A dependency-based study of medicine package inserts in Brazilian Portuguese

Adriana S. Pagano[1], André V. Lopes Coneglian[1], Lucas Emanuel Silva e Oliveira[2]

[1]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil
[2]Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba - PR - Brasil

{apagano,coneglian}@ufmg.br,lucas.oliveira@pucpr.br

Abstract. *This paper reports on a study of medicine package inserts (MPIs) aimed at verifying to what extent texts addressing patients evidence different morphosyntactic patterns from those addressing HC professionals. To that end, we draw on a corpus of sentences manually retrieved and aligned, which were annotated for dependency syntax following the UD guidelines. Results point to clear distinctive patterns in both sets of MPIs, which are in line with guidelines on simplified language for Brazilian Portuguese.*

1. Introduction

Medicine package inserts (henceforth MPIs) are acknowledged as critical texts in healthcare activities, particularly in countries where people have less access to medical advice to clarify doubts on how to take prescription drugs or resort to over-the-counter medications sold with no prescription at all. MPIs are texts generally regulated by governmental institutions, which dictate standards for pharmaceutical companies to follow. This is the case, for instance, in the European Union and the United States (Pires et al., 2015). In Brazil, MPIs are required to follow a standard by the National Agency of Sanitary Surveillance (ANVISA) both for format and content. Pursuant to Resolution 047 (ANVISA, 2009a), MPIs are required to comprise two separate sections, one addressing patients and another one, healthcare professionals (henceforth HC professionals). ANVISA (2009b) has also published guidelines with best practices regarding the language used in MPIs, including recommendations on accessibility for blind and deaf people.

However, despite pharmaceutical companies' efforts to adhere to ANVISA guidelines, studies on the legibility and understandability of MPIs have shown that texts still pose enormous challenges to patients. Pizzol et al. (2019), for one, carried out a national survey of over 28,000 individuals in Brazil, revealing that although almost 60% of respondents found MPIs relevant texts to read, over 50% of them reported difficulties in reading and understanding them. To make matters worse, respondents with a lower literacy level reported greater difficulty in understanding MPIs.

While some linguistic studies have been carried out on the language of MPIs in Brazil (Amorim et al., 2015), to the best of our knowledge no study has been carried out on texts annotated for dependency relations with a view to comparing MPIs targeting patients and those targeting HC professionals. If text in MPIs is purportedly adapted to a specific target audience (patients vs. HC professionals), differences in language patterns

are expected to be found and these patterns are expected to comply with patterns suggested in text simplification tasks. This paper reports on an exploratory study of the syntax in MPIs aimed at verifying to what extent texts addressing patients evidence different morphosyntactic patterns from those addressing HC professionals. To that end, we draw on a corpus of sentences manually annotated for dependency syntax following the UD guidelines.

The remainder of this paper is structured as follows. Section 2 provides an overview of previous work on syntax complexity indicators and text simplification with a special focus on Brazilian Portuguese. Section 3 describes the corpus compiled and annotated in our study and the steps followed to analyze our data. Section 4 presents the results of our analysis and our findings regarding MPIs comparison. Section 5 discusses our findings with regard to the available literature. Finally, Section 6 presents the main conclusions of our study, its limitations and suggestions for further research.

2. Syntax complexity and text simplification

Drawing on cognitive and psycholinguistic research, studies on text simplification have investigated a number of language indicators for text simplification tasks. These include choices in morphology, lexis and syntax at sentence level as well as in cohesion at discourse level (Siddharthan, 2006). These indicators have been used in different approaches to the simplification task, relying on more or less manual annotation and implementing different solutions such as Phrase-Based Machine Translation, Syntax Based Machine Translation, transformation rules, and methods drawn from other computational tasks, as is text compression (Siddharthan et al., 2014). More recently, studies have begun to explore dependency trees in order to propose rules for simplification. Angrosh, Nomoto & Siddharthan (2014) explore dependency trees to perform lexical and syntactic simplifications. Chatterjee & Agarwal (2021) developed a rule-based tool (DEPSYM) for simplification drawing on dependency trees and focusing on coordinate and subordinate clauses (appositive and relative clauses) and passive-to-active voice conversion.

