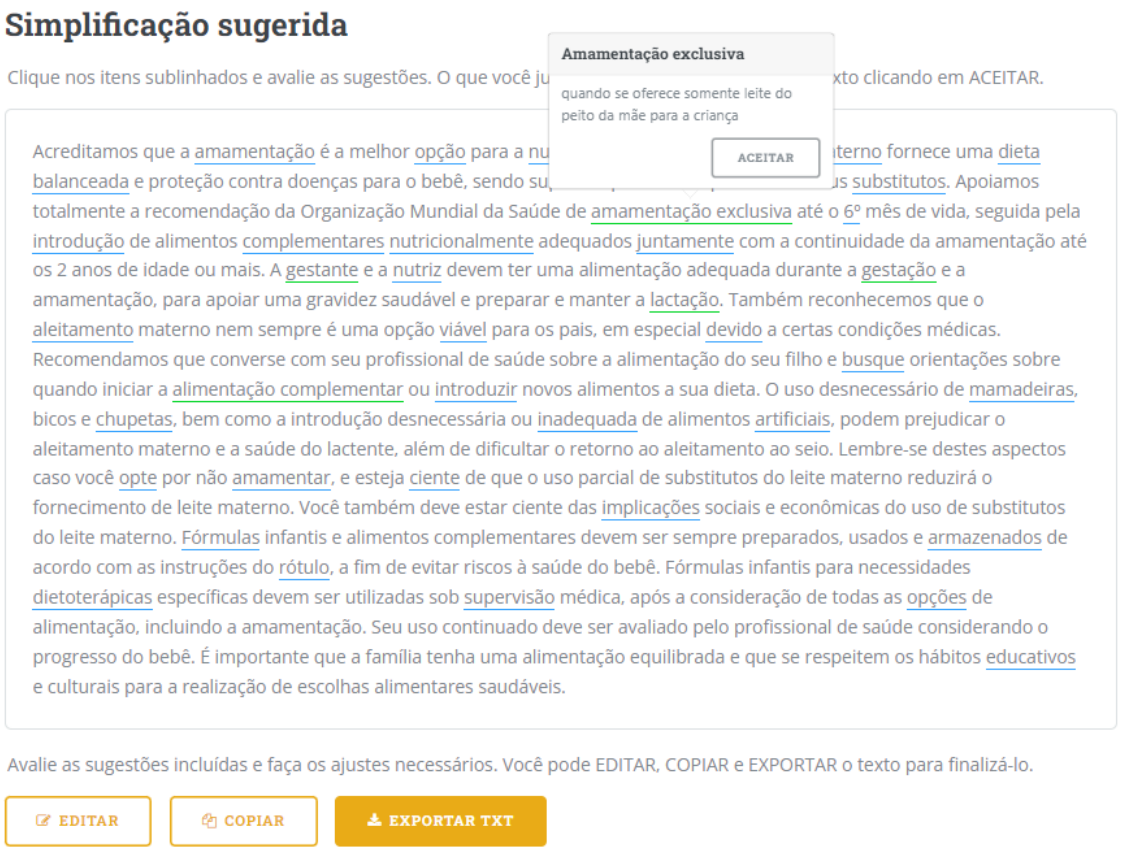
A Ferramenta MedSimples (doravante MedSimples) é um sistema *on-line* gratuito para apoiar a redação facilitada de textos sobre diferentes temas de Saúde para pessoas leigas. É um sistema disponível no *website* (https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/), no sistema de dados da Universidade Federal do Rio Grande do Sul (UFRGS). Sua proposta, internacionalmente premiada no *Google LARA Awards* de 2019, é auxiliar redatores a produzir textos para atender leitores-destinatários adultos com escolaridade limitada e poucas experiências de leitura.

MedSimples integra-se a um ambiente com diferentes insumos para auxiliar a escrita simplificada, como glossários e guias, com funcionamento bastante facilitado. O usuário precisa seguir apenas estas etapas:

1. Inserir, na ferramenta, um texto que julga complexo e que deseja simplificar;
2. Selecionar o **tema médico** em foco e o **perfil do leitor** a quem se dirige;
3. Clicar em “Enviar”.

O texto é, então, processado e reapresentado ao usuário com uma série de sugestões de reescrita e/ou de melhorias em forma de sublinhas coloridas. As palavras e expressões sublinhadas ligam-se a *pop-ups* com simplificações para termos técnicos ou com alternativas de sinônimos mais fáceis para as palavras potencialmente difíceis. Em seguida, o usuário poderá exportar o texto com as sugestões ou alertas gerados pela MedSimples, podendo ou não incorporá-los na sua edição final. Essa descrição está resumida na imagem a seguir. Nela, vemos um texto submetido ao módulo de “cuidados com o recém-nascido” (Pediatria).

Figura 1. Interface da Ferramenta MedSimples



Fonte: elaborada pelos autores. Texto processado na Ferramenta MedSimples, módulo Pediatria, “cuidados com o recém-nascido”, leitor tipo 1 – Ensino Fundamental.

Segue aqui a transcrição do texto acima processado na ferramenta, para a consideração, pelo nosso leitor, da sua potencial complexidade para o leitor em foco – uma pessoa adulta com escolaridade limitada ao Ensino Fundamental completo:

> Acreditamos que a amamentação é a melhor opção para a nutrição de lactentes, pois o leite materno fornece uma dieta balanceada e proteção contra doenças para o bebê, sendo superior quando comparado aos seus substitutos. Apoiamos totalmente a recomendação da Organização Mundial da Saúde de amamentação exclusiva até o 6º mês de vida, seguida pela introdução de alimentos complementares nutricionalmente adequados juntamente com a continuidade da amamentação até os 2 anos de idade ou mais. A gestante e a nutriz devem ter uma alimentação adequada durante a gestação e a amamentação, para apoiar uma gravidez saudável e preparar e manter a lactação. Também reconhecemos que o aleitamento materno nem sempre é uma opção viável para os pais, em especial devido a certas condições médicas. Recomendamos que converse com seu profissional de saúde sobre a alimentação do seu filho e busque orientações sobre quando iniciar a alimentação complementar ou introduzir novos alimentos a sua dieta. O uso desnecessário de mamadeiras, bicos e chupetas, bem como a introdução desnecessária ou inadequada de alimentos artificiais, podem prejudicar o aleitamento materno e a saúde do lactente, além de dificultar o retorno ao aleitamento ao seio. Lembre-se destes aspectos caso você opte por não amamentar, e esteja ciente de que o uso parcial de substitutos do leite materno

> reduzirá o fornecimento de leite materno. Você também deve estar ciente das implicações sociais e econômicas do uso de substitutos do leite materno. Fórmulas infantis e alimentos complementares devem ser sempre preparados, usados e armazenados de acordo com as instruções do rótulo, a fim de evitar riscos à saúde do bebê. Fórmulas infantis para necessidades dietoterápicas específicas devem ser utilizadas sob supervisão médica, após a consideração de todas as opções de alimentação, incluindo a amamentação. Seu uso continuado deve ser avaliado pelo profissional de saúde considerando o progresso do bebê. É importante que a família tenha uma alimentação equilibrada e que se respeitem os hábitos educativos e culturais para a realização de escolhas alimentares saudáveis.
> Fonte: Nestlé & Me Brasil - "Nota importante", disponível em: https://www.nestlebabyandme.com.br/?gclid=Cj0KCQiAutyfBhCMARIsAMgcRJScUtZMis8R3sPiU6cwsUKqZNh8AhiDI17Lin4RHUC9n7xmxlXSDkQaArbdEALw_wcB.
> Acesso em 23 fev. 2023.

