Novo CorPop Saúde: Data Collection for Simplified Health Communication in a Brazilian Patient Information Leaflet *Corpus*

Bianca Pasqualini[1,2], **Maria J. B. Finatto**[1,2], **Eduarda Bertotto**[1]
Paula Salem Carpio[2], **Roges Horacio Grandi**[3], **Leandro Krug Wives**[3]

[1] Fundação de Amparo à Pesquisa Estado do Rio Grande do Sul (FAPERGS)
Porto Alegre – RS – Brazil

[2] PPG-LETRAS – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[3] PPGIE – CINTED – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil
{bianca.pasqualini, eduarda.bertotto, mfinatto, roges.grandi,
leandro.wives, paula.carpio}@ufrgs.br

Abstract. *This paper presents exploratory results from the "Novo CorPop Saúde" project, evaluating whether validation by expert linguists improves readability levels compared to AI-generated simplifications. We analyzed original, AI-generated, and human-validated versions of thirty Patient Information leaflets. Using the Flesch Reading Ease Index, two adapted Dale-Chall metrics, and lexicometric analysis, we assessed automatic simplification against human strategies. Results indicate that while AI improves readability scores primarily through text compression, expert linguists prioritize elaborative strategies, maintaining key terminology while providing didactic explanations to fill semantic gaps. This suggests that human intervention plays a key role in balancing terminological precision with communicative accessibility, a nuance that purely reductive AI strategies might overlook. This study provided evidence on the use of the complex and simplified lexical repertoire of medicine leaflets for the "Novo CorPop Saúde".*

1. Introduction

The Novo CorPop Saúde (NCS)[1] is a research and innovation project in Lexicon and Terminology Studies based on Textual and Terminological Accessibility (TTA) and Natural Language Processing (NLP). Its objective is to expand the existing CorPop [Pasqualini 2018][2] by incorporating texts from Health Sciences and Pharmacology to provide resources for simplifying health information to match the country's average literacy

[1]Project conducted at UFRGS. Granted by FAPERGS, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável (INCT TILD-IAR) and supported by healthcare professionals.

[2]Available at https://www.ufrgs.br/textecc/porlexbras/corpop/index.php.

level.[3] The original CorPop is a reference for written popular Brazilian Portuguese, based on the population's literacy level, comprising diverse genres such as newspapers, literature, and informational brochures, and is used to assess general linguistic accessibility.

The NCS research complies with the Brazilian legal framework, specifically Law 15.263 on plain language[4] and Law 14.874[5], which establish clear communication as a legal obligation and linguistic accessibility as a universal design principle. By affirming the citizen's right to understand official communications, this legal landscape positions accessible health information as a basic right intended to protect patient autonomy and citizenship from technical complexity, especially considering the country's low literacy levels. Importantly, 54.1% of Brazilians struggle to comprehend Patient Information leaflets (PILs), which causes 25% of users to stop their medication because of confusion, despite the fact that 98.4% of the population considers PILs necessary and 59.6% report reading them regularly [Dal Pizzol et al. 2019].

The NCS will include diverse written health materials from official sources: PILs ("bulas de medicamentos"), official educational materials (health booklets, guidelines, and texts directed at the general public produced by Anvisa), informed consent forms ("termos de consentimento livre e esclarecido"), and AI-generated content (a curated sample of health-related texts for laypeople produced by generative AI, such as ParkinsonBot [Campos 2025]).

The project is structured into three main phases: Phase 1 (evaluation and validation of AI simplifications), Phase 2 (healthcare professionals' reviews for scientific and clinical accuracy) and Phase 3 (reading tests with patients to assess real-world impact). In this initial phase (Phase 1), the work is focused on a *corpus* of PILs of Summaries of Product Information for the most consumed drugs nationwide. We analyzed the sample in [Scalercio et al. 2026], which includes all hypertension drugs from the 2025 Programa Farmácia Popular do Brasil[6], as detailed in the next sections.