Likewise, in Brazil, the Interinstitutional Center for Computational Linguistics (NILC) in the State of São Paulo has developed corpora and tools to simplify texts (Aluísio et al., 2008a,b; 2010; Leal et al., 2022) focusing on clause complexes involving coordination and subordination. Hence, subordinate noun clauses functioning as apposition, relative clauses, and adverbial clauses are filtered out and turned into individual sentences. Moreover, passive voice constructions are rewritten into their active voice counterparts. With regards to MPIs, ANVISA itself published guidelines for drafting texts targeting patients (Brasil, 2009). To comply with them, pharmaceutical companies are instructed to use colloquial instead of medical terms, avoid coordinate and subordinate clauses, use verbs rather than abstract nominalizations and prefer active voice constructions to passive voice ones.

To the best of our knowledge, no work has reported on studies exploring dependency syntax for the purposes of text simplification in Brazilian Portuguese. Nor has any study been reported on using dependency syntax to compare texts targeting different readerships with different levels of literacy and domain knowledge. In this respect, MPIs offer a valuable source for corpus compilation of monolingual texts and

their annotation with dependency relations, a fertile approach to gather insights for prospective text simplification tasks.

3. Methodology

In order to carry out our study, we first retrieved MPI texts written and published in Brazil, targeting patients and HC professionals. Sentences representative of each target group were manually extracted in order to compile a comparable corpus. Since not all pieces of information in HC professionals MPIs are included in patient MPIs, for each sentence in patient MPIs we manually selected a counterpart sentence construing a closely analogous meaning in HC professional MPIs. Table 1 shows manually retrieved and aligned pairs of sentences illustrating analogous meaning construed in a patient and a HC professional MPI. An English gloss is provided beneath them.

Table 1. Aligned sentence pairs

	Patient MPI	HC professional MPI
(1)	Se ocorrerem reações cutâneas, como vermelhidão na pele, bolhas e erupções cutâneas, ou qualquer outro sinal de hipersensibilidade, ou ainda piora de problemas de pele já existentes, interrompa o uso do medicamento e procure ajuda médica imediatamente.	O uso do medicamento deve ser descontinuado no primeiro aparecimento de erupções cutâneas ou qualquer outro sinal de hipersensibilidade.
gloss	*If skin reactions, such as redness, blistering and eruptions, or any other sign of hypersensitivity occur or, still, if conditions worsen, stop use of this medication and seek medical attention right away*.	*Use of the drug should be discontinued at the first appearance of skin eruptions or any other sign of hypersensitivity.*
(2)	Não utilize NALDECON DIA juntamente com outros medicamentos que contenham paracetamol.	NALDECON DIA não deve ser usado juntamente com outros medicamentos que contenham paracetamol em sua formulação, devido ao risco de toxicidade hepática.
gloss	*Do not take NALDECON DAY together with other paracetamol-containing products.*	*NALDECON DAY should not be administered with other preparations containing paracetamol in their formulations, due to the risk of hepatotoxicity.*

As the examples in Table 1 show, there is variation in the length of the aligned segments, some of the sentences in patient MDIs being, at times, longer than their counterparts in HC professional MDIs, while at other times the reverse is the case. 200 sentences were selected for each target group, making up a corpus of 400 sentences. Table 2 shows basic statistics of our corpus, which reveal that despite variability in length, sentences in HC professional MPIs have a higher overall number of tokens for the whole set of 200 sentences and a higher average number of tokens per sentence.

Table 2. MPIs corpus

	Patient	HC professional
Total number. of sentences	200	200
Total number of tokens	2762	4816
Average number of tokens per sentence	13.81	24.08

Sentences were parsed using a freely available neural network pipeline for tokenization, tagging, lemmatization and dependency parsing (UDPipe[1]) with a Portuguese language model (Portuguese-bosque-UD-2.10). The output CoNLL-U files were then uploaded into the Arborator Grew NILC[2] tool developed by ICMC/USP and manually revised following the latest annotation guidelines for Brazilian Portuguese (Duran, 2021, 2022).

In order to compare both sets of texts, we computed the total number of tokens and the average number of tokens per sentence. We then computed POS and dependency relation tags and their relative frequencies in order to allow for comparability between the two sets. We focused on tags indicative of coordinated, subordinated and passive constructions in order to verify if texts followed the guidelines available for simplified Portuguese.

4. Results

Figure 1 shows the number of POS tags annotated for each set of MPIs and their relative frequency.

