1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 70–76
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Lexical Simplification Tool for Promoting Health Literacy

Leonardo Zilio[1], Liana Braga Paraguassu[2], Luis Antonio Leiva Hercules[3],
Gabriel L. Ponomarenko[2], Laura P. Berwanger[2], Maria José Bocorny Finatto[2]
[1] Centre for Translation Studies, University of Surrey, United Kingdom, l.zilio@surrey.ac.uk
[2] PPG-LETRAS, Federal University of Rio Grande do Sul, Brazil,
liana@linguatraducoes.com, mariafinatto@gmail.com
[3] Automatic Data Processing, Brazil

Abstract

This paper presents MedSimples, an authoring tool that combines Natural Language Processing, Corpus Linguistics and Terminology to help writers to convert health-related information into a more accessible version for people with low literacy skills. MedSimples applies parsing methods associated with lexical resources to automatically evaluate a text and present simplification suggestions that are more suitable for the target audience. Using the suggestions provided by the tool, the author can adapt the original text and make it more accessible. The focus of MedSimples lies on texts for special purposes, so that it not only deals with general vocabulary, but also with specialized terms. The tool is currently under development, but an online working prototype exists and can be tested freely. An assessment of MedSimples was carried out aiming at evaluating its current performance with some promising results, especially for informing the future developments that are planned for the tool.

Keywords: Healthcare, Health Literacy, Lexical Simplification, Natural Language Processing, Corpus Linguistics, Terminology

1. Introduction

Most health professionals in Brazil have no specific or even complementary training in the area of communication. However, when it comes to health-related information, as Cambricoli (2019) points out, based on a study made by Google, 26% of Brazilians have the Internet as their first source to look for information about their own or their relatives' illnesses, which puts Brazil in the number one position in health-related searches on Google and the YouTube. In a scenario like that, it is important to have support for improving health communication and patient understanding, and this is directly related to health literacy. Health literacy is about communication and understanding; it affects how people understand wellness and illness, and participate in health promotion and prevention activities (Osborne, 2005).

Adding to the question of health literacy, Brazil presents a panorama where functional illiteracy[1] rates are critical. According to a recent INAF[2] report (Lima and Catelli Jr, 2018) published by the Paulo Montenegro Institute, 29% of Brazilians (38 million people) with ages ranging from 15 to 64 years old are considered functional illiterates. Also according to this INAF report, only 12% of the Brazilian population at working age can be considered proficient.

Even though literacy skills are low on the country, Brazil has perceived a significant increase of Internet access in the past years, and information has become available to a much larger number of people. According to the Brazilian Institute of Geography and Statistics (IBGE)[3], in 2017, 67% of the Brazilian population have access to the Internet, as opposed to less than half of the population in 2013.

As it is now, the Brazilian scenario shows a considerable number of people looking for health-related information on the Internet, while only a small percentage of the population can be considered proficient. Adding to that, health professionals don't usually receive the necessary training for providing information that matches the literacy level of a large number of people. In this scenario, a tool that aims at making information more accessible to different audience profiles and that respects the choices of a specialized writer can provide a relevant service both for professionals in charge of communication and for the society in general.

MedSimples[4] was conceived for supporting the involvement of health professionals and health communication professionals and for helping them to write information that can be understood by a large part of the population. It is a tool that was designed to help professionals in the task of improving the communication of health-related information to lay people that have low literacy skills. In that way, MedSimples works as a text simplification tool that highlights lexical items and offers suggestions that could improve the accessibility of a health-related text for the Brazilian population. The project is currently focused on the Parkinson's disease domain, and in this paper our aim is to conduct an initial evaluation of the tool, so that we can draw some considerations for its future improvements, especially bearing in mind that the current working structure of MedSimples will be later adjusted for other topics from the Health Sciences.

This paper is divided as follows: Section 2 presents information about text simplification in general and about the PorSimples project, which deals with text simplification for Portuguese; in Section 3, we present how MedSimples was

[1] People are considered functionally illiterate when they cannot use reading, writing, and calculation skills for their own and the community's development.

[2] INAF is a Brazilian literacy indicator. More information about INAF can be found at: `http://www.ipm.org.br/inaf`

[3] `https://bit.ly/2HBwmND`

[4] Freely available at: `http://www.ufrgs.br/textecc/acessibilidade/page/cartilha/`.

build, how it works and what are its main features and resources; Section 4 discusses the methodology we applied for evaluating MedSimples and presents its results; in Section 5 we further discuss the evaluation by presenting some data from an error analysis; finally, Section 6 reports on the main findings of this paper and discusses future improvements and changes to the online tool.