The analytical approach of this study is grounded in research bridging Textual and Terminological Accessibility (TTA), Health Literacy (HL), and Natural Language Processing (NLP). [Finatto and Tcacenco 2021] define text simplification as intralingual translation, showing that accessibility requires elaborative strategies, such as explanatory paraphrasing, rather than simple reductive summarization. [Finatto and Paraguassu 2022] apply this to the Brazilian context, where high rates of functional illiteracy require the mediation of medical terminology through familiar vocabulary to support patient autonomy. To evaluate these transformations, [Menezes et al. 2023] used NLP for Automatic Readability Assessment (ARA), noting that traditional formulas fail to capture the didactic expansions of human curation. Recently, [Wives and Finatto 2025] and [Scalercio et al. 2025] explored Large Language Models (LLMs) for simplification and conceptual modeling, showing that Generative AI can generate accessible versions, but

[3]According to INAF, 58% of the Brazilian population has rudimentary to elementary literacy levels, creating a significant demand for accessible materials. Available at: https://alfabetismofuncional.org.br/alfabetismo-no-brasil/.

[4]Law 15.263, November 14th 2025.

[5]Law 14.874, May 28th 2024, regulates clinical research involving human subjects and establishes the national system of ethics in research with human beings in Brazil

[6]https://www.gov.br/saude/pt-br/composicao/sectics/farmacia-popular

expert validation is essential for accuracy. It's worth mentioning [Basu et al. 2023] as well, who dealt specifically with controllable simplification of medical texts.

2. Methodology and analysis

The methodology is based on the comparative analysis of 30 hypertension PILs. We established three parallel *corpora* along with a reference *corpus*. The first *corpus*, C1 (original PILs), consists of the original texts provided by the manufacturers. The second, S1 (AI simplification), consists of versions automatically simplified by AI. The third, S2 (human validation), consists of versions reviewed by linguists. Additionally, the CorPop [Pasqualini 2018] was used as the baseline for accessible Brazilian Portuguese vocabulary (Table 1).

Table 1. Description and lexical characterization of the parallel *corpora*.

***Corpus*/set**	**Tokens**	**Types**	**TTR**
CorPop (reference)	688,788	32,311	0.0469
C1 (original PILs)	77,306	3,116	0.0403
S1 (AI simplification)	41,991	2,279	0.0543
S2 (human validation)	48,532	2,513	0.0518

The dataset S1 used ChatGPT (version 5.2) to process the original PILs, guided by a specific set of prompt instructions to: (1) replace technical terminology with simpler synonyms, keeping the original term in parentheses only when unavoidable; (2) convert passive voice constructions into the active voice; (3) split lengthy sentences into shorter, more direct ones; (4) omit irrelevant or excessive details; (5) resolve anaphoras and ambiguous pronouns; and (6) strictly preserve all numerical data, such as dosages, frequencies, and routes of administration[7]. The automatic simplifications (S1) were then refined by three linguists specialized in plain-language techniques to determine validated versions (S2) [Scalercio et al. 2026]. The human validation (S2) was done by linguists who had access to both C1 and S1.

The Flesch Reading Ease Index was calculated using the Vidya Text Tool[8], based on the formula adapted for Brazilian Portuguese [Martins et al. 1996]. In this scale, higher scores indicate lower textual complexity, with values above 75 (easy) being desirable for readers with low literacy. As shown in Table 2, the original PILs (C1) present a very difficult baseline (22.4). While AI simplification (S1) improves this score to 40.9, the human-validated versions (S2) achieve the highest overall readability score (42.8). Despite the improvement, both simplified versions remain within the "difficult" range (equivalent to "university level" of complexity; see Table 2).

To complement the analysis, we used two variations of the Dale-Chall formula [Chall and Dale 1995] adapted for Brazilian Portuguese. The first, Vidya-Dale-Chall (Vidya-DC), uses the technological validation from the Vidya Text Tool. The second, CorPop-Dale-Chall (CorPoP-DC), uses the 3,000 most frequent words from CorPop as its reference for easy words (excluding grammatical words). The results show that S2

[7]This repository contains the code and the resources used: https://github.com/scalercio/med-simple-docs; [Scalercio et al. 2026].

[8]Available at http://vidyatext.nuvem.ufrgs.br/.

Table 2. Flesch Reading Ease Index metrics across the parallel *corpora*.

Corpus/set	Flesch	Interpretation	Vidya-DC	CorPop-DC	Interpretation
C1 (orig. PILs)	22.4	Very difficult	10.08	10.61	Post-graduate level (very difficult)
S1 (AI simp.)	40.9	Difficult	9.63	9.93	University level (difficult)
S2 (hum. valid.)	42.8	Difficult	9.34	9.71	University level (low)

consistently achieves the most accessible scores in both metrics (Table 2). This assessment suggests that human elaborative strategies tend to improve lexical accessibility, as indicated by the consistency across these metrics.