Embora centrada apenas no enfrentamento da potencial complexidade do vocabulário dos textos sobre temas médicos, a MedSimples funciona em um ambiente que oferece também:

a) informações suplementares e didáticas sobre como avaliar e contornar a complexidade frasal-sintática do texto;

b) glossários simplificados, em formato resumido, organizados por temas médicos específicos;

c) dados de pesquisa sobre o tema da Acessibilidade Textual Terminológica (ATT) e da simplificação de textos sobre diferentes temas científicos, como grupo de estudos e publicações da equipe de trabalho: https://www.ufrgs.br/textecc/acessibilidade/page/index/

A Ferramenta MedSimples foi inspirada e construída com base em princípios da ATT (FINATTO, PARAGUASSU, 2022), da Terminografia Didático-Pedagógica (FADANELLI, 2017), da Linguística de *Corpus* (BERBER SARDINHA, 2004), do Processamento da Linguagem Natural (CASELI, NUNES, 2023) e da Lexicometria (SANROMÁN, DOCÍO, 2022). Essas referências, especialmente a ATT e a Lexicometria, dão suporte às escolhas automáticas que incidem sobre a simplificação do vocabulário de textos considerados complexos. Afinal, a partir de um enfoque e tratamentos estatísticos, aplicado a um conjunto de textos de estudo, que servem como referência de vocabulário em um dado tema médico, serão selecionadas e apontadas, automaticamente, alternativas para a melhoria de um texto de entrada de interesse do usuário. Assim, o Processamento da Linguagem Natural e a Linguística de *Corpus* fornecem orientações metodológicas tanto para o tratamento estatístico do vocabulário dos *corpora* de estudo que subsidiam a ferramenta, quanto para o que, enfim, a MedSimples consegue realizar.

Com tais referências teórico-práticas, o sistema MedSimples apoia-se em um conjunto de bases com dados interconectadas, associado aos diferentes assuntos médicos cobertos em seus módulos. Essas bases contêm, em cada tema:

i) um sistema de *parsing*, o PassPort (ZILIO, WILKENS, FAIRON, 2018), que apenas separa e etiqueta cada uma das palavras de um texto de entrada – o texto que se deseja simplificar;

ii) listagens de palavras/itens/expressões fáceis, que, quando identificadas no texto-entrada, serão ignoradas pela ferramenta após o *parsing*;

iii) informações sobre palavras/itens/expressões potencialmente difíceis a marcar no texto de entrada, com possíveis sinônimos mais simples;

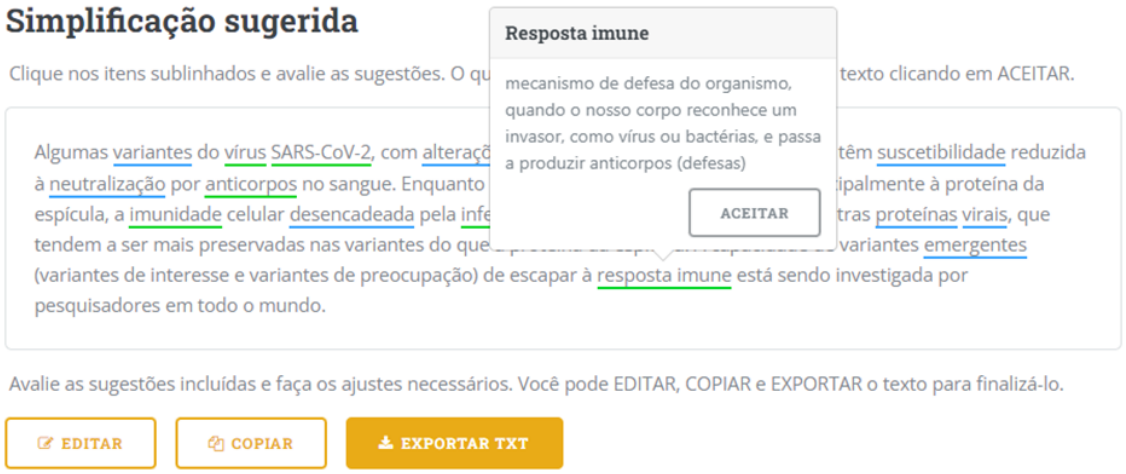
iv) listagens de terminologias e de expressões técnicas e respectivas definições e/ou explicações simplificadas, que são os "glossários internos" de cada um dos assuntos médicos cobertos.

Conforme mencionado, essas bases são previamente alimentadas pelo processamento computacional de diferentes *corpora* textuais tomados como fonte para o reconhecimento do vocabulário empregado. Isto é, essas bases são fruto do reconhecimento de usos linguísticos mais e menos frequentes em diferentes conjuntos de textos que tratam de temas médicos. Foram compilados, especialmente, textos de divulgação científica para leigos, divididos por diferentes temas. Esses temas correspondem aos três atuais módulos, já implantados na MedSimples:

A) Doença de Parkinson (Neurologia);
B) Cuidados com o recém-nascido (Pediatria);
C) COVID-19 (Infectologia).

Com apoios do programa PROBIT-FAPERGS-UFRGS, do CNPq e do PPG-Letras-UFRGS, está sendo construído um novo módulo da MedSimples, agora no tema Câncer/Oncologia. Como já citado, a nossa ferramenta apenas mostra, como *output*, sugestões de escrita/reformulação geradas automaticamente, conforme o tema selecionado e um perfil de escolaridade do leitor destinatário: pessoa com Ensino Fundamental ou com Ensino Médio. Assim, não é feita, diretamente, uma simplificação automática. As sugestões de reescrita geradas pela MedSimples estão exemplificadas na Figura 2, que oferece uma visualização do atual funcionamento do módulo sobre COVID-19.

Figura 2. Interface da Ferramenta MedSimples



Fonte: elaborada pelos autores. Texto processado na Ferramenta MedSimples módulo de “COVID-19”, leitor tipo 1 – Ensino Fundamental, teste em julho de 2023.

Segue aqui a transcrição do texto acima processado:

> Algumas variantes do vírus SARS-CoV-2, com alterações importantes na proteína da espícula, têm suscetibilidade reduzida à neutralização por anticorpos no sangue. Enquanto os anticorpos neutralizantes visam principalmente à proteína da espícula, a imunidade celular desencadeada pela

> infecção natural também tem como alvo outras proteínas virais, que tendem a ser mais preservadas nas variantes do que a proteína da espícula. A capacidade de variantes emergentes (variantes de interesse e variantes de preocupação) de escapar à resposta imune está sendo investigada por pesquisadores em todo o mundo.
>
> Fonte: Informe Científico, de 10 de maio de 2021, da Organização Pan-Americana de Saúde (OPAS), Imunidade natural contra a COVID-19. Principais mensagens, Item 3. Disponível em: https://iris.paho.org/bitstream/handle/10665.2/54855/OPASWBRAPHECOVID-19210074_por.pdf. Acesso em 19 Dez. 2023.

Conforme a figura 2, o *output* traz itens com sublinhas azuis e verdes, que assinalam prováveis pontos a melhorar/simplificar. Em verde, são assinaladas automaticamente as terminologias do texto (como **vírus**, **anticorpos** e **resposta imune**), com *pop-ups* para sugestões de definições facilitadas. Em azul, a ferramenta identificou palavras complexas de tipo geral (como **suscetibilidade**) associadas a *pop-ups* com possíveis sinônimos mais fáceis. As sugestões, assim categorizadas, podem então ser avaliadas e incorporadas ou não na edição do texto, o que é feito pelo usuário, no ambiente da ferramenta. Feito isso, exporta-se o trabalho para finalizá-lo.

