MedSimples: An Automated Simplification Tool for Promoting Health Literacy in Brazil*

Liana Paraguassu[1][0000−0003−4043−1836], Leonardo Zilio[2][0000−0002−6101−0814], Luis Antonio Leiva Hercules[3], and Maria José Bocorny Finatto[1][0000−0002−6022−8408]

[1] Universidade Federal do Rio Grande do Sul, UFRGS, Brazil
liana@linguatraducoes.com, mariafinatto@gmail.com
[2] University of Surrey, United Kingdom, l.zilio@surrey.ac.uk
[3] Automatic Data Processing, ADP, Brazil

Functional illiteracy rates in Brazil are critical. According to a recent study (2018) conducted by the Paulo Montenegro Institute, 3 out of 10 Brazilians are considered functional illiterates. Also according to INAF[4], only 12% are truly proficcient readers. On the other hand, with the significant increase of Internet access in Brazil in the past years, information is available to a much larger number of people. According to the Brazilian Institute of Geography and Statistics (IBGE), in 2017, 67,00% of the Brazilian population had access to the Internet. Although a search on the Internet by a layperson cannot replace going to the doctor, the growing number of Internet users who rely on it as a source of information is a reality that cannot be overlooked. Therefore, it is important that the source be reliable, but also that the information provided on these sources be linguistically accessible and understandable by people with low levels of literacy. In this scenario, our research problem is this: How can we render health-related information available on the Web in a linguistically accessible way to people with limited education and low literacy skills? Our project combined Natural Language Processing, Corpus Linguistics and Terminology Studies to develop MedSimples, an online tool that automatically identifies complex phrases in health-related texts presented by the user and offers suggestions of simplification. MedSimples is an example of how research on medical language, associated with NLP, can enrich the current scenario of Digital Humanities in its broadest scope.

The prototypical user for the tool is the Health or Communication professional interested in producing accessible texts graduated according to the needs of their audience. However, it can also be used by anyone who is interested in getting simplified health-related information for themselves. Our Text Simplification approach is focused on lexical and terminological levels, and the goal is to reduce complexity while preserving meaning. Lexical simplification can have

Supported by LARA 2019 (Google Inc.), CNPq - Brasil (Productivity in research 305625/2016-0), Universal (403521/2016-5), SEAD-UFRGS, and Expanding Excellence in England (E3).

[4] INAF is a Brazilian literacy indicator. http://www.ipm.org.br/inaf

two approaches [3]: modifying the vocabulary on a text by selecting words that are more adequate to the reader's reading skills, or adding explanations or definitions to the vocabulary that cannot be replaced. MedSimples combines both approaches, using the suggestions of modifications mostly for complex phrases, and then using simpler definitions for the terminology that is present in the text. We believe that this way the information is preserved, being more accurate and reliable, and, in addition to that, by explaining the complex vocabulary, we can promote health literacy by educating our readers on health issues.

The tool was initially built based on a Parkinson's Disease (PD) corpus. The PD corpus is a representative collection of original texts available on the Internet and published by reliable sources. These texts have been simplified by Linguistic graduate and undergraduate students and validated by health specialists to build a parallel simplified PD corpus. This corpus was then further analysed and converted into an initial list of terms and a list of example-sentences with varying degrees of complexity, that can be used for consultation by the health professionals.

In terms of automatic processing of texts, for identifying complex phrases and terms, and offering suggestions of simplification or simpler definitions for the user, MedSimples currently relies on three lexical resources and one parser. The three lexical resources comprise a list of words that are considered simple, a list of words that are considered complex along with simpler synonyms, and a glossary of terms with simple definitions. The first step in MedSimples process is to parse the text that was selected by the user, we used the PassPort parser [4] for annotating morphological and lemma information on the text. This annotated text is then processed and, when MedSimples recognizes a term that is listed in our glossary it automatically provides a simplified explanation. For the complex words that are not considered health-related terms, we used the synsets present in the Thesaurus of Portuguese (TeP 2.0) [1], which were filtered based on a list of simple words extracted from CorPop [2], a corpus of texts considered fairly accessible to the average Brazilian reader.

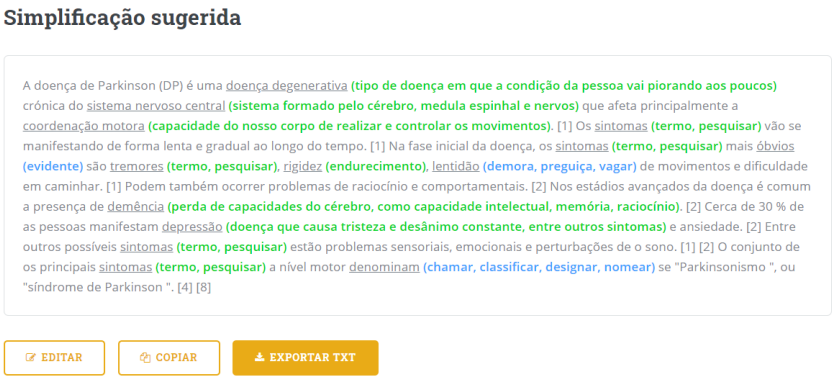

Fig. 1. Parkinson's Disease Simplification. Source text: Wikipedia.

Figure 1 shows an excerpt of a text about PD extracted from the Internet and processed on MedSimples. The synonyms in blue are from the filtered TeP 2.0 synsets and the explanations in green are from our term glossary. Examples: (term from glossary) **doença degenerativa** (tipo de doença em que a condição da pessoa vai piorando aos poucos), (filtered item, difficult word) **denominam** (chamar, classificar, designar, nomear).

Although MedSimples is at its early stages it has shown promising results with PD material. Currently, we are expanding our database to cover a wider range of health-related topics. In addition, we are working on the qualification of our glossary of health-related terms and our list of complex words. For instance, we are refining and expanding our list of terms to cover words that should be considered terms, such as "demência" and "tremores" (Fig. 1), and that were previously listed as complex words, which was usually leading to inaccurate suggestions of simplification.

References

1. Maziero, E., Pardo, T.: Interface de acesso ao tep 2.0–thesaurus para o português do brasil. Relatório técnico. University of Sao Paulo (2008)
2. Pasqualini, B.F.: CorPop: um corpus de referência do português popular escrito do Brasil. Ph.D. thesis, Universidade Federal do Rio Grande do Sul (2018)
3. Saggion, H.: Automatic text simplification: Synthesis lectures on human language technologies, vol. 10 (1). California, Morgan & Claypool Publishers (2017)
4. Zilio, L., Wilkens, R., Fairon, C.: Passport: A dependency parsing model for portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 479–489. Springer (2018)