2. Related Work

There are several studies regarding text simplification in general and regarding areas that are directly related to text simplification, such as readability assessment (e.g. Vajjala and Meurers (2014), complex word identification (e.g. Wilkens et al. (2014)), intralingual translation (e.g. Rossetti (2019)). However, in this section, we will first focus on briefly introducing the task of text simplification in general, presenting different levels of simplification, and proceed to describe some more applied related work that was developed in the form of a tool that deals with the task of simplifying texts written in Portuguese.

2.1. Text Simplification

In Natural Language Processing, the text simplification task focuses on rewriting a text, adding complementary information (e.g. definitions), and/or discarding irrelevant information for minimizing the text's complexity, but all the while trying to assure that the meaning of the simplified text be not greatly altered, and that the new, rewritten version seem natural and fluid for the reader (Siddharthan, 2002; Siddharthan, 2014; Paetzold and Specia, 2015). This simplification usually occurs by replacing complex words or phrases with simpler ones, in what is called lexical simplification, and/or by modifying the text syntactical structure to render it more simple, which is called a syntactical simplification.

Different types of simplification architectures have been proposed (e.g. Siddharthan (2002; Gasperin et al. (2009; Coster and Kauchak (2011; Paetzold and Specia (2015)), dealing with either or both levels of simplification, generally going from the syntactical level to the lexical level. In this paper, we are focusing on the lexical level, following the bases described by Saggion (2017). MedSimples addresses words, phrases and terms that may be complex for people with low literacy and presents simpler suggestions or term explanations. However, it is important to point out that MedSimples does not focus on trying to automatically replace complex phrases. It is designed to help communicators of health-related information to write more simplified texts. As such, it only presents suggestions of changes, in the form of simpler words or term explanations, that may or may not be accepted by the author of the text.

2.2. Simplification for Portuguese

For Portuguese, there are studies focusing on the classification of complex texts, such as Wagner Filho et al. (2016), and Gazzola et al. (2019), and others that aim at evaluating sentence complexity, such as Leal et al. (2019). However, for the purposes of text simplification, i.e., identifying complex structures of a text and suggesting simpler replacement structures, in the way that we are looking for in MedSimples, project PorSimples (Aluísio et al., 2008; Aluísio and Gasperin, 2010) is the one that currently exists with the most similarities.

The project PorSimples deals with the challenges of text simplification and has an online tool called Simplifica (Scarton et al., 2010) that helps authors to write simpler texts. Simplifica uses lexical resources allied with automatically extracted features to identify complex parts of a text and make suggestions on how to make it more readable for people with low literacy. It presents a module for lexical simplification and another module for syntactical simplification, allowing for some customization in terms of which resources are used and which types of syntactical structures are target of the simplification.

While Simplifica serves as an interesting model as a simplification authoring tool, it focuses on the general language, and, as such, it usually cannot suggest befitting simplifications for specialized terms, and this is where the main strength of MedSimples lies. By drawing on specialized resources, MedSimples aims at focusing on different areas of the human knowledge for providing more suitable suggestions for simplifications, and, by aiming at health-related texts, it addresses a widely recognized issue for text simplification (Rossetti, 2019).

3. System Description

MedSimples relies on different corpora and lexical resources, and uses a parsing system at its core. By combining these resources, it can identify complex words and present suggestions for lexical simplification. In this section, we first discuss the lexical resources that were created for MedSimples and then present the pipeline.

3.1. Simple Corpus and Lexical Resources

One of the challenges of text simplification is to identify what kind of vocabulary could be complex to the target audience and try to suggest simpler replacement words or definitions. At this stage of the project, MedSimples deals with the specialized, health-related area of Parkinson's disease[5], so it has to identify not only phrases that are complex from the point of view of the general language, but also terms. It also has to treat complex phrases and terms differently, because offering a simpler lexical suggestion for a term may not help for preserving approximately the same semantic content for the reader, which could lead to serious consequences in a text with information about a health-related subject. For instance, it is possible to substitute the word *involuntário* [involuntary] with *inconsciente* [unconscious] without much semantic difference. However, substituting the term *dopamina* [dopamine] with a simplified version would render the information much less precise, and this could have serious, life-impacting consequences. Considering this different treatment for complex phrases and terms, MedSimples relies on two lexical resources: a list with simpler suggestions for complex phrases from the general language, and a list of simpler definitions for terms (and, when possible, simpler lexical variants).

