CorPop26: Uma atualização das 3.000 Palavras Mais Frequentes do CorPop de 2018

Roges Horacio Grandi[1], Bianca Pasqualini[2], Maria Jose Bocorny Finatto[2], Leandro Krug Wives[1], Raquel Salcedo Gomes[1]

[1]PPGIE – CINTED – UFRGS
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

roges.grandi@gmail.com, wives@ufrgs.br, raquel.salcedo@ufrgs.br

[2]PPG-LETRAS-UFRGS – Campus do Vale - Porto Alegre – RS – Brazil

roges.grandi@gmail.com, leandro.wives@ufrgs.br, raquel.salcedo@ufrgs.br

Abstract. *This study aimed to update the list of the 3,000 most frequent words in Brazilian Portuguese based on the CorPop (2018), composed of texts reflecting the average literacy level of readers in the country. The update methodology consisted of comparing the frequencies from original CorPop with those of two other corpora, one based on the Carolina corpus from USP and the other on Wikipedia, thereby expanding the textual base. As a result, 659 words were replaced, representing 21.97% of the original list. Comparative tests across different text analysis tools — NILC Metrix, Vidya Text with original CorPop, and Vidya Text with CorPop26 — demonstrate the influence of the most frequent word list on the results of the Dale and Chall (1995) equation.*

Resumo. *Este estudo teve como objetivo atualizar a lista das 3.000 palavras mais frequentes do português do Brasil com base no CorPop de 2018, formado por textos baseados no letramento médio dos leitores do país. A metodologia de atualização utilizada foi comparar frequências do CorPop original com as de outros dois corpora, um baseado no Carolina da USP e outro na Wikipédia, ampliando a base textual. Como resultado, obteve-se uma renovação de 659 palavras, representando 21,97% das originais. Testes comparativos em diferentes ferramentas de análise textual (NILC Metrix, Vidya Text com CorPop original e Vidya Text com CorPop26) evidenciam a influência da lista de palavras mais frequentes no resultado da fórmula de Dale e Chall de 1995.*

1. Introdução

A fórmula de leiturabilidade de Dale-Chall é um modelo matemático de predição de dificuldade de leitura de um texto que correlaciona uma variável lexical, a *familiaridade vocabular*, a uma variável sintática, o *comprimento médio* das sentenças. Seu modelo original foi desenvolvido em 1948 por Dale e Chall como uma alternativa ao método proposto por Flesch (1948), que se baseia na variável sintática de comprimento médio das sentenças e na variável lexical de número médio de sílabas por palavra. Ambos os modelos foram originalmente desenvolvidos para o inglês norte-americano [Dale and Chall 1948, Flesch 1948].

Dale e Chall (1995) publicaram uma versão revisada de sua fórmula com o objetivo de aprimorar a acurácia preditiva e a adequação a diferentes níveis de leitura. A principal revisão consistiu na atualização da lista de palavras familiares, que passou a refletir de forma mais fiel o vocabulário efetivamente conhecido por estudantes contemporâneos. A lista original, baseada em dados da década de 1940, foi substituída por uma versão revisada e fundamentada em evidências empíricas mais recentes sobre reconhecimento lexical, o que tornou a identificação de palavras frequentes mais precisa [Chall and Dale 1995].

Além disso, a fórmula revisada de Dale e Chall introduziu um fator de ajuste, que adiciona 3,6365 pontos ao índice de legibilidade quando a proporção de palavras "difíceis"(isto é, não frequentes) no texto excede 5%, com a finalidade de compensar a subestimação da dificuldade de ler textos com alta densidade de vocabulário não familiar. A Equação 1 mostra a revisão de Dale e Chall (1995), onde:

- ***PDW (percentage of difficult words)*** = porcentagem de palavras difíceis (não presentes na lista de palavras frequentes)
- ***ASL (average sentence length)*** = quantidade média de palavras por frase.

Conforme apresentado na Equação 1, o cálculo da fórmula revisada de Dale e Chall requer o uso de uma lista de palavras frequentes. Na versão original de 1948, essa lista era composta por aproximadamente 763 palavras familiares. Na versão revisada de 1995, esse conjunto foi expandido para cerca de 3.000 palavras, com o objetivo de melhorar a estimativa da dificuldade lexical [Chall and Dale 1995]. Desde então, a literatura tem adotado esse valor como referência para a equação [DuBay 2004].

$$\textbf{Equação 1: escore bruto} = 0{,}1579 \cdot PDW + 0{,}0496 \cdot ASL + 3{,}6365 \qquad (1)$$

No Brasil, temos um *corpus* do português popular escrito [Pasqualini 2018] composto por textos oriundos de cinco fontes:

1. textos do projeto PorPopular, principalmente do *Diário Gaúcho*[1], mas também de jornais como *Hora de Santa Catarina*[2], voltados às classes C e D, que retratam o leitor médio brasileiro [UFRGS 2016];
2. textos bastante lidos pelo leitor médio brasileiro, de acordo com a pesquisa *Retratos da Leitura no Brasil*[3];
3. textos da coleção *É Só o Começo*, de adaptações de clássicos da literatura brasileira para leitores com baixo letramento[4].
4. textos do jornal *Boca de Rua*[5], produzido por pessoas em situação de rua, com baixa escolaridade e baixo letramento; e

[1]Página inicial: https://diariogaucho.clicrbs.com.br. Acesso em 18 abr. 2026.