Como se pode depreender pela observação atenta das duas figuras anteriores, a marcação de terminologias (em verde) e das palavras potencialmente difíceis (em azul) pode ser bastante melhorada em cada um dos módulos atuais da MedSimples. Como se percebe, na figura 1, temos marcas equivocadas nos itens **mamadeiras, chupetas** e **dietoterápicas**. Ao poder-se implantar um novo módulo na ferramenta abre-se, justamente, uma "janela de oportunidades" para uma revisão geral de cada módulo já em funcionamento, em diferentes frentes. Por isso, em Câncer/Oncologia, aproveita-se toda uma série de insumos preexistentes, ao passo que suas novas bases de dados são criadas. Esse é o processo que buscamos relatar neste artigo.

Na sequência, este texto terá o seguinte encaminhamento: na próxima seção, tratamos do módulo de Oncologia e caracterizamos seus *corpora* de estudo. São apresentados os repertórios de terminologias e de palavras complexas já levantados. Depois, trazemos os desafios do processamento de um texto-exemplo, apontando acertos e problemas a resolver, especialmente quanto à elaboração de definições acessíveis para as terminologias mais recorrentes. Por fim, situamos a inserção dos novos dados obtidos e colocamos questões sobre a compatibilização dos novos insumos com aqueles de outros temas médicos já cobertos pela ferramenta.

O módulo de Oncologia e seu *corpora* de estudo

O novo módulo da MedSimples atenderá necessidades de simplificação do vocabulário de textos voltados para leitores-destinatários do Tipo 1, isto é, adultos com escolaridade limitada ao Ensino Fundamental completo. Esse módulo é construído a partir de dados identificados em uma base textual (os *corpora* de estudo). Essa base contém uma seleção de textos de diferentes fontes de divulgação para leigos sobre Oncologia, as quais foram aprovadas pelo consultor médico da equipe.

Inicialmente, para obter um panorama do vocabulário mais usual ou recorrente, partimos de uma obra de divulgação para leigos intitulada, “Câncer – Uma breve introdução” de Nicholas James (JAMES, 2018). Esse texto foi traduzido e publicado no Brasil pela editora L&PM, em 2018. Sua escolha deu-se também em função desse livro trazer informações para quem desejar uma primeira mirada ao tema e, em tese, poder

espelhar um repertório básico de palavras e de terminologias nesse tema. Houve, ainda, outra razão para a escolha: a tradução brasileira foi feita por uma profissional experiente, autora de trabalhos no tema da simplificação textual, Bianca Franco Pasqualini. Seu doutorado ocupou-se de reconhecer e compilar um *corpus* de referência para representar padrões vocabulares do Português Popular Escrito, *a priori*, mais acessível para pessoas de escolaridade limitada. Seu trabalho está concretizado nas ferramentas de acesso livre da base *CorPop* (https://www.ufrgs.br/textecc/porlexbras/corpop/index.php). Essa base voltará a ser citada mais adiante.

Entretanto, em uma primeira apreciação do texto de James (2018), ainda detectamos potenciais pontos de complexidade para o nosso perfil de leitor. Isso é o que exemplifica o trecho a seguir:

> As proteínas, como o DNA, são constituídas por cadeias de moléculas mais simples. Os componentes essenciais das proteínas, chamados aminoácidos, podem ser ligados entre si para formar cadeias praticamente infinitas. A molécula básica de aminoácido tem três características principais – denominadas carbóxi-terminal e (daí o nome) amino-terminal, além de uma ramificação lateral variável, que confere a cada aminoácido suas propriedades distintas (ilustrada como R na Figura 11). (JAMES, 2018)

Esse breve trecho sobre o DNA, sua formação e replicação, demonstra a necessidade de encontrarmos alternativas para facilitar a compreensão de um leitor adulto com escolaridade limitada. Também fica clara a necessidade de trazer informações facilitadas sobre palavras e sobre conceitos associados às terminologias empregadas.

Após uma apreciação geral do conteúdo vocabular da obra, passamos a uma testagem pontual de verificação com trechos na MedSimples, em seus módulos atuais. Mesmo em temas médicos diferentes (COVID-19, Doença de Parkinson e Pediatria), essa verificação sobre a potencial complexidade forneceu uma ideia sobre a cobertura e precisão das diferentes bases de dados atuais. Além disso, essa etapa foi útil para reconhecer o vocabulário comum aos diferentes temas já cobertos, indicando pontos de sobreposição, melhorias e ajustes.

Mesmo fora do tema de Oncologia, o sistema da ferramenta MedSimples reconheceu e assinalou que uma palavra como **molécula** é um item terminológico, acompanhando-o de uma sugestão de definição, em tese, mais acessível. Por outro lado, os itens marcados em azul foram reconhecidos como difíceis, itens como **denominadas,** que é assinalado com seus quatro prováveis sinônimos mais simples. Entretanto, houve alguma confusão, no caso de **amino** e **aminoácidos**, que não constam como terminologias nesse módulo. A seguir, podemos ver algumas das sugestões de reescrita para o trecho inserido na Ferramenta MedSimples. A seguir um exemplo em cores, conforme o sistema aponta.

> A **molécula** (a menor partícula em que um elemento pode ser dividido sem mudar suas propriedades; é como se fossem pequenos tijolos de uma estrutura maior e que não podem ser quebrados) básica de **amino** (Item difícil. Avalie trocar.) ácido tem três características principais - **denominadas** (chamar, classificar, designar, nomear) **carbóxi** (Item difícil. Avalie trocar.) **-terminal** (final, que causa morte) e (daí o nome) amino-terminal, além de uma **ramificação** (ramo, divisão, ação de se dividir em "braços" menores) lateral **variável** (o que muda, elemento ou fator), que confere a cada **aminoácido** (Item difícil. Avalie trocar.) suas propriedades distintas.

Esses testes de cobertura vocabular, com James (2018), indicaram também que essa obra centrou-se, tematicamente, em alguns tipos de câncer, em especial os cânceres de pulmão, de mama e de próstata. Esse destaque foi propositalmente feito, pelo autor, tendo em vista os índices de incidência que apontava, pela ordem: população geral, gênero feminino e gênero masculino. Assim, mostrou-se necessário suprir um déficit temático e de repertório vocabular importante no cenário brasileiro atual: palavras associadas ao tema do câncer de pele e ao câncer de pelo tipo melanoma. Por isso, agregamos ao *corpus* de estudo mais textos. A opção foi por textos de divulgação jornalística nesse tema específico. Assim, com a recomendação e aval do médico consultor, foram selecionados e acrescidos 46 novos textos aos *corpora* de estudo.

A seguir, vemos um trecho que exemplifica a natureza e os encaminhamentos desses 46 textos jornalísticos sobre melanoma, os quais foram agregados aos *corpora* de estudo:

> Se encontrar qualquer coisa estranha, por menor que seja, é preciso procurar um médico. O diagnóstico final é feito por meio de um exame de laboratório chamado biópsia. Na biópsia, o médico corta um pedacinho pequeno de pele, na parte suspeita, para ser analisado em um microscópio. Apenas o médico consegue chegar ao diagnóstico e dizer qual a melhor opção para tratamento. A escolha do tratamento depende principalmente se o câncer está só na pele ou se já se espalhou. Além desse exame de biópsia, conhecido como anatomopatológico, técnicas de exame de DNA também são boas opções para determinar o melhor tipo de tratamento para cada paciente.