[5] The inclusion of other health-related areas are already in development.

Resource	Source	# of Items
List of simple words	CorPop	6,881
List of complex words	TeP	15,427
List of terms	Handcrafted + Validation	439

Table 1: Lexical resources used by MedSimples for identifying complex lexical items and suggesting simpler alternatives.

For deciding what should be considered as a complex phrase, we decided to look at the problem from a different perspective. By relying on CorPop (Pasqualini, 2018; Pasqualini and Finatto, 2018), a corpus composed of texts that were written for and/or by people with low literacy skills, we were able to estimate which words could be considered simple for our target audience. The corpus was tagged using the PassPort parser (Zilio et al., 2018), and a frequency-ranked word list was generated considering both lemma and part of speech. From this word list, we selected all words with frequency of five or more to be part of our list of simple words. CorPop is a small corpus, containing around 740k tokens and 24k lemmas associated to different word classes, but it was positively evaluated in terms of adequacy for people with low literacy, so we considered that even a low frequency such as five would be enough to warrant the status of simple word to a lemma that is present in this corpus, this led to a list of almost 7k lemmas (associated to the respective word class).

We used this list from CorPop to then filter the Thesaurus of Portuguese (TeP) 2.0 (Maziero and Pardo, 2008) and generate a list of complex words with simpler synonyms. TeP is a language resource that contains WordNet-like synsets for Portuguese. We automatically analyzed each synset and set complex words (i.e. those which were not in the CorPop list of simple words) as entries, while the other words in the synset that were present in our list of simple words were set as simpler synonyms. This list of complex words with simpler synonyms contains more than 15k entries, and also includes some multiword structures, such as *a favor* [in favor], *abóbada celeste* [celestial dome], *curriculum vitae*, *de súbito* [suddenly].

In addition to the list of complex words with simpler synonyms generated from TeP and the list of simple words extracted from CorPop, MedSimples also relies on a list of terms related to Parkinson's disease. This list is still in the process of being completed and simplified, for achieving definitions that are suitable for our target audience. It is being manually built by linguists and also manually validated by a specialist in Medicine[6].

These three lexical resources are used for the automatic process of complex word identification and suggestion of simplifications, as we explain in the next subsection. Table 1 shows the precise numbers of items in each of them.

3.2. Identification and Suggestions

The MedSimples online tool uses automatic text processing and relies on the PassPort parser (Zilio et al., 2018) for first tagging the text that is used as input by the user. It then analyses each sentence by matching the items first to the list of terms, then to the list of simple words and, finally, to the list of complex words. For matching the list of terms, MedSimples uses the surface forms of words, based on the terminological principle that terms can differentiate themselves by their surface realization (Krieger and Finatto, 2004). Then, it uses the lemma forms to either ignore the word (if it is present in the list of simple words), or to identify it as complex and present a simpler suggestion (if it is present in the complex word list).

MedSimples is still under development, but all the steps mentioned above were already implemented, and the system can visually highlight terms and complex words with suggestions in different colors (depending on whether it is a term or complex word). As it is now, the system is only visually flagging words as complex if there are simpler suggestions in our lexical resources, otherwise, they are ignored. This can be modified, and the idea in the future is to be able to annotate as complex also some types of words that are not in the list of complex words, so as to at least indicate their complexity to the user. Here, for the purpose of this evaluation, we wanted the system to only identify complex words for which we have suggestions, so that we could more easily verify how our suggestions were fitting the context. However, this decision also means we are not currently presenting all the info that we can, and this is reflected in the evaluation process, as will be seen in the next section. This same approach was not used for terms, which we are marking as recognized even if we don't yet have a definition for them. We took this different approach for each type of automatic annotation because the list of terms is much smaller than the number of out-of-vocabulary words, and we expect to have definitions in place for them in the foreseeable future. Figure 1 shows how the system is currently presenting the information about terms and complex phrases. As explained above, this presentation was chosen to speed up the current evaluation, but, in the future, the suggestions will be shown in a different way, in order to not pollute the text for the user.