[1] Available at https://lindat.mff.cuni.cz/services/udpipe/
[2] Available at https://arborator.icmc.usp.br/#/

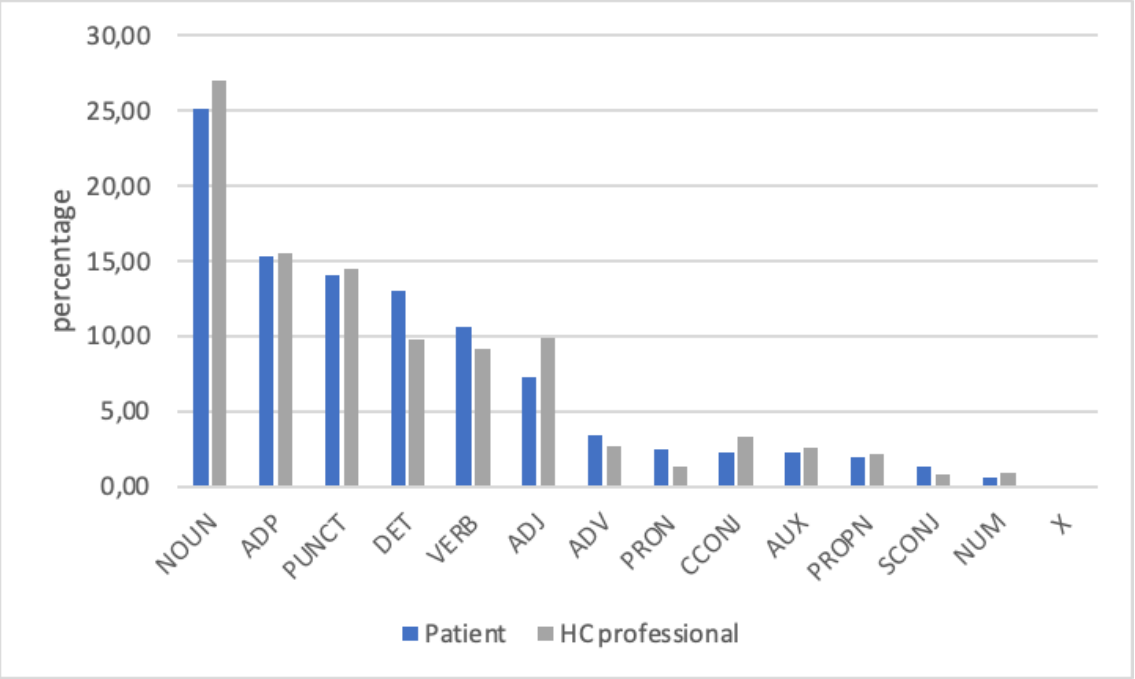

Figure 1. Relative frequency of POS tags

The frequency of POS tags shows that patient MPIs have a higher number of verbs when compared to HC professionals, which, instead, have a larger number of nouns and adjectives. This is in line with simplified language guidelines suggesting patient MPIs make use of less abstract nominalizations. Patient MPIs also have a higher number of pronouns, which can be accounted for by the fact that they address the reader with a second-person singular pronoun ("você"), while HC professional MPIs refer to patients with the noun "patient". The number of coordinating conjunctions is lower in patient MPIs, a finding also in line with guidelines. However, the number of subordinating conjunctions, which would be expected to be lower, is actually higher in patient MPIs. This and other findings need to be interpreted in light of the frequency of dependency relation tags.

Figure 2 shows the number of dependency relation tags annotated for each set of MPIs and their relative frequency.