[2]Página inicial: https://www.nsctotal.com.br/veiculos/hora. Acesso em 18 abr. 2026.

[3]Página inicial: https://www.prolivro.org.br/pesquisas-retratos-da-leitura/as-pesquisas-2. Acesso em 18 abr. 2026.

[4]Essa coleção foi criada a partir do projeto de valorização do livro, uma iniciativa do Serviço Social da Indústria (SESI), com incentivos e auxílio da Unesco e do Programa Brasil Alfabetizado, do Ministério da Educação e Cultura (MEC) [Vieira 2010].

[5]Página inicial: https://jornalbocaderua.wordpress.com. Acesso em: 18 abr. 2026.

5. textos do *Diário da Causa Operária*[6], imprensa operária brasileira produzida também por pessoas dentro da faixa média de letramento do país.

A partir desse *corpus*, compilado de acordo com o nível médio de letramento dos brasileiros, [Pasqualini 2018] criou uma lista ordenada de frequência de 9.904 palavras, da qual se pode extrair a lista das 3.000 palavras mais frequentes e usá-la nas equações de Dale e Chall no português do Brasil.

Este estudo propõe uma reordenação das 3.000 palavras mais frequentes do CorPop por meio de uma investigação cruzada com palavras frequentes encontradas em dois *corpora*:

1. ***Corpus* Carol Pop BR**: derivado de uma seleção de textos do *corpus* Carolina[7] da Universidade de São Paulo (USP), do qual se procurou excluir textos de vocabulário específico e textos com origem em Portugal, mais textos do Diário Gaúcho [Grandi 2026a].
2. ***Corpus* Wikipédia PTLimpo**: palavras extraídas de artigos da Wikipédia em português [Grandi 2026b].

Tanto o *Corpus* Carol Pop BR quanto o *Corpus* Wikipédia PTLimpo foram processados para extrair suas palavras mais frequentes, por meio de um processamento automatizado de remoção de tremas, exclusão de sequências de caracteres formadas somente por números romanos e separação de nomes. As palavras restantes foram as selecionadas para computar frequências, das quais foram selecionadas as 10.000 mais frequentes. Por fim, foi realizada uma análise visual dos três *corpora*: CorPop, Carol Pop BR e Wikipédia PTLimpo, para eliminar abreviaturas, letras soltas remanescentes, como "b"e "s", palavras em português não pertencentes à grafia brasileira, como "património", e estrangeirismos não incorporados ao português do Brasil, conforme o Vocabulário Ortográfico Comum da Língua Portuguesa (VOC[8]). Chamamos as palavras remanescentes ao final desse processo de *CarolPopBR10mil* e *WikipédiaPTLimpo10mil*.

A atualização de frequências do CorPop original para formar o CorPop26 foi realizada por meio da leitura, palavra por palavra, do CorPop, em ordem de frequência, das palavras mais frequentes às menos frequentes. Se a palavra lida estivesse simultaneamente no *CarolPopBR10mil* e no *WikipédiaPTLimpo10mil*, a palavra era selecionada para formar a lista de palavras mais frequentes do CorPop26. Essa etapa foi necessária para garantir a evolução do CorPop original, evitando vieses muito intensos do vocabulário jornalístico da imprensa paulista presente no *corpus* Carolina e do vocabulário lusitano também bastante presente no *corpus* original da Wikipédia.

Assim, este artigo foi organizado da seguinte forma: a seção *Fundamentação Teórica* discorre sobre leiturabilidade, frequência de palavras e representatividade de textos em *corpora*. A seção *Metodologia* relata o processo de renovação adotado. A seção *Processo de Criação do CorPop26* detalha a criação dos *corpora* Carol Pop BR e Wikipédia PTLimpo, assim como critérios e os passos de seleção das 3.000 palavras do CorPop26. A seção *Resultados* comenta um teste em que se comparam resultados de Dale-Chall obtidos em diferentes plataformas de análise de leiturabilidade de textos em português do

[6]Página inicial: https://causaoperaria.org.br. Acesso em 18 abr. 2026.
[7]Página inicial: https://sites.usp.br/corpuscarolina. Acesso em 18 abr. 2026.
[8]Página inicial: https://voc.cplp.org/index.php?action=von&csl=br. Acesso em 18 abr. 2026.