Fixado um *corpus* de estudo final - que serve de referência vocabular para o novo módulo - passamos ao seu processamento para reconhecer, com o devido detalhamento estatístico, seu repertório vocabular. Esse trabalho de Lexicometria foi feito com apoio do *software* gratuito AntConc (ANTHONY, 2022), que nos fornece, automaticamente, listagens diversas e dados estatísticos sobre o vocabulário em um conjunto de textos. O *software* está disponível no *website*: <https://www.laurenceanthony.net/software/antconc/>. Esse recurso foi escolhido pelo seu bom desempenho e multifuncionalidades embutidas. Essas funcionalidades estão descritas e exemplificadas, com um passo a passo detalhado, em Finatto, Esteves, Villar (2022).

Assim, o *corpus* de estudo ficou composto por um universo de 74.408 *tokens* (palavras totais) e 7.287 *types* (formas diferentes de palavras). Além disso, temos um universo de palavras que mostra uma *TTR* (*type/token ratio,* ou a razão de formas diferentes de palavras pelo total de palavras no *corpus)* de 10,28%. De uma listagem geral, passamos para a produção de listagens de agrupamentos de 3 ou 4 palavras mais utilizadas ao longo dos textos, conhecidas como *clusters*. Com esses passos, pudemos observar as frequências pontuais também de expressões multipalavra, como seria o caso de **câncer do cérvice uterino.** O quadro a seguir resume os dados do *corpus* de estudo, conforme finalizado até julho de 2023.

Tabela 1. Resumo do *corpus* reunido para referência de repertório vocabular

Modo ou meio	Escrito
Tempo	Sincrônico
Seleção	Estático (não será atualizado) e de amostragem
Conteúdo	Monolíngue (português do Brasil) e de gênero misto sobre Oncologia
Autoria	Tradução e de língua nativa
Finalidade	De estudo
Fonte	01 Livro para leigos sobre Oncologia, 46 textos de instituições e organizações e portais de notícias
Volume	47 textos
Tokens (total de palavras)	74.408
Types (palavras diferentes)	7.287
TTR (*type-token* ratio/riqueza lexical)	10,28 %

Os itens das listas obtidas sejam terminologias, palavras potencialmente difíceis e/ou agrupamentos de palavras (*clusters*), passaram então para uma nova fase de análises e de leituras contextuais, na qual simulamos a interpretação do leitor tipo 1. Depois disso, cada elemento ou expressão identificados foram cotejados com as listas de palavras do *CorPop* (PASQUALINI, 2018). O *CorPop* nos fornece uma referência do repertório de palavras que seria potencialmente mais fácil ou mais complexo. Esse *corpus*, funcionando como um *corpus* de exclusão e guia, é acessado no endereço: https: //www.ufrgs.br/textecc/porlexbras/corpop/index.php. O cotejo entre o perfil do vocabulário do nosso *corpus* e o perfil de itens do *CorPop* está exemplificado na tabela a seguir:

Tabela 2. Exemplo de contrastes entre o *corpus* de estudo de Oncologia e o *corpus* de exclusão – verde: palavras potencialmente difíceis; azul: termos a definir ou explicar

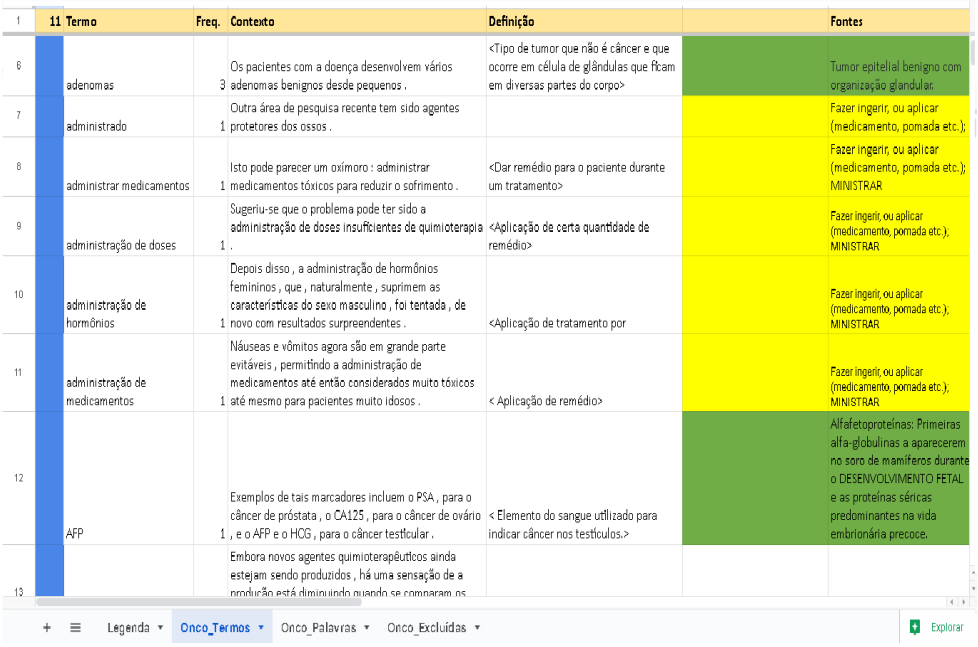
Palavra/ Terminologia/ Expressão Multipalavra	Frequência *corpus* Oncologia	Frequência *CorPop*
anticorpos	18	0
assimetria	17	0
benefício	18	14
câncer	863	7
melanoma acral	8	0

A validação de elementos a incorporar ao módulo de Oncologia dá-se por meio do comparativo de frequências acima exemplificado. Isso é feito via uma métrica de corte: todo item/palavra que ocorrer de zero até cinco vezes no *CorPop*, é marcado como potencialmente difícil. Retomando a tabela 2, vemos que isso ocorre no caso dos itens **anticorpos, assimetria e melanoma acral.** Assim esses elementos deverão constar na lista de vocabulário potencialmente difícil do novo módulo. Por outro lado, nos casos de **câncer** e **benefícios**, ambos possuem uma frequência dentro do *CorPop* superior à linha

de corte estabelecida (de zero até cinco). Para o caso do item **câncer**, sendo palavra tópico, permanecerá no repertório de termos com necessidade de alguma definição acessível. No caso de **benefícios**, a decisão foi pela exclusão do item do rol das palavras potencialmente difíceis, supondo-se que seja de uso corriqueiro e conhecido, o que consideramos validado pelo resultado de frequência observado no *CorPop*.