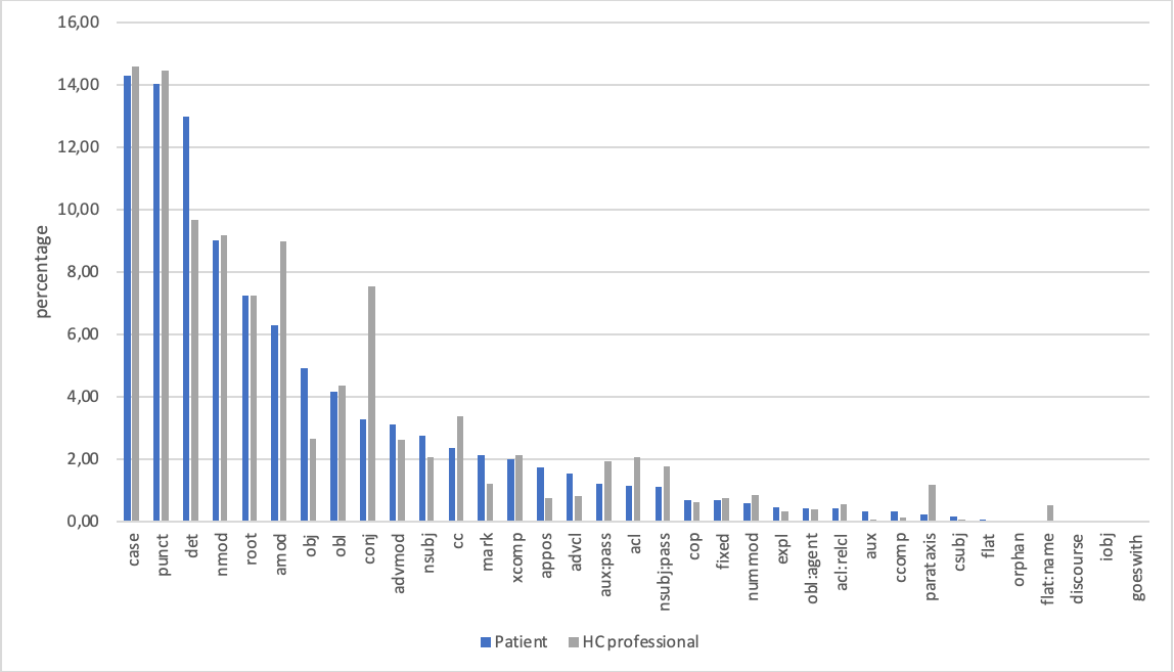

Figure 2. Relative frequency of deprel tags

Figure 2 shows differences for some of the dependency relations in the two sets of texts, many of which can be correlated to frequencies in Figure 1. HC professional MPIs have a higher number of conjuncts (conj) and coordinating conjunctions (cc), a finding that can be related to the higher frequency of the POS coordinating conjunction. Likewise, the higher number of adjectival modifiers in HC professional MPIs can be related to the frequency of the POS adjective. There is also a higher frequency of dependency relations pertaining to passive voice constructions (aux:pass; nsubj:pass) in HC professional PMIs. There is a higher frequency of adjectival clauses (acl) in HC professional MPIs as well. Also remarkable is the higher number of parataxis in HC professional MPIs, a relation established between main clauses and intersected or parenthetical explanations. Regarding patient MPIs, adverbial clauses (advcl) outnumber those in HC professional MPIs, which can account for the higher frequency of markers (mark), i.e., words marking a clause as being subordinate to another clause. This, in turn, may account for the higher number of the POS subordinating conjunction in patient MPIs as seen in Table 1. Adverbial clauses are typically used to construe if-then conditionals.

Sentence 3 in Figure 3[3] is an example of an annotated sentence retrieved from a patient MPI, which evidences the use of active voice and a second-person form of address, in this case, through the use of an imperative form ("consulte" [consult]) and a verb inflected for the form of address to the reader/patient ("tenha" [you have]). It also shows the recurrent use of adverbial clauses to construe a conditional meaning in patient MPIs.

[3] Images were obtained with the tikz-dependency package in a LaTeX editor.

(3) Não use este medicamento caso tenha asma ou úlcera no estômago.
[*Do not use this medication if you have asthma or a stomach ulcer.*]

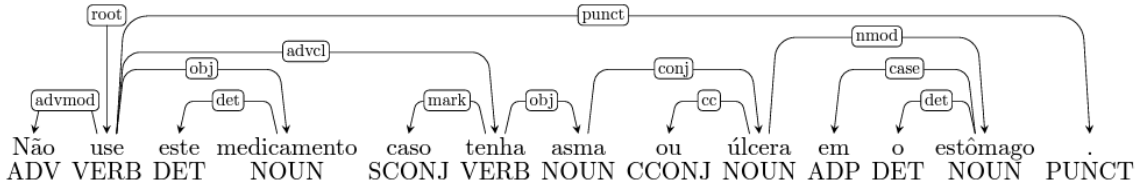

Figure 3. Sample annotated sentence from a patient MPI

Sentence 4 in Figure 4 is a counterpart sentence retrieved from a HC professional MPI.