Brasil. Concluiu-se que as métricas de Dale-Chall que adotam a lista de 3.000 palavras selecionadas do CorPop26 apresentam bons níveis de acurácia em relação aos níveis de escolaridade estimados por esse método.

2. Fundamentação Teórica

Finatto e Paraguassu (2021) definem leiturabilidade como a potencial facilidade ou dificuldade de leitura de um texto, determinada por fatores linguísticos (referentes a escolhas lexicais e sintáticas) que, por sua vez, estão relacionados ao perfil de leitor pretendido do texto, que pode ser alguém com maior ou menor escolaridade, com mais ou menos conhecimento prévio do assunto do texto [Finatto and Paraguassu 2021], entre outros aspectos. Estimativas de leiturabilidade têm sido tradicionalmente modeladas por meio do uso de variáveis sintáticas e lexicais, conforme proposto por pesquisadores como Flesch[9], Dale, Chall e outros [Flesch 1948, Dale and Chall 1948, Chall and Dale 1995, DuBay 2004]. O domínio das palavras mais frequentes de um idioma é fundamental para a compreensão textual, o que reforça a relevância das listas de frequência e dos modelos de leiturabilidade, como os propostos por Dale e Chall [Dale and Chall 1948, Chall and Dale 1995, Nation 2001].

Em relação à validade de análises textuais, a Linguística de *Corpus* enfatiza que a validade de análises linguísticas depende diretamente da representatividade dos dados utilizados. Biber (1993) argumenta que os *corpora* devem refletir a diversidade de usos da língua, enquanto McEnery e Hardie (2012) destacam que vieses de domínio podem comprometer generalizações linguísticas [McEnery and Hardie 2011]. Nesse sentido, a combinação de vários *corpora* constitui uma estratégia relevante para mitigar vieses e ampliar a cobertura lexical.

3. Metodologia

Idealmente, a construção de listas de frequência lexical deve buscar minimizar vieses amostrais. Entre os critérios fundamentais, destaca-se a abrangência regional do *corpus* utilizado. Assim, no caso do português do Brasil, recomenda-se a adoção de *corpora* de cobertura nacional, constituídos por textos produzidos em português brasileiro, de modo a assegurar maior representatividade linguística.

Para a estimação da leiturabilidade por faixa etária, é igualmente desejável a disponibilidade de textos anotados por nível de ensino, a fim de viabilizar a condução de testes de acurácia. Idealmente, tais dados devem apresentar abrangência nacional, contemplando diferentes regiões do Brasil, de modo a assegurar maior representatividade educacional e linguística. No caso específico da fórmula de Dale e Chall (1995), se a classificação dos textos por nível de ensino não estiver de acordo com a realidade brasileira, é possível, a partir de um conjunto de textos representativos, ajustar os parâmetros da equação, para minimizar a margem de erro. Um desses parâmetros refere-se ao tamanho da lista de palavras frequentes, cujo valor de referência na literatura é de aproximadamente 3.000 vocábulos.

Diante do exposto, na ausência de *corpora* ideais, optou-se por atualizar a lista das 3.000 palavras mais frequentes do CorPop, tomando como base sua compilação de 9.940

[9] [Martins et al. 1996] adaptaram a fórmula de Flesch para o português do Brasil.

ocorrências, devido à representatividade sociolinguística e do nível médio de letramento da população e registro do português popular escrito. Para a atualização, selecionaram-se *corpora* de textos em língua portuguesa, preferencialmente do português do Brasil, buscando-se a maior abrangência possível e evitando-se vocabulários especializados. Esses *corpora* foram utilizados para o cruzamento de frequências, com o objetivo de reordenar as frequências observadas no CorPop original. A estratégia de reordenação consistiu na iteração sobre o léxico do CorPop original, verificando-se, para cada palavra, sua ocorrência nos demais *corpora* selecionados. Quando identificada uma intersecção, o item foi mantido na lista final até que se alcançasse o conjunto de 3.000 palavras.

4. Processo de Criação do CorPop26

Com base na metodologia apresentada, buscaram-se *corpora* abrangentes e bem elaborados para realizar a intersecção com as palavras contabilizadas no CorPop original. Dentre os *corpora* disponíveis para uso acadêmico, foram encontrados o *Brazilian Portuguese Web as Corpus* (BrWaC[10]), o *Corpus* Geral do Português Brasileiro Contemporâneo (Carolina), um *corpus* do *Diário Gaúcho*, um *corpus* da Wikipédia e um *corpus* da Google, o mC4[11].