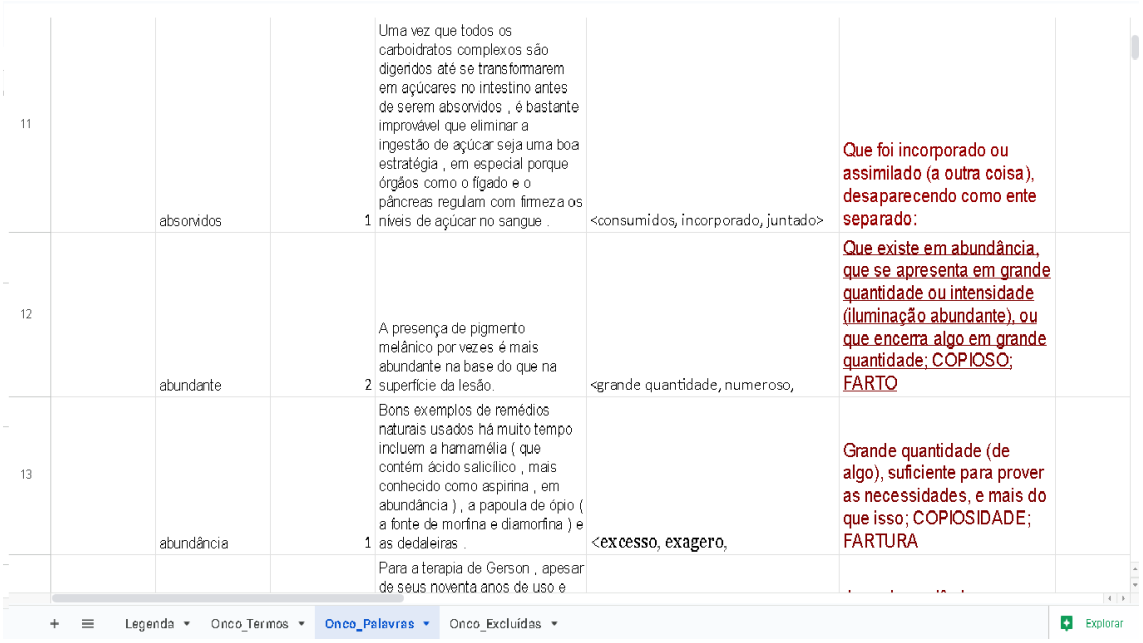
Planilhas de trabalho

Os dados levantados nos *corpora* de estudo e cotejados com o *CorPop* constituem itens potencialmente aproveitáveis no novo módulo de Oncologia. Esse universo vocabular é então organizado em uma planilha de trabalho que nos ajuda a categorizar os itens entre "palavras potencialmente difíceis" e "terminologias". Cada categoria de itens é armazenada em uma seção à parte. As terminologias recebem informações definicionais ou explicativas; os itens lexicais considerados difíceis são acompanhados de uma série sinonímica hierarquizada, denominada *synset*. Isso é o que se exemplifica nas Figuras a seguir:



1	11 Termo	Freq.	Contexto	Definição		Fontes
6	adenomas	3	Os pacientes com a doença desenvolvem vários adenomas benignos desde pequenos .	<Tipo de tumor que não é câncer e que ocorre em célula de glândulas que ficam em diversas partes do corpo>		Tumor epitelial benigno com organização glandular.
7	administrado	1	Outra área de pesquisa recente tem sido agentes protetores dos ossos .			Fazer ingerir, ou aplicar (medicamento, pomada etc.);
8	administrar medicamentos	1	Isto pode parecer um oxímoro : administrar medicamentos tóxicos para reduzir o sofrimento .	<Dar remédio para o paciente durante um tratamento>		Fazer ingerir, ou aplicar (medicamento, pomada etc.); MINISTRAR
9	administração de doses	1	Sugeriu-se que o problema pode ter sido a administração de doses insuficientes de quimioterapia .	<Aplicação de certa quantidade de remédio>		Fazer ingerir, ou aplicar (medicamento, pomada etc.), MINISTRAR
10	administração de hormônios	1	Depois disso , a administração de hormônios femininos , que , naturalmente , suprimem as características do sexo masculino , foi tentada , de novo com resultados surpreendentes .	<Aplicação de tratamento por		Fazer ingerir, ou aplicar (medicamento, pomada etc.), MINISTRAR
11	administração de medicamentos	1	Náuseas e vômitos agora são em grande parte evitáveis , permitindo a administração de medicamentos até então considerados muito tóxicos até mesmo para pacientes muito idosos .	< Aplicação de remédio>		Fazer ingerir, ou aplicar (medicamento, pomada etc.), MINISTRAR
12	AFP	1	Exemplos de tais marcadores incluem o PSA , para o câncer de próstata , o CA125 , para o câncer de ovário , e o AFP e o HCG , para o câncer testicular .	< Elemento do sangue utilizado para indicar câncer nos testículos.>		Alfafetoproteínas: Primeiras alfa-globulinas a aparecerem no soro de mamíferos durante o DESENVOLVIMENTO FETAL e as proteínas séricas predominantes na vida embrionária precoce.
13			Embora novos agentes quimioterapêuticos ainda estejam sendo produzidos , há uma sensação de a produção está diminuindo quando se comparam os			

Figura 3. Recorte da lista de trabalho (A) para os termos selecionados no *corpus* de estudo



11		absorvidos	1	Uma vez que todos os carboidratos complexos são digeridos até se transformarem em açúcares no intestino antes de serem absorvidos , é bastante improvável que eliminar a ingestão de açúcar seja uma boa estratégia , em especial porque órgãos como o fígado e o pâncreas regulam com firmeza os níveis de açúcar no sangue .	<consumidos, incorporado, juntado>	Que foi incorporado ou assimilado (a outra coisa), desaparecendo como ente separado:	
12		abundante	2	A presença de pigmento melânico por vezes é mais abundante na base do que na superfície da lesão.	<grande quantidade, numeroso,	Que existe em abundância, que se apresenta em grande quantidade ou intensidade (iluminação abundante), ou que encerra algo em grande quantidade; COPIOSO; FARTO	
13		abundância	1	Bons exemplos de remédios naturais usados há muito tempo incluem a hamamélia (que contém ácido salicílico , mais conhecido como aspirina , em abundância) , a papoula de ópio (a fonte de morfina e diamorfina) e as dedaleiras .	<excesso, exagero,	Grande quantidade (de algo), suficiente para prover as necessidades, e mais do que isso; COPIOSIDADE; FARTURA	
				Para a terapia de Gerson , apesar de seus noventa anos de uso e			

Legenda | Onco_Termos | Onco_Palavras | Onco_Excluídas | Explorar

Figura 4. Recorte da lista de trabalho (B) para a recolha de palavras e expressões potencialmente difíceis.

A partir do trabalho com os *corpora d*e estudo e cotejo com o *corpus* de exclusão (*CorPop*), temos já uma coleta avançada. Em julho de 2023, já alcançávamos 693 itens na lista de termos técnicos (A), enquanto a lista de palavras potencialmente difíceis (B) continha 300 itens. Dessas listagens, fizemos duas bases de dados internas para a alimentação da MedSimples. Um futuro produto dessas listagens é a organização de um conjunto de fichas terminológicas e fichas lexicais, que serão exemplificadas mais adiante, organizadas conforme a proposta de Esteves (2023).

Listagens em detalhe

Para ambas as listagens, de terminologias (A) e de palavras potencialmente difíceis (B), estabelecemos as seguintes categorias e/ou *tags* de identificação: **item; contexto; definição coletada; definição sugerida; observação**. São exemplos de **item,** em (A): **angiogênese, apoptose, célula** e **cromossomo**. No grupo (B), temos, por exemplo: **abrangente, acometer, detecção precoce** e **eficácia**.

O **contexto** corresponde a um excerto retirado dos textos em que o **item** ocorre. Houve um esforço para que o excerto selecionado contivesse algum elemento explicativo ou qualificador. Como exemplo, podemos apresentar o contexto para o item **apoptose**:

> Isso exige que as células desnecessárias sejam excluídas e eliminadas com o mínimo de perturbação (um processo chamado apoptose, a partir da palavra grega que significa "queda de pétalas"). Fonte: *corpus* Oncologia

Nesse segmento, temos uma descrição do fenômeno de forma clara e, quase como um bônus, um pouco de etimologia sobre o **item** em foco.