(4) Este medicamento não deve ser utilizado por pacientes que tenham asma ou úlcera estomacal.
[*This medication should not be used in patients who have asthma or a stomach ulcer.*]

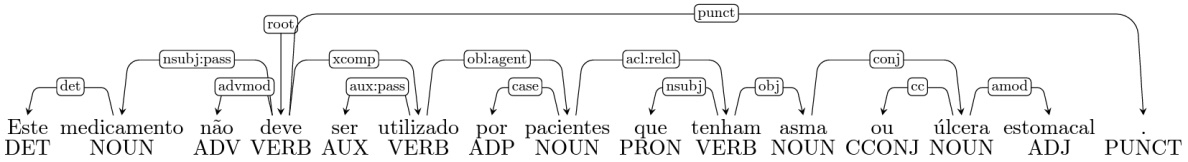

Figure 4. Sample annotated sentence from a HC professional MPI

Unlike (3), sentence 4 shows the use of a passive voice construction, a third-person form of address ("pacientes" [patients]), a defining relative clause ("que tenham asma..." [who have asthma]) and the use of an adjectival modifier ("estomacal" [stomach]).

A further distinction worth noting is the differential use of appositional modifiers in patient and HC professional MPIs. Besides being more frequent in patient MPIs, appositional modifiers (appos) are frequently used to provide synonyms intended to facilitate patient understanding of medical terms. This is illustrated by sentence 5 in Figure 5.

(5) Fissura na retina (rasgo na retina).

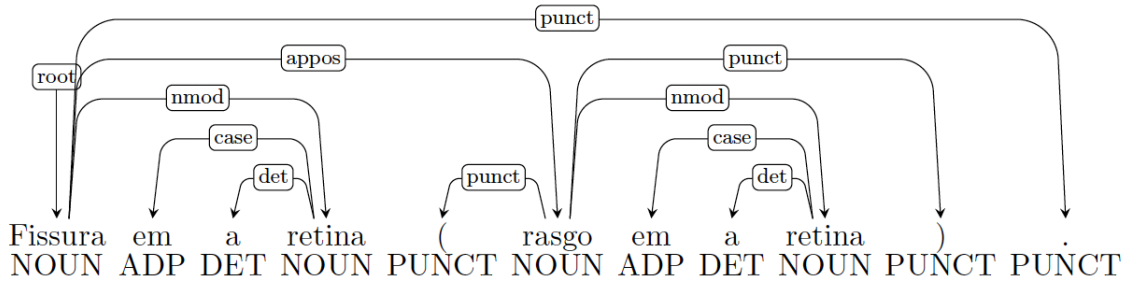

Figure 5. Appositional modifier in annotated sentence from patient MPI

In HC professional MPIs, appositional modifiers are mostly used for abbreviations and acronyms of diseases, as illustrated by example 6 in Figure 6.

(6) Tratamento da hipercalcemia induzida por tumor (HIT).

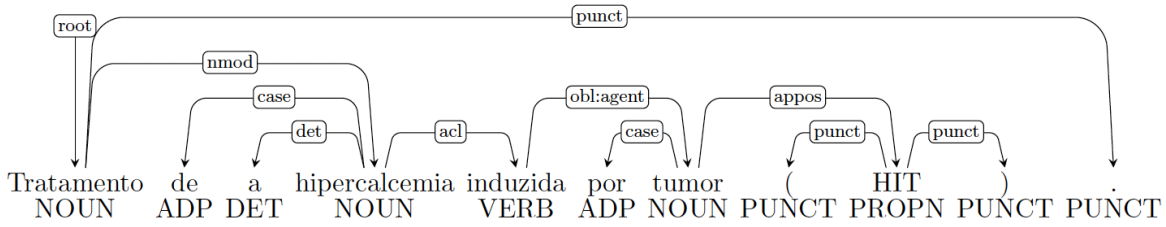

Figure 6. Appositional modifier in annotated sentence from HC professional MPI

5. Discussion

Results for patient MPIs, namely lower frequency of POS tags and dependency relations indicating coordinating constructions and low frequency of passive constructions, suggest that the texts are in compliance with the guidelines for simplified language (Aluísio et al., 2008a,b; 2010; ANVISA, 2009b). So is the use of appositional modifiers to provide synonyms for medical terms. The high frequency of adverbial clauses is accounted for by mostly conditional (if-then) clauses. Simplification guidelines do not advise to split this type of adverbial clause.