4. Evaluation

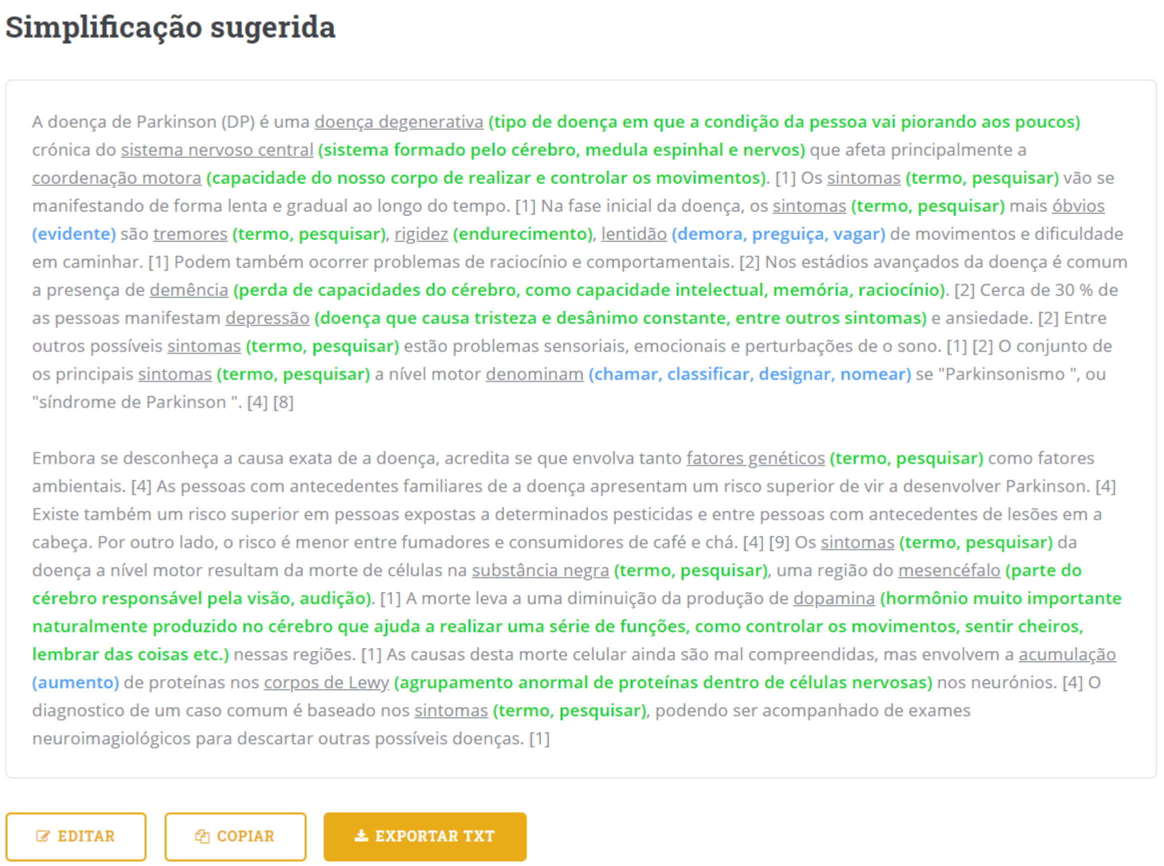
In this paper, one of our aims is to measure how MedSimples is performing in its current state, and what areas should be the focus of our next efforts. To that end, we designed a strict evaluation using a gold standard that was created using authentic online material. In the next subsections, we discuss the creation of the gold standard, then explain the evaluation methodology and, finally, present the results.

4.1. Gold Standard

The first step for creating a gold standard for the evaluation of MedSimples was to create a corpus with texts related to the Parkinson's disease domain. To achieve this, we crawled the web using trigram-combinations of 7 terms related to the target domain: "doença de Parkinson" [Parkinson's disease], "Parkinson", "mal de Parkinson" [alternative denomination for Parkinson's disease[7]], "cuidador"

[6] Ricardo Eizerik Machado, M.D., CRMRJ 52-0110079-3.