As **definições coletadas** provêm tanto de dicionários especializados, como também de variados textos de divulgação científica e fontes de apoio, como materiais do

INCA (Instituto Nacional do Câncer) e de dicionários de uso geral. Frisamos que uso de material mais técnico, voltado especialmente para profissionais da área médica, é feito apenas na composição do repertório de lista de termos técnicos. Para nosso uso, focamos em alguns recursos de apoio de acesso gratuito, tais como: o *site* Descritores de Ciências da Saúde (DeCS), acessado em: https://decs.bvsalud.org/ ; a Biblioteca Virtual em Saúde (BVS), disponível em https://bvsalud.org/; o portal do médico Drauzio Varella, em https://drauziovarella.uol.com.br/ ; o *site* do INCA, https://www.gov.br/inca/pt-br ; o dicionário *on-line* Caldas Aulete, em https://aulete.com.br/ ; assim como no dicionário Priberam *on-line*, que consulta-se no endereço https://dicionario.priberam.org/ .

Para ilustrar informações coletadas para **apoptose**, vejamos o conteúdo do *site* DeCS – antes mencionado:

> Mecanismo regulado de morte celular caracterizado por alterações morfológicas distintas no núcleo e no citoplasma, incluindo a clivagem endonucleolítica do DNA genômico em regiões internucleossômicas regularmente espaçadas, isto é, FRAGMENTAÇÃO DE DNA. É programada geneticamente e funciona como um equilíbrio à mitose na regulação do tamanho dos tecidos animais e na mediação de processos patológicos associados com crescimento tumoral. (Grifos do original)

A **definição sugerida**, por sua vez, é um dos principais campos da nossa base. A parte das terminologias e sua apresentação com definições potencialmente mais compreensíveis são o cerne do trabalho que estamos desenvolvendo, o objetivo final do projeto, por assim dizer. É o elemento mais almejado e também o de produção mais complexa, afinal envolve transformar definições especializadas em enunciados mais simples, com o cuidado de não deturpar a informação.

Como alternativa para a formulação definicional simples, recorremos, novamente, à base de palavras do CorPop. Conforme já mencionado, palavras que ocorram mais do que cinco vezes no CorPop são consideradas válidas para integrar o conjunto vocabular do enunciado definicional. Entretanto, em determinados casos, é incontornável a utilização de palavras fora desse repertório. É o caso de nosso exemplo:

[**apoptose**]: </def< Processo de autodestruição natural das células. > <def/>

Nesse caso, as palavras **autodestruição** e **célula** têm frequências inferiores ao nosso ponto de corte no CorPop, mas foram mantidas nesse enunciado que ainda carece de revisão conceitual-médica.

Por fim, como indexador da base de dados para ferramenta operar, temos a *tag* **observações.** Esse campo da base contém anotações específicas sobre os itens coletados, sendo um conteúdo de uso interno da nossa equipe. No momento, tal espaço vem sendo utilizado para a coleta de simplificações já contidas dentro da Ferramenta MedSimples, disponíveis nos outros módulos (Doença de Parkinson, Cuidados de recém-nascido e Covid-19). Esse campo possibilita um contraponto e crítica de novos enunciados e os já utilizados em outros temas, sobre um mesmo item.

Glossários internos

Partindo das listagens de trabalho e dos dados categorizados, é necessário produzir uma síntese de informações objetivas que guiam o funcionamento da ferramenta. Essas informações são: o **item** a apontar via sublinha, o seu estatuto (termo ou palavra – cor da sublinha), suas informações gramaticais e sintáticas (*features* úteis para o reconhecimento item em meio a um texto inserido pelo usuário) e os elementos substitutivos associados ao item: conjunto de sinônimos ranqueados OU definição simplificada.

Essa sistemática de registro, com as palavras que são terminologias, pode ser resumida na tabela 3 a seguir, com o item **angiogênese**. Depois, temos o caso das palavras que não correspondem a terminologias, representado pelo verbo **mitigar**.

Tabela 3. Recorte do glossário interno MedSimples

Item	*Feature 1* Categoria gramatical	*Feature 2* Função sintática	**definição sugerida**
angiogênese (singular) **angiogêneses** (plural) gerar=> [sublinha azul]	Substantivo Sing. Substantivo Pl.	Sujeito SUBJ Núcleo do sujeito SN Objeto direto Complemento nominal	<Formação de novos vasos sanguíneos. >

Item	*Feature 1* Categoria gramatical	*Feature 2* Função sintática	**definição sugerida** **[*synset*]**
mitigar [lista: flexões possíveis do item, tempos e modos] gerar => [sublinha azul]	Verbo [Infinitivo] OU Substantivo [deverbal nominalizado]	Sujeito Predicado Núcleo do predicado [várias possibilidades]	<conter; reduzir; amortecer>

Como explicar terminologias de modo acessível?

As definições que estamos desenvolvendo para os itens terminológicos seguem uma idealização teórica denominada por Esteves (2023) como "definição acessível". Em termos de estrutura do enunciado, buscamos utilizar uma sistemática definicional considerada "clássica", trazendo duas categorias semânticas conhecidas como *gênero próximo* e *diferenças específicas*.

Vejamos um exemplo com o termo **apoptose.** Esse termo corresponde a um aspecto intrinsecamente vinculado ao câncer, especialmente para a compreensão do seu surgimento e desenvolvimento acelerado. A definição coletada em dicionário especializado (DECS, 2022), como já vimos é:

> **Mecanismo** regulado de **morte celular** caracterizado por alterações morfológicas distintas no núcleo e no citoplasma, incluindo a clivagem endonucleolítica do DNA genômico em regiões internucleossômicas regularmente espaçadas, isto é, **FRAGMENTAÇÃO** DE DNA. É programada geneticamente e funciona como um equilíbrio à mitose na regulação do tamanho dos tecidos animais e na mediação de processos patológicos **associados com crescimento tumoral**. (DECS, 2020, grifos no original)

Em função de sua complexidade, tal enunciado não poderia ser diretamente aproveitado, ainda que contenha alguns elementos informacionalmente úteis (ou pistas semânticas) para quem elabora uma versão acessível – que grifamos. Assim sendo, foi necessário, buscar outras fontes de apoio, mais simples. Até o momento, temos a seguinte proposta de definição:

> <Processo de autodestruição [gênero próximo] natural das células [diferença específica]>

Como ocorre em todo processo de simplificação, que associamos a algo semelhante a uma "tradução intralinguística", sempre há espaço para melhorias. Utilizar a palavra **célula** pode ser um complicador, entretanto, conforme já mencionado, pareceu algo inevitável. Esse enunciado é um esboço e ainda deverá ser avaliado por consultores especializados.

Alternativas de sinônimos – como ajudar o redator a substituir palavras potencialmente difíceis?

As palavras/expressões potencialmente difíceis, por sua vez, requerem outro tratamento. Como já reiterado, apresentam-se apenas sugestões substitutivas, que o redator precisará "encaixar" no seu texto, conforme sua escolha e/ou decisão. Quando não temos o conteúdo associado ao item identificado na base de dados, é exibido apenas um tipo de alerta: [**item difícil, avalie trocar**].

Nesse caso, como um exemplo ilustrativo, temos o item **alastrar**. Esse é um verbo frequentemente encontrado, em diferentes formatos e/ou flexões, geralmente acompanhado de partícula reflexiva SE, em nosso *corpus* de textos sobre câncer. Esse item associa-se ao caráter de contínua expansão de alguns tipos da doença. No módulo da doença de Parkinson, a ferramenta já mostra a seguinte sugestão, que podemos considerar, hoje, bastante criticável: <**espalhar, circular, correr, passar, rever, rolar**>. No novo

módulo de Oncologia, a opção, até o momento, para o *synset* de **alastrar (-se)** é a seguinte: <**espalhar, ampliar, alongar**>.