Table 3. Consolidated lexical analysis: frequency, ratio and difference across parallel *corpora*.

Word	S1 (%)	S2 (%)	CorPop (%)	Ratio (S2/CP)	Diff. (S2-S1)
médico	1.329%	1.263%	0.029%	43.9	-0.066%
pressão*	0.867%	0.938%	0.006%	161.4	+0.071%
você	0.914%	0.870%	0.420%	2.1	-0.045%
dose	0.953%	0.870%	0.001%	1197.8	-0.083%
medicamento	0.779%	0.826%	0.002%	569.1	+0.048%
coração*	0.562%	0.777%	0.052%	14.8	+0.215%
sangue*	0.572%	0.680%	0.011%	60.0	+0.108%
usar	0.564%	0.525%	0.014%	38.5	-0.039%
dia	0.641%	0.556%	0.172%	3.2	-0.084%
rins	0.367%	0.451%	0.001%	444.0	+0.085%
tratamento	0.398%	0.398%	0.020%	19.8	0.000%
problemas	0.181%	0.383%	0.049%	7.9	+0.202%
aumentar	0.441%	0.394%	0.007%	54.2	-0.047%
potássio	0.393%	0.354%	0.000%	—	-0.039%
remédios	0.324%	0.330%	0.004%	90.8	+0.006%

*Usage predominantly metaphorical in CorPop.

To evaluate the lexical choices made during the simplification process, we compared the relative frequencies of key terms (excluding grammatical words and multi-word expressions, making sure that the analysis captures the primary lexical shifts) across S1, S2, and CorPop. Table 3 presents a consolidated analysis of the fifteen most frequent content words, which were selected by calculating the relative frequency of unigrams in the S2 dataset, along with relevant terminological units (such as “rins” and “potássio”). By cross-referencing these high-frequency tokens across the parallel *corpora*, we can observe how human curation adjusts word distribution to match medical terminology with the accessible vocabulary patterns of the CorPop baseline.

The data indicate that the human elaborative strategy (S2) increases the frequency of concrete, familiar terms (such as “coração”, “sangue”, “problemas”, and “remédios”) compared to the AI output (S1). Notably, the word “problemas” is used more than twice as frequently in S2 as in S1, a shift of +0.202% that underscores the human preference for high-coverage lexical choices over more specialized words. These terms also show a stronger baseline presence in CorPop, confirming their accessibility to lay readers. In contrast, the linguists tended to reduce the frequency of certain technical or prescriptive terms, like “dose”, “usar”, and “aumentar”, which were more prominent in the AI output.

The qualitative impact of the elaborative strategy is best observed through direct

comparisons of specific sentences. Table 4 provides eight representative examples that highlight the differences between reductive automatic methods and human validation.

Table 4. Microstructural analysis: sentence transformation examples organized by NCS layers.

#	C1 (original PILs)	S1 (AI simplification)	S2 (human validation)
	LEXICAL LAYER		
1	O atenolol é indicado para o controle da hipertensão arterial.	O atenolol serve para controlar a pressão alta.	O atenolol serve para controlar a pressão alta.
2	Este fármaco é excretado primordialmente por via renal.	O remédio sai do corpo pela urina.	O remédio sai do corpo principalmente através da urina (pelos rins).
3	A eficácia deste medicamento depende da posologia correta.	O remédio só funciona se tomar direito.	Para o remédio fazer efeito, você deve tomar as doses nos horários certos.
	DISCURSIVE LAYER		
4	O fármaco atua bloqueando os receptores beta-adrenérgicos.	O remédio age no coração diminuindo a pressão.	O remédio ajuda o coração a bater mais devagar e com menos força.
5	Não tome duas doses ao mesmo tempo.	Não tome duas doses juntas.	Nunca tome duas doses ao mesmo tempo para compensar a que esqueceu.
	PRAGMATIC LAYER		
6	A pressão arterial deve ser aferida regularmente.	Meça a pressão sempre.	Por isso, você deve medir sua pressão regularmente.
7	Podem ocorrer tonturas, fadiga e cefaleia.	Pode causar tontura, cansaço e dor de cabeça.	Você pode sentir tontura, cansaço e dor de cabeça.
8	Pacientes com disfunção renal requerem ajuste posológico.	Quem tem problema nos rins precisa de doses menores.	Pessoas com problemas nos rins podem precisar de doses diferentes.