[7] "Mal de Parkinson" is an alternative denomination for which the use is currently not recommended by the World Health Orga-



Simplificação sugerida

A doença de Parkinson (DP) é uma doença degenerativa (tipo de doença em que a condição da pessoa vai piorando aos poucos) crónica do sistema nervoso central (sistema formado pelo cérebro, medula espinhal e nervos) que afeta principalmente a coordenação motora (capacidade do nosso corpo de realizar e controlar os movimentos). [1] Os sintomas (termo, pesquisar) vão se manifestando de forma lenta e gradual ao longo do tempo. [1] Na fase inicial da doença, os sintomas (termo, pesquisar) mais óbvios (evidente) são tremores (termo, pesquisar), rigidez (endurecimento), lentidão (demora, preguiça, vagar) de movimentos e dificuldade em caminhar. [1] Podem também ocorrer problemas de raciocínio e comportamentais. [2] Nos estádios avançados da doença é comum a presença de demência (perda de capacidades do cérebro, como capacidade intelectual, memória, raciocínio). [2] Cerca de 30 % de as pessoas manifestam depressão (doença que causa tristeza e desânimo constante, entre outros sintomas) e ansiedade. [2] Entre outros possíveis sintomas (termo, pesquisar) estão problemas sensoriais, emocionais e perturbações de o sono. [1] [2] O conjunto de os principais sintomas (termo, pesquisar) a nível motor denominam (chamar, classificar, designar, nomear) se "Parkinsonismo ", ou "síndrome de Parkinson ". [4] [8]

Embora se desconheça a causa exata de a doença, acredita se que envolva tanto fatores genéticos (termo, pesquisar) como fatores ambientais. [4] As pessoas com antecedentes familiares de a doença apresentam um risco superior de vir a desenvolver Parkinson. [4] Existe também um risco superior em pessoas expostas a determinados pesticidas e entre pessoas com antecedentes de lesões em a cabeça. Por outro lado, o risco é menor entre fumadores e consumidores de café e chá. [4] [9] Os sintomas (termo, pesquisar) da doença a nível motor resultam da morte de células na substância negra (termo, pesquisar), uma região do mesencéfalo (parte do cérebro responsável pela visão, audição). [1] A morte leva a uma diminuição da produção de dopamina (hormônio muito importante naturalmente produzido no cérebro que ajuda a realizar uma série de funções, como controlar os movimentos, sentir cheiros, lembrar das coisas etc.) nessas regiões. [1] As causas desta morte celular ainda são mal compreendidas, mas envolvem a acumulação (aumento) de proteínas nos corpos de Lewy (agrupamento anormal de proteínas dentro de células nervosas) nos neurónios. [4] O diagnostico de um caso comum é baseado nos sintomas (termo, pesquisar), podendo ser acompanhado de exames neuroimagiológicos para descartar outras possíveis doenças. [1]

Figure 1: Suggestions of simplifications for a text excerpt about the Parkinson's disease on MedSimples. Source: https://pt.wikipedia.org/wiki/Doen%C3%A7a_de_Parkinson

[caretaker], "DP" [acronym for Parkinson's disease], "sintoma motor" [motor symptom], and "qualidade de vida" [quality of life]. These terms were manually selected based on word and n-grams lists extracted from the book *Entendendo a Doença de Parkinson* [Understanding Parkinson's Disease] (Rieder et al., 2016). We used slate3k[8] to scrape PDF documents and jusText[9] to exclude boilerplate and non-interesting content. We also made sure to only scrape content from different Websites, by not repeating previously scraped URLs.

From the resulting crawled corpus, we created 8 random samples of 120 medium-to-long sentences[10] each and distributed them to 8 annotators[11]. Each sample had 30 sentences that were annotated by all annotators and 90 sentences that were annotated only by each individual annotator, totaling 750 sentences. Annotators were asked to annotate any word, phrase or term that they deemed to be complex or terminological, making an explicit distinction between terms and complex phrases.

The result of the annotation was then analysed in terms of a pairwise Cohen's kappa inter-annotator agreement (Cohen, 1960) by using the agreement verified on the 30 sentences that were annotated by all. Since it was a free-flow annotation, in which any part of a sentence could be selected for annotation and there was also a classification task (complex phrase or term) on top of it, this can be considered a very complicated task, so we did not expect to achieve high levels of kappa, but we set .20 as a bare minimum. After calculating the agreement (Table 2), two annotated samples were excluded from the gold standard for not achieving a minimum mean kappa score of 0.20. The final Fleiss' kappa score (Fleiss, 1971) for the remaining annotators' samples was 0.25. This filtering process generated a final gold standard with 570 annotated sentences, and 2080 annotated instances. These final instances were thoroughly checked for inconsistencies (errors resulting from the manual annotation) by one of the authors.