Cabe dizer que as diferenças de sugestões de outros módulos/temas podem ser justificadas em função dos seus diferentes "encaixes" substitutivos, conforme o emprego desse verbo em diferentes enunciados, em cada especialidade médica. Isso é exemplificado a seguir, na situação em que um sinônimo **alongar** não seria uma alternativa adequada quando o tema do texto é o câncer:

> Medicamentos quimioterápicos, tais como dacarbazina e temozolomida, podem ser administrados pela veia para tratar os melanomas que se **alastraram** [**espalharam**/**ampliaram**], mas não prolongam a sobrevida e são normalmente administrados a pessoas que não têm outras opções. (Fonte: *Corpus* Oncologia)

Assim, em resumo, as sugestões devem ser ordenadas por critério de acessibilidade/facilidade, balizadas por suas frequências no *CorPop*. Mas, além disso, o conjunto de alternativas precisa ter compatibilidade semântica com o campo temático. Assim, ao trabalhar nos *synsets* do módulo de Oncologia, encontrando o item em outros módulos, podemos criticar o que já temos e melhorar o potencial de aproveitamento das formas sugeridas neste novo cenário de textos e também no contexto preexistente.

Extração de itens da base textual de Oncologia – conexões que se revelam

Um aspecto que merece ser comentado é a continuidade do processo de melhoramento da ferramenta. Como a MedSimples tem já três módulos temáticos implantados, há todo um extenso repertório atual de terminologias e de palavras potencialmente difíceis alocado em três bases diferentes. Assim, para o esforço de melhoria, impõe-se também a crítica e o reaproveitamento dos insumos já aplicados.

Como exemplo, temos o item **biópsia**, que integra o repertório de terminologias do módulo sobre a Doença de Parkinson. Nesse módulo, a sugestão de definição é a seguinte:

> **procedimento no qual se colhe uma amostra de tecidos ou células do nosso corpo para ser estudada em laboratório; exame muito usado para verificar câncer**.

Em Oncologia, após o exame contextual dos usos desse item, chegamos à seguinte formulação:

> **quando se corta um pedacinho de uma parte do corpo para ser examinado em um laboratório**.

Essa nova sugestão de definição facilitada para o item **biópsia** é ainda um rascunho. De todo modo, caso seja considerada conceitualmente adequada, poderá ser aproveitada em Oncologia e compartilhada nos outros módulos temáticos e/ou bases de dados da ferramenta, como Pediatria, COVID-19 e Pediatria.

O mesmo vale para listagem de palavras potencialmente difíceis. O método da sugestão de sinônimos pode ser calibrado para um maior ou menor grau de especialidade temática do texto. Um exemplo disso é o item "**abordagem**". No módulo da Doença de Parkinson, temos as seguintes sugestões: **"enfoque, visão, perspectiva".** Mas, na base de dados do módulo de Oncologia, temos um novo *synset*: **"tratamento, orientação, procedimento"**.

Novamente, ressaltamos que a atual composição da base de dados da MedSimples, especialmente no que se refere ao conjunto das definições simplificadas, é um rascunho a ser validado pelos consultores da área da Saúde que apoiam o nosso grupo de pesquisa. Somente com tal aval a informação é tornada pública para o nosso usuário.

Perspectivas do projeto com Oncologia

Cabe agora persistir no trabalho de refino, aprimorando os dados sobre o vocabulário e expandindo a qualificação dos itens na base de dados sobre Oncologia da MedSimples. Estimamos que os bancos de dados coletados serão formados por algo próximo a 800 itens na categoria das terminologias e por outros 350 itens na categoria de palavras potencialmente difíceis. A parte de elaboração de listas, de definições e de sinônimos em *synsets* deve estar concluída até o final de dezembro de 2023. Depois disso, em uma nova cota de auxílio que já conquistamos junto à FAPERGS, até setembro de 2024, passaremos à parte da indexação ou categorização das terminologias reunidas (o que facilitará as buscas do usuário nos glossários anexos à MedSimples) e concluiremos a etapa de avaliação da adequação conceitual dos conteúdos a serem oferecidos, a cargo de profissionais da Saúde.

O estágio atual do projeto é a formalização dos bancos de dados e sua implementação técnica, dentro da ferramenta, por profissionais de Computação. Para tanto, nossas listas devem ser "tratadas", ou seja, todos os itens de vocabulário reunidos devem ser transformados, lematizados e indexados, e preparados para o passo seguinte. Nesse passo, ocorre a aplicação do *parser* PassPort (ZILIO, WILKENS, FAIRON, 2018), uma ferramenta automática de reconhecimento gramatical de cada item conforme sua ocorrência nos textos de estudo. É um *parser*, embutido à ferramenta, que permite o *matching* entre as palavras do texto de entrada, que o usuário quer simplificar, e as listas internas de palavras e definições a sugerir nas bases de dados. Para essa etapa futura, já que estamos avaliando a utilização de outro tipo de *parser*.

Como já dito na introdução, o *parser* funciona como um grande "motor" interno da Ferramenta MedSimples, viabilizando que o vocabulário do texto do usuário seja reconhecido e os itens difíceis sejam apontados e filtrados pelo banco de dados. A imagem a seguir mostra, em forma de esquema, as etapas do processamento pelo qual passará um texto sobre Oncologia selecionado por um usuário.

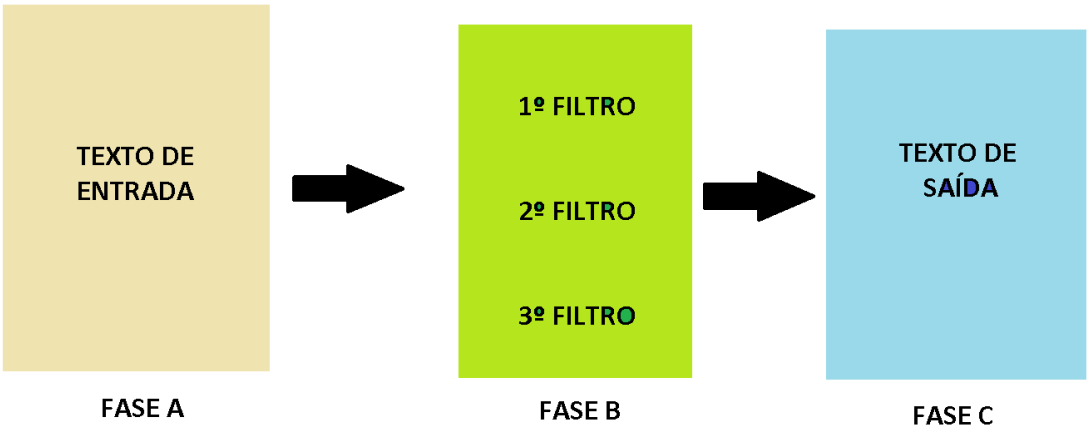


Figura 5. Exemplificação do sistema interno da Ferramenta MedSimples

A fase A representa a entrada do texto escolhido, acompanhada da informação do tema e do tipo de leitor a atender na simplificação. Na fase B, estão as três etapas de filtragem de todas as palavras do texto inserido pelo utilizador: identificação se o item é complexo ou simples; classificação do item entre palavras e terminologias associação com *synsets* ou definições. Por fim, a fase C representa o resultado do processamento, no qual o texto inserido é reapresentado com as sugestões e/ou alertas de provável complexidade em forma de sublinhas e *pop-ups*.