4.1. *Corpus* BrWac

O BrWac é um *corpus* que foi disponibilizado publicamente pelo *Neurocognition and Natural Language Processing Research Lab* do Instituto de Informática da Universidade Federal do Rio Grande do Sul (UFRGS) em 2017, composto por 3,53 milhões de documentos e 2,68 bilhões de *tokens*. Embora seja bastante extenso, não foi possível extrair adequadamente as palavras desse *corpus*, pois o arquivo apresenta muitos problemas de codificação de caracteres decorrentes da concatenação dos milhões de documentos que o compõem, o que inviabilizou seu uso.

4.2. *Corpus* Carolina

O Carolina é um *corpus* desenvolvido pela Universidade de São Paulo (USP) contendo textos em português coletados de 1970 a 2021 [USP 2023]. Neste estudo, foi utilizada a versão de 2023, que contém mais de dois milhões de textos e 823 milhões de *tokens*. Com a finalidade de minimizar vieses, os *corpora* e *datasets* do Carolina listados na Tabela 1 não foram utilizados.

Tabela 1. Análise de Corpora e Datasets Desconsiderados

Corpus / Dataset	Características	Motivo
Brazilian Stock Market Tweets with Emotions	*Corpus* para análise de emoções em *tweets* sobre a Bolsa de Valores de São Paulo em 2018 [UNICAMP 2018].	Terminologia especializada do mercado financeiro.

[10]Página inicial: https://huggingface.co/datasets/UFRGS/brwac. Acesso em 19 abr 2026.
[11]Página inicial: https://aaas.blog/dataset/mc4-dataset. Acesso em 27 abr. 2026.

Corpus / Dataset	Características	Motivo
Datasets of Neuropsychological Language Tests in Brazilian Portuguese	*Datasets* de neuropsicologia elaborados da Pontifícia Universidade Católica do Rio Grande do Sul.	Terminologia especializada da área médica [Casanova et al. 2025].
HAREM	*Corpora* mantidos pela Linguateca[12].	Forte influência do português europeu [Santos et al. 2015].
Leg2Kids	*Corpus* de 36.413 legendas de filmes e séries dos gêneros "família"e "animação".	Vocabulário direcionado a crianças [Hartmann 2019].
Portuguese Tweets for Sentiment Analysis	*Dataset* de 235,3 MB de *tweets* voltados às análises de polaridade de sentimentos.	Muitos dos *tweets* têm erros de grafia, abreviaturas e jargões de internet.
Revista Pesquisa FAPESP *Parallel Corpora*	Link informado no Carolina[13] está quebrado.	Impossibilidade de acesso.
Judiciário	*Corpora* do STF com 38.068 textos.	Vocabulário especializado do Direito.
Legislativo	*Corpora* do Senado brasileiro constituídos por 14 textos.	Vocabulário especializado do Poder Legislativo.
Social Media	O link informado no Carolina[14] leva a uma página sem opções para *download*.	Impossibilidade de acesso.
Wikis	*Corpora* formados por 960.139 textos da Fundação Wikimedia.	O *corpus* completo de artigos da Wikipédia em português utilizado neste estudo tornou esta coleção desnecessária.

4.3. *Corpus* Bruto do *Diário Gaúcho*

Também foi utilizado o *corpus* bruto do *Diário Gaúcho*, de 2008, disponibilizado pelo projeto PorPopular da UFRGS [UFRGS 2016]. Esse *corpus* reúne textos coletados ao longo de 116 dias daquele ano, totalizando 2,89 MB. Nesse mesmo *site*, há um arquivo de 2010 com extrações de textos do mesmo periódico. Todavia, esse arquivo não pode ser aproveitado devido a problemas de codificação de caracteres.

4.4. Formação do Corpus Carol Pop BR

O *Corpus* Carol Pop BR [Grandi 2026a] foi formado pela unificação dos textos selecionados do Carolina aos textos selecionados do *Diário Gaúcho*, conforme descrito anteriormente. Após a concatenação dos textos em um *corpus* unificado com 510 MB e 90,1 milhões de *tokens*, foi executada uma rotina de remoção de tremas.

[12]Centro de pesquisa e desenvolvimento, repositório de *corpora*, de ferramentas e recursos linguísticos e um polo de articulação científica entre pesquisadores de países lusófonos.

[13]Página: http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm. Acesso em 19 abr. 2026.

[14]Página inicial: https://www.wattpad.com. Acesso em 19 abr. 2026.