While AI (S1) focuses on lexical substitution, the human strategy (S2) incorporates discursive and pragmatic layers to help provide clearer comprehension. For instance, in example 5, the human strategy provides meaning to a semantically hollow AI construction ("pressão que cai") by describing the actual physical sensation ("tontura ou vista escura"). In example 6, the addition of the justification "para compensar a que esqueceu" and the emphatic "nunca" provides a more complete functional conceptualization for the reader, helping to avoid confusion. Similarly, the shift from impersonal third-person structures (e.g., "pacientes" and "podem ocorrer") to direct engagement ("você pode sentir") in examples 8 and 9 reflects the strategy of establishing communicative proximity. This approach supports the reader's understanding by personalizing the health information and making the content more accessible.

The strategies shown in these layers of intralingual translation [Finatto and Tcacenco 2021] represent the core elements to be incorporated into the NCS repertoire:

1. **Lexical layer**: technical terms changed into accessible equivalents (such as "pressão alta" instead of "hipertensão"). This also includes specific words and terms that are infrequent in CorPop, such as "medicamento".
2. **Discursive layer (explanatory paraphrases)**: didactic explanations that serve as a reference standard for explaining complex mechanisms, such as how the medicine acts on the heart.
3. **Pragmatic layer**: direct engagement strategies, including the use of "você" and the active voice, which bring the text closer to the reader, as validated by the linguists

3. Discussion

Human validation (S2) yields higher readability than S1 (Table 2) by prioritizing terminological precision through elaboration. The modest gain in the Flesch index (42.8 versus 40.9) is a direct consequence of the strategy to expand the text with didactic explanations. Because these formulas are sensitive to sentence length, adding context through

explanatory paraphrasing limits improvements in the score. Still, S2 achieved better results in both Flesch and Dale-Chall metrics despite its larger token volume (Table 1). This suggests that choosing accessible terms compensates for the longer sentence length, reinforcing the importance of all three layers in the NCS repertoire.

Beyond global scores, the analysis of frequency ratios and lexical ranking (Table 3) is key to determining the terms to be included in the NCS repertoire. A qualitative screening in CorPop shows that, while biological terms like "coração", "sangue", and "pressão alta" are highly frequent, they often appear in metaphorical contexts. However, this high frequency in common language indicates **reinforced familiarity**: because these terms are already deeply integrated into the reader's vocabulary, they serve as ideal anchors for intralingual translation [Finatto and Tcacenco 2021]. By prioritizing familiar biological terms over specialized terminology (such as "hipertensão"), this selection process takes advantage of existing linguistic habits to reduce cognitive load and improve accessibility. This strategy is particularly relevant because reading difficulties tend to intensify under stress, such as in self-medication contexts, where linguistic complexity may contribute to 25% of users stopping their medication prematurely [Dal Pizzol et al. 2019].

In contrast, the analysis also identifies essential terms with low frequency in CorPop, such as "medicamento" and "dose" (Table 3). These units exhibit extreme frequency ratios: "medicamento" is 569.1 times more frequent in the PILs (S2) than in the general reference *corpus*, while "dose" is nearly 1,200 times more frequent (ratio 1197.8). In context, these are **terminological units** that carry technical concepts and are candidates for the NCS repertoire. Their consistent presence in S2 (Table 3) highlights that even units from common language require conceptual disambiguation to prevent errors, particularly when they are potentially polysemic [Finatto 2022]. Within the scope of the simplification process as a kind of intralingual translation, these terms form essential parts of the NCS vocabulary, and the discursive layer intentionally offers explanations to bridge semantic gaps for the reader.

4. Final remarks and future work

The analysis of the hypertension PILs sample suggests that human elaborative strategies, organized into lexical, discursive, and pragmatic layers, helped to resolve semantic and structural gaps often found in the AI simplifications. These strategies will be incorporated into the NCS repertoire as the basis of future validation efforts in the next phases of the project. Expanding the NCS will also require a standardized validation protocol for consistency across linguists. This will guide the adaptation of specialized terminology based on the CorPop reference while preserving clinical accuracy through layered explanations. We also intend to investigate multi-word expressions (MWEs) to better capture complex medical concepts in simplified health communications.