4.2. Methodology

Having a gold standard for the evaluation, we randomized the sentences in it and divided all the instances among the authors for evaluation. Since the evaluation was a somewhat more straightforward process, we did not duplicate sentences for calculating the agreement on the evaluation process (as we did for the generation of the gold standard). Some of the gold standard annotators worked as evaluators as well.

For the evaluation, we asked evaluators to check three aspects of the automatic annotation: first, if the word or

nization, because it can cause discrimination or prejudice. Still it can easily appear in online texts about the subject of Parkinson's disease, so we decided to include it as well.

[8] https://pypi.org/project/slate3k/

[9] http://corpus.tools/wiki/Justext

[10] Each sentence in the gold standard has a minimum of 15 space-separated tokens.

[11] All annotators are linguists or undergraduate students of Linguistics. Some of the authors also contributed as annotators.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	1.0000	0.3828	0.4292	0.3823	0.3355	0.4725	0.2259	0.0765
A2	0.3828	1.0000	0.3568	0.2982	0.2290	0.3534	0.2389	0.1667
A3	0.4292	0.3568	1.0000	0.2625	0.3232	0.5775	0.2946	0.0480
A4	0.3823	0.2982	0.2625	1.0000	0.3854	0.2165	0.1121	0.0465
A5	0.3355	0.2290	0.3232	0.3854	1.0000	0.2090	0.1390	0.0237
A6	0.4725	0.3534	0.5775	0.2165	0.2090	1.0000	0.2235	0.1284
A7	0.2259	0.2389	0.2946	0.1121	0.1390	0.2235	1.0000	0.0734
A8	0.0765	0.1667	0.0480	0.0465	0.0237	0.1284	0.0734	1.0000
Mean	0.3292	0.2894	0.3274	0.2433	0.2350	0.3115	0.1868	0.0805

Table 2: Cohen's kappa pairwise agreement among all annotators. The mean scores ignore the lines where annotators are paired with themselves.

phrase was recognized as complex or as a term; second, if it was correctly recognized as either term or difficult phrase; and, third, to check if the suggestion semantically fitted the context[12]. For the evaluation of the semantic and the recognition task, there was an option for a partial match[13]. In order to simplify the process for the human evaluators, we did not further divide the classification of the partially recognized instances into mismatch for term or complex phrases. In addition to the recognition and the semantic evaluation, in cases where MedSimples failed to recognize the target phrase (either no recognition or only partial recognition), evaluators were asked to proceed with an error analysis, by checking if there were no typos (such as numbers attached at the beginning or end of an instance, spelling errors, etc.), foreign words[14] or unrelated terms[15]. The phrases on the gold standard were also compared with the words on the list of simple words to see if there were any matches.

4.3. Results

As we explained in the previous sections, we used a hard test to see how MedSimples is currently performing, especially because the aim of this study was to look for points in which we need to improve in the future. As shown on Table 3, one of the negative results that we got from this evaluation is that MedSimples currently does not achieve a good coverage. From all the instances, 67.88% were not taken into account for simplification in any way. However, there is also positive information coming from these results: for all the instances that were correctly recognized, MedSimples provided the correct meaning on 67.04% of the cases (with a slightly better performance for terms, as expected, which have their suggestions coming from a handcrafted glossary).

When there was a partial recognition of an instance (which could only happen for multiword instances) or a mismatch, we see that MedSimples struggles to provide a suggestion that fits the context. This is especially true in the case of mismatches, where the number of suggestions that do not fit the context (bad suggestions) is 3.5 times higher than the number of good suggestions. By further analyzing the partially recognized instances, we see that the vast majority of unfitting suggestions come from our list of complex words (the one that was automatically created using TeP (Maziero and Pardo, 2008) and CorPop (Pasqualini, 2018)).

5. Discussion

After looking at the results, especially the ones from unrecognized and partially recognized instances, we can look at an error analysis to better understand what was missing.

Table 4 shows information about out-of-scope terms (i.e. terms that do not belong to the area of Parkinson's Disease), foreign words present on the target instances, and typos. The number of out-of-scope terms accounted for 13.05% of the terms that were not recognized by the tool (counting also the ones that were partially recognized or mismatch). The number of foreign words and typos, on the other hand, are almost negligible, accounting for only 4.67% of the unrecognized instances.