4.5. Formação do Corpus Wikipédia PTLimpo

Além da compilação de textos que formou o Corpus Carol Pop BR, buscou-se uma coleção de textos com foco em abrangência. As opções encontradas foram:

1. a coleção de todos os artigos da Wikipédia em português, disponibilizada em seu repositório oficial[15], com mais de um milhão de artigos, os quais ocupam 7,2 GB após descompactados; e
2. o extrato em língua portuguesa do *Multilingual Colossal Clean Crawled Corpus* (mC4), um *corpus* textual multilíngue de 27 TB desenvolvido pela Google para treinar modelos de linguagem. O extrato em português do mC4 tem aproximadamente 524 GB.

Analisadas as alternativas, optou-se pela utilização de artigos da Wikipédia, uma vez que esses textos passam por revisão colaborativa contínua, com tendência à clareza e à padronização linguística, ainda que apresentem variação quanto ao nível de complexidade. A extração de palavras do Corpus Wikipédia PTLimpo foi realizada em Python, com o módulo Gensim da biblioteca Wikicorpus[16]. Para essa extração, letras maiúsculas e minúsculas foram preservadas, bem como palavras compostas por uma única letra. Finalizando o tratamento, tanto no Corpus Carol Pop BR como no Corpus Wikipédia PTLimpo foi executada uma rotina de remoção de tremas.

4.6. Extração das 10 mil Palavras mais Frequentes dos *Corpora* Carol Pop BR e Wikipédia PTLimpo

Os *corpora* Carol Pop BR e Wikipédia PTLimpo passaram pela mesma sequência de linhas de comando em um sistema operacional Linux para extrair suas 10 mil palavras mais frequentes.

4.6.1. Extração de Palavras

Para extrair as palavras dos *corpora*, foi executado o seguinte comando, onde `$CORPUS_FILE` é o *corpus* Carol Pop BR ou o Wikipédia PTLimpo. Foi usada uma expressão regular no comando `grep` para extrair todas as sequências contínuas de caracteres alfabéticos do *corpus*, retornando cada ocorrência como um *token* individual em uma nova linha. Em seguida, esses *tokens* foram direcionados por meio de um pipe para o comando `sort`, que ordenou os termos alfabeticamente. Por fim, o resultado foi armazenado no arquivo `$ORDENADAS`:

```
grep -oE
    "[[:alpha:]]
    "$CORPUS_FILE" | sort --parallel=4 > "$ORDENADAS"
```

Para liberar espaço, o `$CORPUS_FILE` foi removido. A etapa seguinte foi aplicar o comando `uniq` com a opção -c sobre os arquivos previamente ordenados de *tokens*, com o objetivo de agrupar ocorrências consecutivas idênticas e contabilizar suas frequências. Desse modo, para cada palavra, foi gerada uma linha contendo o número de ocorrências seguido do respectivo *token*:

[15] Página inicial: https://dumps.wikimedia.org/ptwiki/latest. Acesso em 19 abr. 2026.

[16] Página inicial: https://github.com/piskvorky/gensim/tree/develop/gensim. Acesso em 19 abr. 2026.

```
uniq -c "$ORDENADAS" > "$CONTADAS"
```

A próxima etapa foi criar dois arquivos, o `$COMUNS_FREQ` e o `$NOMES_FREQ`, por meio do comando `awk`, tendo como parâmetro de entrada o arquivo `$CONTADAS` de cada *corpus*. Inicialmente, cada palavra foi normalizada para minúsculas para agrupar variações de capitalização sob a mesma chave, assim como suas frequências de ocorrência. Em seguida, foram descartadas palavras com frequência inferior a 20 e calculou-se a proporção de usos com letra maiúscula. Palavras cuja escrita com inicial maiúscula era superior a 95% foram gravadas no arquivo `$NOMES_FREQ`. As demais foram consideradas comuns e armazenadas no arquivo `$COMUNS_FREQ`. Por fim, foi utilizado o comando *head* para se obter as 10 mil palavras mais frequentes dos respectivos arquivos `$COMUNS_FREQ`:

```
head -n 10000 "$COMUNS_FREQ" > "$TOP10K"
```

O último trabalho realizado nessa etapa foi remover visualmente, não apenas nos *corpora* Carol Pop BR e Wikipédia PTLimpo, mas também na lista de frequências do CorPop original, abreviaturas, letras soltas, como "b"e "s", estrangeirismos e palavras cuja grafia não corresponde à do português do Brasil, conforme o VOC.

4.7. Finalização da Criação do CorPop26

A última etapa da criação do CorPop26 consistiu em uma análise comparativa entre os vocábulos do CorPop original e os vocábulos dos *corpora* Carol Pop BR e Wikipédia PTLimpo. O CorPop original continuou sendo a base, como afirmamos anteriormente, por sua representatividade sociolinguística e registro do português popular escrito. Observa-se que o *corpus* Carol Pop BR, por ser composto predominantemente por textos do *Jornal Folha de São Paulo* (97,2%), tem um forte viés do vocabulário jornalístico paulista. Os artigos em língua portuguesa da Wikipédia, por sua vez, componentes do *corpus* Wikipédia PTLimpo, apesar de serem escritos, em sua maioria, por brasileiros (em torno de 78%), ainda têm um viés considerável do português lusitano.