Results for HC professional MPIs are also in line with characteristics assumed to pertain to increased text complexity, as they evidence patterns clearly in contrast to those in patient MPIs.

6. Conclusion

This paper reported on a study exploring dependency syntax for the purposes of comparing texts targeting different readerships with different levels of literacy and

domain knowledge. MPIs were found to differ in their morphosyntactic patterns, which are in line with guidelines for simplified text in Brazilian Portuguese.

A corpus of 400 sentences in Brazilian Portuguese manually selected and aligned was compiled and a treebank of 400 sentences annotated for POS and dependency relations following the UD guidelines was developed. Both will be made available for public use.

The UD framework for morphosyntactic annotation proved adequate to retrieve text annotations that can be interpreted in terms of characteristics of simplified texts. Given its potential for comparability, the UD framework is expected to be useful, not only for monolingual aligned sentences, as is the case of our corpus, but also for multilingual sets. Corpora of aligned monolingual texts annotated for dependency relations are useful resources to gather insights for prospective text simplification tasks. In this sense, a further step in our project is to align our corpus with a corpus of MPIs addressing patients and HC professionals written and published in English.

Acknowledgements

The authors would like to thank two anonymous reviewers for their valuable comments. Adriana S. Pagano holds a research productivity grant awarded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (Processo CNPq 313103/2021-6).

References

Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E.G., & Fortes, R.P. (2008a) Towards Brazilian Portuguese Automatic Text Simplification Systems. In: Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), pages 240-248, São Paulo, Brazil. https://doi.org/10.1145/1410140.1410191.

Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E.G., Caseli, H. & Fortes, R.P. (2008b) A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems In: Proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pages 15-22.

Aluísio, S.M., & Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. North American Chapter of the Association for Computational Linguistics.

Amorim, C. M. da S., Rocha, L. H. P. da, & Costa, M. J. (2015) A linguagem da bula: um estudo de estruturas linguísticas do gênero. Letrônica, 8(2), pages 467–479. https://doi.org/10.15448/1984-4301.2015.2.20401.

Angrosh, M., Nomoto, T., and Siddharthan, A. (2014) Lexico-syntactic text simplification and compression with typed dependencies. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014), Dublin, Ireland, pages 1996–2006..

ANVISA - Agência Nacional de Vigilância Sanitária. (2009a) Resolution RDC nº 47. Brasil: Agência Nacional de Vigilância Sanitária. Available at: http://www.anvisa.gov.br/medicamentos/bulas/rdc_47.pdf. Access on 29 June 2023.

ANVISA - Agência Nacional de Vigilância Sanitária. (2009b) Guia de Redação de Bula Gerência-geral de Medicamentos. GGMED. Brasília. Available at: https://www.gov.br/anvisa/pt-br/setorregulado/regularizacao/medicamentos/bulas-rotulos-e-nome-comercial/arquivos/copy8_of_GuiadeRedaodeBula.pdf. Access on 29 June 2023.

Duran, M. S. (2021) Manual de anotação de PoS tags. Relatório Técnico, n. 434. NILC-ICMC/USP. Available at: https://sites.google.com/icmc.usp.br/poetisa. Access on 29 June 2023.

Duran, M. S. (2022) Manual de Anotação de Relações de Dependência: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Available at: https://sites.google.com/icmc.usp.br/poetisa. Access on 25 June 2023.

Leal, S.; Duran, M.; Scarton, C.; Hartmann, N.; Aluísio, S. (2022) NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. CoRR abs/2201.03445. Available at: https://arxiv.org/abs/2201.03445. Access on 14 August 2023.

Pires, C., Vigário, M., & Cavaco, A. (2015) Readability of medicinal package leaflets: a systematic review. Revista De Saúde Pública, 49. Available at: https://doi.org/10.1590/S0034-8910.2015049005559. Access on 29 June 2023.

Pizzol, T. da S. D., 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., & Mengue, S. S. (2019) Medicine package inserts from the users' perspective: are they read and understood?. Revista Brasileira De Epidemiologia, 22, e190009. https://doi.org/10.1590/1980-549720190009.

Siddharthan, A. (2011) Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In Proceedings of the 13th European Workshop on Natural Language Generation, pages 2–11, Nancy, France. Association for Computational Linguistics. Available at: https://aclanthology.org/W11-2802/. Access on 14 August 2023.