As a second part of this error analysis, we looked at our own list of words that are assumed to be simple (this is the list of words that was extracted from CorPop, which was already tested by Pasqualini (2018) in terms of complexity) and matched it against instances that were considered as complex phrases by the annotators. In total, we found out that 393 instances that were not recognized in any form contained words that were in our list of simple words, this accounts for 55.11% of the unrecognized complex phrases in the evaluation.

This comparison revealed a complicated, but expected (as pointed out by Cabré (1993), Krieger and Finatto (2004)), aspect of the lexical simplification: there are words or phrases with a generally simple meaning that can have a complex meaning in specific contexts (for instance, "administração" [administration] in general has a fairly simple meaning, but in the context of "administration of medicines to patients", it takes a more complex meaning). However, by looking further into this comparison, it also

[12] In those cases where the suggestion was a whole synset, only one of the suggested replacement words should fit to be considered a good suggestion. This decision take into consideration that we rely on the user to decide which one of the suggested replacement words would fit the context.

[13] For instance, if only part of a term was identified or if a suggestion of simplification would only partially fit in the context.

[14] Since we are using lexical resources for the Brazilian Portuguese variant, the evaluators were instructed to mark European Portuguese variants as foreign words as well.

[15] Since the corpus was crawled from the internet, there is always the possibility of having sentences that do not belong to the Parkinson's disease domain, even if the keywords used were heavily linked to the domain.

	Recognized			Partially Recognized			Mismatch			Unrecognized	Total
	Good	Bad	Partial	Good	Bad	Partial	Good	Bad	Partial		
Terms	125	47	0	47	87	22	10	35	4	699	1076
Complex phrases	172	73	26	5	11	4	0	0	0	713	1004
Total	297	120	26	52	98	26	10	35	4	1412	2080

Table 3: Evaluation results. The labels “Good”, “Bad” and “Partial” reflect the evaluation of the meaning of MedSimples’ suggestions in the given context.

	Out-of-Scope Terms	**Foreign Words**	**Typos**	**Total**
Terms	118	19	22	159
Complex phrases	0	19	6	25
Total	118	38	28	184

Table 4: Error Analysis

revealed that the number of complex instances in the evaluation may as well have been overestimated (for instance, words like “demonstrar” [demonstrate], “interferir” [to interfere], and “promover” [to promote] were annotated as complex, even if the context in which they appear does not imply a more complex meaning). This observation requires some further analyses that we haven’t yet carried out, to better estimate what could be considered to be included in our current lexical resources and what can be viewed as an overestimation of complexity from the annotation.

The case of words that assume a more complex meaning in context is the one that poses an interesting challenge for MedSimples. Since we are currently not using any type of disambiguation, we have no way of distinguishing between the “administration of a business” and the “administration of medicines”, and this should be a matter to take into account for the future steps of the tool.

6. Final Thoughts and Future Work

In this paper we presented MedSimples, an authoring tool that is mainly focused on helping producers of content from the healthcare industry to provide more accessible texts to Brazilian people with low literacy. MedSimples is currently under development, but has a working online prototype for testing. By accessing the Website, a user can input a text and, after having selected the domain and type of target reader and submitting it for processing, receive suggestions of simpler words or definitions for terms that could be taken into consideration for formulating a more accessible text.

In order to expand MedSimples, an evaluation was developed to assess the current state of the system and to provide useful information for the steps going forward. One of the results of the evaluation was that MedSimples is still lacking in terms of good suggestions that would fit the context of a text dealing with Parkinson’s disease. That is one of the reason’s why the list of complex words and simple suggestions is going to be target of a major review, that intends on checking for entries that are not very helpful and trying to provide suggestions that would potentially present a better fit for the specialized context, considering meanings that would be more in line with the domain. This evaluation also presented some interesting information for expanding MedSimples’ term base, which currently contains almost 450 terms, but that could be expanded to have a broader coverage of the area, possibly including terms that are not directly linked to the Parkinson’s disease, but that deals with more general terminology of the healthcare area.