Isto posto, uma intersecção de palavras dos três *corpora* ajudou a diminuir os vieses dos *corpora* Carol Pop BR e Wikipédia PTLimpo. O resultado foi o CorPop26, um conjunto das 3.000 palavras mais frequentes do português do Brasil, com base no CorPop original, mas influenciado pelos *corpora* Carol Pop BR e Wikipédia PTLimpo para uma ordenação que reflita uma abrangência maior, tanto textual como regional. A atualização das palavras mais frequentes resultou na substituição de 659 vocábulos, correspondendo a 21,97% do conjunto original de 3.000 itens[17].

5. Resultados

Para avaliar o impacto da atualização do vocabulário, foram realizados testes comparativos com três configurações distintas da fórmula de Dale-Chall (1995). Com base na pontuação bruta resultante da aplicação da equação, é possível estimar tanto o número de anos de escolarização (contados a partir do primeiro ano do Ensino Fundamental) quanto o respectivo nível de escolaridade, conforme apresentado na Tabela 2 [Paraguassu 2024, p. 174].

[17] A lista das palavras substituídas está disponível em [Grandi and Pasqualini 2026].

Tabela 2. Interpretação de Dale-Chall (1995) por nível de escolaridade

Pontuação	Anos	Escolaridade
< 5	≤ 4	Nível 1: até o 4º ano do Ensino Fundamental
5 a < 6	5 a 6	Nível 2: do 5º ao 6º ano do Ensino Fundamental
6 a < 7	7 a 8	Nível 3: do 7º ao 8º ano do Ensino Fundamental
7 a < 8	9 a 10	Nível 4: do 9º ano do Ensino Fundamental ao 1º ano do Ensino Médio
8 a < 9	11 a 12	Nível 5: do 2º ano do Ensino Médio aos primeiros anos do Superior
9 a < 10	13 a 14	Nível 6: Ensino Superior
≥ 10	≥ 15	Nível 7: Ensino Superior e Pós-graduação

A Tabela 3 mostra um comparativo entre os resultados obtidos utilizando-se o sistema NILC-Metrix [Leal et al. 2023] e a Vidya Text [Grandi et al. 2026] em duas configurações: uma com o vocabulário do CorPop e outra com o vocabulário do CorPop26. Foram selecionados cinco trechos de caráter institucional, produzidos por órgãos de saúde e cidadania e voltados à comunicação direta com o cidadão. Conforme detalhado nessa Tabela, embora as ferramentas usem a mesma equação [Chall and Dale 1995], os resultados indicam que as variações nas listas de palavras mais frequentes influenciam a métrica final. Essa sensibilidade demonstra que a escolha da base de referência é uma variável importante e pode modificar significativamente os índices de legibilidade e a classificação do nível de complexidade dos textos. A interpretação da pontuação dos textos analisados na Tabela 3 pode ser consultada na Tabela 2.

A atualização lexical é refletida especialmente nos resultados das amostras 2 e 3, em que houve reclassificação do nível dos textos. Essa mudança ocorreu porque termos como "oportunidade", "atendimento"e "respeito"(em azul) passaram a constar do CorPop26. Vê-se, assim, que a nova base de referência oferece uma estimativa de leiturabilidade voltada à realidade do português brasileiro contemporâneo. Na amostra 5, observa-se uma divergência significativa de 2,3 pontos entre os resultados: o NILC-Metrix classificou o texto como Nível 5 (8,5 pontos), ao passo que a Vidya Text o classificou como Nível 3 (6,2 pontos). Essa diferença exemplifica o impacto direto da escolha da base textual na sensibilidade da ferramenta. Enquanto o NILC-Metrix tende a elevar os índices de dificuldade de itens lexicais que, embora de fácil compreensão intuitiva (como "decisões"e "apoiada", em vermelho), não estão em sua base de referência, o CorPop26 reflete uma seleção lexical mais próxima da realidade do letramento brasileiro atual, como mencionado anteriormente.