Desafios para pleno funcionamento do módulo de Oncologia

O trecho a seguir corresponde a um possível texto a ser inserido pelo usuário na Ferramenta MedSimples. Depois dele, trazemos a nossa expectativa de um bom processamento no que se refere à eficiente identificação de elementos complexos e alternativas simplificadas:

> Embora o câncer de pele seja o mais frequente no Brasil e corresponda a cerca de 30% de todos os tumores malignos registrados no país, o melanoma representa apenas 3% das neoplasias malignas do órgão. É o tipo mais grave, devido à sua alta possibilidade de provocar metástase (disseminação do câncer para outros órgãos). O prognóstico desse tipo de câncer pode ser considerado bom se detectado em sua fase inicial. Nos últimos anos, houve grande melhora na sobrevida dos pacientes com melanoma, principalmente devido à detecção precoce do tumor e à introdução dos novos medicamentos imunoterápicos.

A seguir, a suposição do processamento do trecho acima, com a seguinte legenda: as marcações em verde representam os itens que seriam marcados pela

Ferramenta MedSimples como terminologias; os itens em azul são as palavras potencialmente difíceis; o símbolo [=>] representa o comando *gerar pop-up* com a sugestão de simplificação ou de sinônimos. Por questões de espaço, a definição simplificada não está apresentada:

> Embora o câncer de pele[=>] seja o mais frequente[=> sin.: comum]no Brasil e corresponda a cerca de 30% de todos os tumores malignos[=>] registrados[=>] no país, o melanoma representa apenas 3% das neoplasias malignas[=>] do órgão[=>]. É o tipo mais grave, devido à sua alta possibilidade de provocar metástase [=>] (disseminação[=>] do câncer para outros órgãos [=>]). O prognóstico[=>][definição simples]] desse tipo de câncer pode ser considerado bom se detectado[=>]sin: descoberto, identificado] em sua fase inicial. Nos últimos anos, houve grande melhora na sobrevida[=>] dos pacientes com melanoma[=>], principalmente devido à detecção=>] precoce [=>] do tumor e à introdução dos novos medicamentos imunoterápicos [=>]

Naturalmente, essas marcações representam um resultado ideal. Por fim, vale dizer que trabalho aqui relatado possibilitou a geração de dados promissores em termos do reconhecimento, via técnicas de Lexicometria (SANROMAN, DOCIO, 2022) de todo um repertório vocabular mais e menos utilizado em diferentes fontes textuais (de palavras que podem ser difíceis, como **detectados**, a terminologias como **neoplasias malignas**). Acreditamos que, na parte do reconhecimento linguístico, já temos um bom conjunto de dados para alimentar a ferramenta e colocá-la em funcionamento no seu novo módulo de Oncologia. Além disso, a revisão dos diferentes registros das outras bases de dados da MedSimples permitiu detectar problemas e projetar possíveis soluções, além das apontadas por Esteves (2023).

Esperamos ter demonstrado que o trabalho linguístico, que dá suporte à implementação computacional, envolve conhecer técnicas e passos do processamento lexical dos textos-fonte. A Lexicometria aplicada aos *corpora* criteriosamente reunidos e o cotejo com as referências de simplicidade vocabular do *CorPoP*, balizados por um critério de frequência simples, têm se mostrado bastante produtivos para apoiar, com sugestões pertinentes, uma escrita simplificada. Conforme se depreende, desde as testagens sobre os conteúdos mais ou menos cobertos pelo *corpus* até a feição de uma definição simples que busca explicar o que é uma **biópsia**, são as pessoas e suas subjetividades que apontam os melhores insumos e alternativas para guiar tal tipo de simplificação. Afinal, somente pessoas conseguem ter empatia verdadeira pelos destinatários finais dos textos, pessoas adultas de escolaridade limitada que muito precisam dessa informação compreensível, enunciada de acordo com suas condições de entendimento e suas experiências de vida.

Agradecimentos

Este trabalho somente foi possível realizar com o apoio da nossa colega Francine Facchin Esteves, cujo mestrado, na UFRGS, ocupou-se, justamente de padrões definição simplificada em temas de Saúde. Nosso obrigada também à FAPERGS e ao CNPq, pelas bolsas concedidas, e à SEDETEC-UFRGS, que segue com o programa de iniciação científica tecnológica em projetos de inovação.

Referências

ANTHONY, Laurence. **AntConc** (Versão 4.2.0) [*Software* de computador]. Tóquio, Japão: Universidade Waseda. Disponível em: https://www.laurenceanthony.net/software Acesso em dez 2023.

BERBER SARDINHA, Tony. **Linguística de *corpus***. Barueri - São Paulo: Manole, 2004.

BONQUEVES FADANELLI, Sabrina. A Terminografia didático-pedagógica e as sequências didáticas no ensino de leitura em ESP. **The ESPecialist**, 39 (1). São Paulo, 2018. https://doi.org/10.23925/2318-7115.2018v39i1a2

CASELI, Helena M.; NUNES, Maria das Graças.V. (org.) **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em dez 2023

ESTEVES, Francine Facchin. Definições acessíveis: por uma linguagem simples em cuidades paliativos. 157 p. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2023.
Disponível em: https://lume.ufrgs.br/handle/10183/258591
Acesso em Dez. 2023.

FINATTO, Maria Jose Bocorny; ESTEVES, Francine Facchin; VILLAR, Guillermo Silva. Construindo uma terminologia de raiz: textos legislativos sob exploração terminológica. PLATÔ – Revista do Instituto Internacional da Língua Portuguesa, n. 9, v. 5, Praia – Cabo Verde. p. 76-97, 2022.
Disponível em: https://iilp.cplp.org/plato/numeros_publicados.html
Acesso em Dez. 2023.

FINATTO, Maria José Bocorny; PARAGUASSU, Liana Braga. (orgs). **Acessibilidade textual e terminológica** - Universidade Federal de Uberlândia, Uberlândia: EDUFU, 2022.
Disponível em:
https://repositorio.ufu.br/bitstream/123456789/35193/1/eClasse_Acessibilidade_Textual.pdf
Acesso em Dez. 2023.

PASQUALINI, Bianca Franco. CorPop: um *corpus* de referência do português popular escrito do Brasil**.** 250 p. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2018.

SANROMÁN, Álvaro Iriarte; DOCÍO, Susana Sotelo. Análise lexicométrica: algumas técnicas aplicadas a entrevistas a visitantes de Santiago de Compostela. In: FEIJÓ, Elias J. Torres; PRADO, Felisa Rodrigues; SANROMÁN, Álvaro Iriarte (Eds.). **Contar o caminho de Santiago: Literatura, discurso(s) e efeitos sociais na comunidade local**. Lisboa: Colibri, 2022. p. 233-260.

ZILIO, Leonardo; WILKENS, Rodrigo; FAIRON, Cédrick. **PassPort: A Dependency Parsing Model for Portuguese.** In: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings.
Disponível em:
https://www.researchgate.net/publication/327223413_PassPort_A_Dependency_Parsing_Model_for_Portuguese_13th_International_Conference_PROPOR_2018_Canela_Brazil_September_24-26_2018_Proceedings
Acesso em Dez 2023.

Submetido em 11 de julho de 2023.
Aprovado em 18 de dezembro de 2023.