Going forward, we have several improvements planned for the tool. Along with the changes planned for the lists of terms and of complex words explained above, we are also studying, for instance, the possibility of expanding the identification of complex words to some of those for which we currently don’t have a simpler suggestion, for it might help the user to identify possible challenges for their target audience. The changes are not only planned for the backend, but also for the interface. By presenting a more visually appealing interface (for instance, without the presentation of suggestions within the text), the tool can be made more suitable for helping health professionals and communicators of the health industry in their tasks of writing texts for people with low literacy.

7. Acknowledgments

The authors would like to thank these funding sources that contributed to this research: Expanding Excellence in England (E3), Google LARA Award 2019, PIBIC-PROPESQ-UFRGS, CNPq, and CAPES.

8. Bibliographical References

Aluísio, S. M. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics.

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., and Fortes, R. P. (2008). Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. Antártida/Empúries.

Cambricoli, F. (2019). Brasil lidera aumento das pesquisas por temas de sa[ude no google.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Coster, W. and Kauchak, D. (2011). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of Text-To-Text Generation, ACL Workshop*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gasperin, C., Maziero, E., Specia, L., Pardo, T. A., and Aluísio, S. M. (2009). Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *Proc. of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401.

Gazzola, M., Leal, S. E., and Aluisio, S. M. (2019). Prediç ao da complexidade textual de recursos educacionais abertos em português.

Krieger, M. d. G. and Finatto, M. J. B. (2004). *Introdução à terminologia: teoria e prática*. Editora Contexto.

Leal, S. E., de MAGALHAES, V., Duran, M. S., and Aluísio, S. M. (2019). Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. In *Embrapa Gado de Leite-Artigo em anais de congresso (ALICE)*.

Lima, A. and Catelli Jr, R. (2018). Inaf brasil 2018: Resultados preliminares. ação educativa/instituto paulo montenegro, 2018.

Osborne, H. (2005). *Health Literacy from A to Z. Practical Ways to Communicate Your Health Message*. Jones and Bartlett Publishers.

Paetzold, G. H. and Specia, L. (2015). Lexenstein: A framework for lexical simplification. *ACL-IJCNLP 2015*, 1(1):85.

Rieder, C. R. M., Chardosim, N., Terra, N., and Gonzatti, V. (2016). Entendendo a doença de parkinson: Informações para pacientes, familiares e cuidadores. *Aspectos Cognitivos na Doença de Parkinson. Porto Alegre, RS: EDIPUCRS*, 2016:97–104.

Rossetti, A. (2019). Intralingual translation and cascading crises. *Translation in Cascading Crises*.

Saggion, H. (2017). Automatic text simplification: Synthesis lectures on human language technologies, vol. 10 (1). *California, Morgan & Claypool Publishers*.

Scarton, C., Oliveira, M., Candido Jr, A., Gasperin, C., and Aluísio, S. (2010). Simplifica: a tool for authoring simplified texts in brazilian portuguese guided by readability assessments. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 41–44.

Siddharthan, A. (2002). An architecture for a text simplification system. In *Language Engineering Conference*, pages 64–71.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics. Special Issue on Readability and Text Simplification. Peeters Publishers, Belgium*.

Vajjala, S. and Meurers, D. (2014). Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL*, volume 14.

Wagner Filho, J. A., Wilkens, R., Zilio, L., Idiart, M., and Villavicencio, A. (2016). Crawling by readability level. In *International Conference on Computational Processing of the Portuguese Language*, pages 306–318. Springer.

Wilkens, R., Dalla Vecchia, A., Boito, M. Z., Padró, M., and Villavicencio, A. (2014). Size does not matter. frequency does. a study of features for measuring lexical complexity. In *Advances in Artificial Intelligence–IBERAMIA 2014*, pages 129–140. Springer.

Zilio, L., Wilkens, R., and Fairon, C. (2018). Passport: A dependency parsing model for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 479–489. Springer.

9. Language Resource References

Maziero, E. and Pardo, T. (2008). Interface de acesso ao tep 2.0–thesaurus para o português do brasil. *Relatório técnico. University of Sao Paulo*.

Pasqualini, B. and Finatto, M. J. B. (2018). Corpop: a corpus of popular brazilian portuguese. In *Latin American and Iberian Languages Open Corpora Forum - OpenCor*.

Pasqualini, B. F. (2018). *CorPop: um corpus de referência do português popular escrito do Brasil*. Ph.D. thesis, Universidade Federal do Rio Grande do Sul.