As mentioned in previous sections, this study only reports on the initial stages of the project (Phase 1 - PILs). Its conclusions are limited to describing and analyzing the validation strategies applied by linguists to review the automatically generated simplifications. This serves to support the evaluation of the potential use, in the new *corpus*, of AI-generated simplifications and reviewed simplifications. The next stage involves a final technical review by healthcare professionals (Phase 2) so that, in the future *corpus*, the

simplified language alternatives also have scientific quality and conceptual accuracy.

5. Acknowledgments

This research was supported by CNPq (National Council for Scientific and Technological Development), grant PQ 307088/2023-5, National Institutes of Science and Technology – CNPq-TILD-IAR (grant 408490/2024-1), PROBIC FAPERGS-UFRGS, and FAPERGS Ed. 06/2025 - PROGRAMA PESQUISADOR GAÚCHO – PqG (grant 25/2551-0002702). We thank the Programa de Pós-Graduação em Letras/UFRGS for the support. We also thank Silvana de Jesus and Felipe L. Duarte Pacheco, who helped review some simplifications, and Arthur Scalercio, who provided the data files used in this research.

6. Limitations

This exploratory study is limited to a sample of 30 hypertension Patient Information Leaflets. The observed validation strategies and lexical shifts may not fully generalize to other medical domains. As previously reiterated, the human-validated simplifications (S2) have not yet been empirically tested with lay readers to assess real-world comprehension improvement.

7. Ethics statement

This study analyzed public Patient Information Leaflets from Anvisa. No human subjects, personal data, or private medical records were involved. The research adheres to ethical standards for linguistic studies, aiming to promote health literacy and social inclusion.

Disclaimer

In compliance with Portaria CNPq 2.664/2026, Gemini CLI version 0.38.2 was used for processing lexical counts, structural organization, and stylistic refinement. ChatGPT (version 5.2) was used to generate the initial automatic simplifications (S1) as part of the research methodology. GPT was used to check text orthography, style, and grammar.

References

Basu, C., Vasu, R., Yasunaga, M., and Yang, Q. (2023). Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12579–12587.

Campos, M. N. (2025). Parkinsonbot: chatbot para auxílio e acessibilidade sobre a doença de parkinson. Trabalho de conclusão de curso (graduação em ciência da computação), Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre. Accessed: 25 Sep. 2025.

Chall, J. S. and Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Brookline Books.

Dal Pizzol, T. d. S., Moraes, C. G., Arrais, P. S. D., Bertoldi, A. D., Ramos, L. R., Farias, M. R., Oliveira, M. A., Tavares, N. U. L., Luiza, V. L., and Mengue, S. S. (2019). Medicine package inserts from the users' perspective: are they read and understood? *Revista Brasileira de Epidemiologia*, 22:e190009.

Finatto, M. J. B. and Paraguassu, L. B. (2022). Comunicação acessível em temas de saúde pública: textos e terminologias sobre epidemias, do passado ao presente. In Moreira, G. L., Costa, L. A. C., and Alves, I. M., editors, *Pesquisas em Lexicologia, Lexicografia e Terminologia*, volume 1, pages 189–323. Pontes Editores, Campinas.

Finatto, M. J. B. and Tcacenco, L. M. (2021). Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. *TradTerm*, 37(1):30–63.

Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. G. V., and Oliveira Jr., O. N. (1996). Readability formulas applied to textbooks in brazilian portuguese. *Notas do ICMSC-USP, São Carlos*.

Menezes, L. C., Paes, A., and Finatto, M. J. B. (2023). Abordagem baseada em aumento de dados para avaliação automática de leiturabilidade. *Domínios de Lingu@gem*, 17:e1721. DOI: 10.14393/DLv17a2023-21.

Pasqualini, B. (2018). Corpop - um corpus do português popular escrito. Thesis, Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre. Accessed: 25 May 2026.

Scalercio, A., Bertotto, E., Jesus, S., Finatto, M. J. B., and Paes, A. (2026). Annotation guidelines and challenges for automatic simplification of portuguese drug leaflets. In *Proceedings of Propor/Lang4Health*.

Scalercio, A., de Souza, E., Finatto, M. J., and Paes, A. (2025). Evaluating LLMs for Portuguese Sentence Simplification with Linguistic Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477. Association for Computational Linguistics.

Wives, L. K. and Finatto, M. J. B. (2025). Usando LLMs para simplificar e representar documentos médicos. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados (SBBD 2025)*, pages 415–425, Porto Alegre, RS, Brasil. SBC.