6. Considerações Finais e Trabalhos Futuros

Diante do objetivo de atualizar as listas das 3.000 palavras mais frequentes do CorPop original de 2018, tendo como metodologia a ampliação da base textual e mantendo o CorPop como *corpus* de referência (e considerando-se que ele é um bom indicativo do nível de letramento médio dos brasileiros), obteve-se uma renovação de 659 (21,97%) das palavras selecionadas. Os testes com a fórmula de Dale-Chall (1995), em três configurações (NILC-Metrix, Vidya Text com o CorPop e Vidya Text com CorPop26), indicam que a lista de frequências influencia o resultado. A ampliação da base textual usada para renovar a lista de palavras mais frequentes, utilizada neste estudo, contribui para sua representa-

Tabela 3. Comparativo de Índices de Legibilidade por Ferramenta

Textos de amostra	FC1	FC2	NILC
(1) O cultivo de bons hábitos alimentares, a prática regular de exercícios físicos, a boa qualidade do sono e o acompanhamento médico periódico são fatores que atuam em conjunto para preservar o bem-estar físico e mental.	Nv: 6 (10,8 pts)	Nv: 6 (9,4 pts)	Nv: 6 (9,89 pts)
(2) A vacinação é essencial para manter a criança saudável. Durante a triagem neonatal, deve-se aproveitar a oportunidade para avaliar se a criança recebeu as vacinas indicadas ao nascer ou se há necessidade de atualizá-las.	Nv: 6 (9,1 pts)	Nv: 5 (8,2 pts)	Nv: 6 (10,05 pts)
(3) O SUS oferece cuidado em saúde para todas as faixas etárias, incluindo exames preventivos, planejamento familiar, atendimento pré-natal e parto. Você tem direito a ser tratado com respeito em todos os serviços.	Nv: 5 (8,4 pts)	Nv: 4 (7,6 pts)	Nv: 6 (9,36 pts)
(4) O Programa Nacional de Saneamento Rural estabelece diretrizes para ações em áreas rurais, visando à universalização do acesso. Entretanto, pouco foi implementado com base em sua orientação para garantir os direitos humanos.	Nv: 5 (8,4 pts)	Nv: 5 (8,4 pts)	Nv: 6 (10,35 pts)
(5) Se a pessoa consegue tomar decisões, dizer o que quer e o que não quer, ela não precisa de um curador (alguém que decide no seu lugar); ela precisa apenas de ajuda. Para isso, a Tomada de Decisão Apoiada é suficiente.	Nv: 3 (6,2 pts)	Nv: 3 (6,2 pts)	Nv: 5 (8,50 pts)

Legenda: FC1 = Vidya Text com CorPop; FC2 = Vidya Text com CorPop26; Nv = nível; pts = pontos. Nível 6: dos últimos anos do Ensino Superior à Pós-graduação; Nível 5: do 2º ano do Ensino Médio aos primeiros anos do Ensino Superior; Nível 3: do 7º ao 8º ano do Ensino Fundamental. Palavras em azul constam no CorPop26 (FC2); palavras em vermelho permanecem classificadas como complexas.

Fontes das amostras: (1) [Hospital Português 2021]; (2) [Brasil. Ministério da Saúde 2023]; (3) [Brasil. Ministério da Saúde 2024]; (4) [Aguiar et al. 2025]; (5) [Ceará. Defensoria Pública do Estado 2024]

tividade. Ressalta-se, no entanto, a dificuldade de se obter uma "lista definitiva"frente às centenas de milhares de palavras que compõem a língua portuguesa no Brasil, conforme registra o Vocabulário Ortográfico Comum da Língua Portuguesa (VOC).

Os trabalhos futuros, em curto e médio prazo, incluem a realização de outros testes com o método de Dale-Chall (1995), em diferentes contextos. Sendo a língua dinâmica, e considerando-se a diversidade de *corpora* existentes, certamente outros *corpora* poderão gerar listas de frequências que reflitam com mais fidelidade o português escrito no cotidiano brasileiro, desde que tenham quantidade e diversidade de tipos textuais empregados no Brasil e que sejam tomados os devidos cuidados para minimizar vieses.

7. Agradecimentos

Esta pesquisa foi apoiada pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), processo PQ 307088/2023-5, pelos Institutos Nacionais de Ciência e Tecnologia – CNPq-TILD-IAR (processo 408490/2024-1), pela PROBIC FAPERGS-UFRGS e pela FAPERGS Ed. 06/2025 – PROGRAMA PESQUISADOR GAÚCHO –

PqG (processo 25/2551-0002702).

8. Declaração de Ética

Este estudo utilizou dados linguísticos de domínio público e corpora acadêmicos de prestígio (CorPop, *Corpus* Carolina e Wikipédia). Não houve envolvimento de seres humanos, de prontuários médicos ou de dados pessoais sensíveis no desenvolvimento das listas de palavras nem nos testes comparativos. A pesquisa segue os padrões éticos da Linguística Computacional, buscando melhorar ferramentas de acessibilidade na comunicação em saúde.

9. Limitações

A principal limitação deste estudo é o dinamismo da língua portuguesa; qualquer lista fixa de palavras frequentes necessita de atualizações regulares para manter sua relevância. Apesar do uso de três *corpora* diferentes ter ajudado a reduzir vieses regionais e estilísticos, o CorPop26 criado é uma base de linguagem geral e pode precisar de ajustes adicionais ao ser usado em áreas técnicas muito específicas.

10. Declaração sobre o Uso de IA Generativa

Em conformidade com a Portaria CNPq 2.664/2026, a versão 0.38.2 do Gemini CLI foi usada para o refinamento estilístico, organização estrutural e processamento de contagens lexicais específicas do artigo durante a preparação deste manuscrito. Nenhuma ferramenta de IA foi utilizada para gerar os dados de frequência, nem para a metodologia central de atualização da lista de palavras.

Referências

Aguiar, A. M. S. et al. (2025). *Experiências sobre água e saneamento rural na América Latina e no Caribe*. Instituto René Rachou, Fiocruz Minas, Belo Horizonte. Amostra de texto 4.

Brasil. Ministério da Saúde (2023). *Caderneta da criança: passaporte da cidadania (menina)*. Ministério da Saúde, Brasília, 6 edition. Amostra de texto 2.

Brasil. Ministério da Saúde (2024). *Conheça seus direitos no SUS! (Guia para migrantes e refugiados)*. Ministério da Saúde, Brasília. Amostra de texto 3.

Casanova, E., Treviso, M., Hübner, L. C., and Mansur, L. L. (2025). Dnlt-bp: Dataset para avaliação de complexidade textual em português brasileiro. `https://github.com/nilc-nlp/DNLT-BP`. Dataset. Acesso em: 5 jan. 2026.

Ceará. Defensoria Pública do Estado (2024). *Cidadania sem Barreiras: curatela e tomada de decisão apoiada*. Defensoria Pública do Estado do Ceará, Fortaleza. Amostra de texto 5.

Chall, J. S. and Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, pages 37–54.

DuBay, W. H. (2004). *The Principles of Readability*. Impact Information, Costa Mesa.

Finatto, M. J. B. and Paraguassu, L. B. (2021). *Acessibilidade textual e terminológica*. EDUFU.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Grandi, R. H. (2026a). Corpus Carol Pop BR. Kaggle. Dataset.

Grandi, R. H. (2026b). Corpus wikipédia PT limpo. Kaggle. Dataset.

Grandi, R. H. et al. (2026). Vidya text: an educational tool to support the teaching of writing and the study of brazilian portuguese vocabulary. https://vidyatext.nuvem.ufrgs.br. Acesso em: 19 jul. 2026.

Grandi, R. H. and Pasqualini, B. (2026). CorPop 2018/2026. Kaggle. Dataset.

Hartmann, N. S. (2019). Leg2kids. `https://sites.google.com/view/nilc-usp/leg2kids`. Corpus. Acesso em: 5 jan. 2026.

Hospital Português (2021). *Check-up da pessoa idosa: acompanhamento médico e exames de rotina*. Salvador. Amostra de texto 1. Informativo institucional.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*.

Martins, T. B. F. et al. (1996). Readability formulas applied to textbooks in brazilian portuguese. Technical report, ICMC-USP.

McEnery, T. and Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge.

Paraguassu, L. B. (2024). *Linguagem Simples na Saúde: contribuições dos estudos da linguagem para o Letramento em Saúde no Brasil*. Tese (doutorado em letras), Universidade Federal do Rio Grande do Sul, Porto Alegre.

Pasqualini, B. (2018). Corpop: Um corpus do português popular escrito. Tese, Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Santos, D., Cardoso, N., Sarmento, Luís e Rocha, P., and Vieira, R. (2015). Harem. `https://www.linguateca.pt/HAREM`. Corpus. Lisboa: Linguateca. Acesso em: 5 jan. 2026.

UFRGS (2016). Porpopular: Léxico do português popular brasileiro. `http://www.ufrgs.br/textecc/porlexbras/porpopular/index.php`. Acesso em: 18 abr. 2026.

UNICAMP (2018). Brazilian stock market tweets with emotions. `https://www.kaggle.com/datasets/fernandojvdasilva/stock-tweets-ptbr-emotions`. Dataset. Acesso em: 5 jan. 2026.

USP (2023). Carolina: Corpus geral do português brasileiro contemporâneo. `https://sites.usp.br/corpuscarolina/repositorios-2023`. Corpus. Acesso em: 5 jan. 2025.

Vieira, G. d. O. (2010). Adaptação para novos leitores: como a literatura clássica adaptada fornecida às escolas do ensino público e utilizada pelos professores no processo de ensino estimula a leitura de obras originais.