Impact of the Text Simplification on Understanding

Oksana IVCHENKO [a] and Natalia GRABAR [a]
[a] *CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France*

Abstract. The reduction of the linguistic complexity of medical texts to make them more understandable to a larger population is an important task. The simplification of texts involves several steps, among which our study focuses on the definition of complex constructions and on study of the impact of the simplification. For this study, we selected 20 texts from the medical domain on different topics, namely drugs, diseases, substances, and medical institutions. We identified complex linguistic constructions and carried out their manual simplification at syntactic, lexical and semantic levels. We then designed a questionnaire to test comprehension of the texts and conducted a study with 26 participants. The results of this study shows that simplified texts obtained higher number of correct answers than technical texts. This difference is statistically significant. The self-evaluation questionnaire, done at the beginning of the test, indicates that the participants tend to overestimate their understanding of medical information. Besides, there is no correlation between the time taken to complete the interview and the correct answers provided.

Keywords. Simplification, Readability, Evaluation, Correlation, France

1. Introduction

The purpose of text simplification is to reduce the linguistic complexity of the content and to make it more understandable for a given population. In the case of medical texts, non-specialist speakers are not familiar with technical terms and expressions, which makes it very difficult to read and understand these texts [1]. Thus, even though medical literature is becoming more freely available online, these documents generally require a level of understanding beyond the capabilities of the average reader [2]. This can lead to a lack of understanding of medical information by patients. Indeed, medicine is currently one of the most rapidly evolving branches of science. There is consequently a growing need for automatic methods to make medical texts more accessible [3]. Yet, before creating automatic simplification systems, it may be useful to know what is the real impact of simplification on comprehension of medical texts.

The simplification of texts usually involves several aspects (detection of complex sequences, lexical and syntactic simplification, evaluation...). The objective of our work is to study where lies the complexity of medical texts and to observe the impact of the text simplification on their understanding. We propose a qualitative study based on interviews with non-specialist speakers.

In what follows, we first introduce the material used and the methods. We then present the results obtained and we discuss them. Finally, we conclude with some issues for future work.

2. Material and Methods

2.1. Source Corpus and Creation of Working Data

The texts studied are randomly selected from the CLEAR corpus [4] and its three genres (encyclopaedic articles, drug leaflets and Cochrane scientific summaries). These texts address diseases (leukaemia, anaemia), drugs (Tracarium), substances (antigens, anti-cholinergics), and institutional facilities. From these, we selected 20 segments with 1 to 3 sentences. Then, we manually identified the complex terms and constructions, and simplified them manually according to three linguistic levels: lexical, syntactic, and semantic.

As has been noticed, lexical complexity of terms has an important impact and, even in short sentences, may prevent the understanding of the whole sentence [5]. Contrary to the general language, short medical terms (such as *ectasia* or *anthrax*) do not make the text more understandable or informative [6]. Hence, we exploit the frequency of words and terms as an indicator of their complexity, as frequent words are read more often. The frequency is provided by *Lexique.org* [7]. Unfrequent and rare terms are replaced by their synonyms or explanations, like *gériatrique* (*geriatric*) replaced by *les personnes âgées* (*[for] elderly people*). These are searched in the *bio-top* resource[1], which is created by medical specialists and qualified professionals. Similarly, the abbreviations are developed using this resource as well, like for *la molécule AMPc* (*the cAMP molecule*) meaning *adénosine monophosphate cyclique* (*cyclic adenosine monophosphate*).

Syntactic simplification involves reducing the grammatical complexity of a text while preserving its information content and meaning [8]. In our sample, we observe that the texts are generally long, with complex subordinate and coordinated sentences. It is therefore necessary to rewrite them into shorter sentences to improve their readability [9]. The syntactic rules implemented can be divided into several types: (1) segmentation of subordinate, coordinated and conjunctive clauses; (2) rewriting of sentences according to a simple word order (Subject-Verb-Complement); (3) transformation of passive sentences into active sentences; (4) modification of negative sentences into positive sentences; (5) preference for the present tense of verbs.

At the semantic level, the most important issue is that the texts remain coherent and clear, and that the information can be understood in its context [10]. We make three additional types of transformations: (1) reorganization of sentences for a better presentation of information, (2) deletion of secondary segments that do not affect the general meaning of the text, (3) addition of examples or explanations for a better understanding.

Indicator	*Technical texts*		*Simplified texts*	
	words	sentences	words	sentences
min	7	1	16	1
max	78	3	75	4
average	33.85	1.25	38.85	2.15

Table 1. Number of words and sentences: minimal, maximal and average values

Table 1 presents the results of the simplification. We can see that simplified texts become longer: they contain more sentences and words. Indeed, the simplification of medical texts often requires addition of information.

[1] bio-top.net/Terminology

2.2. Creation of the Questionnaire and the Interviews

To evaluate the comprehension of technical and simplified texts, we create a questionnaire with multiple choice questions (MCQs). We use different types of questions: definitional, factual, with requests for precision or description and thus we formulated 40 questions. The comprehension of each segment is addressed with two questions: one question on the beginning of the text and one on the end of the text. Four responses are proposed for each question in random order (one correct answer, two wrong answers (distractors), and *I don't know*).

We did a pre-test of the questionnaire with two participants. Following the pre-test, we did not notice any ambiguities in the questions or in the answers. However, we modified three questions to have a greater variety in the types of questions.

At the beginning of the interviews, the participants had to complete a self-evaluation test HLS-EU16 that focuses on the overall understanding of medical information [11]. In this self-evaluation form, the participant indicates, on a scale going from very easy (1) to very difficult (5), how easy it is for him/her to understand medicine-related information. For example, how well the participant understands what the doctor tells him/her, what the results of laboratory tests show, etc. Then, the participants had to answer the main questionnaire after reading each text. The answers collected are analyzed statistically.

3. Results and Discussion

A total of 26 volunteers took part in the study. Their native language was French and they lived in different parts of France as well as abroad. They had no medical education. The correct answers collected for all segments are presented in Figure 1. We obtain on average 16.8 (median 18.5) correct answers for the technical texts (in red), and 22.3 (median 23) for the simplified texts (in blue). The participants spent an average of 14.19 minutes answering the questionnaire.

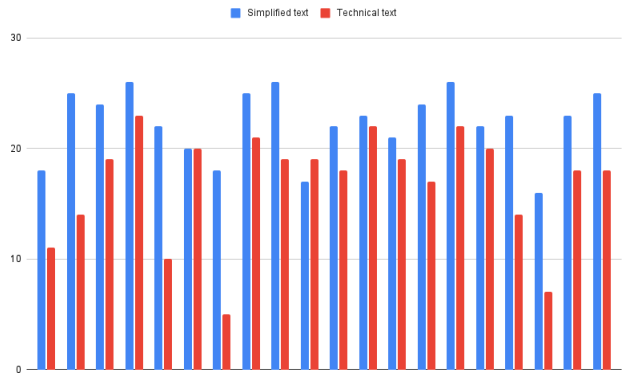

Figure 1. Correct answers collected for each segment for technical (red) and simplified (blue) texts

To test the hypothesis H0 that the observed random variable follows the theoretical distribution law, we conducted a Chi2 test. The analysis of the responses with the Chi2 test shows that there is a statistical significance in the comprehension between the two versions of the texts with $p < 0.00001$. Therefore, this indicates that simplified texts are indeed easier to understand. The analysis for each pair of technical and simplified texts by Student's t-test shows that $p = 0.00009$.

This value is statistically significant and also indicates that the medical texts are easier to understand after simplification. However, if we look at the pairs of texts individually, 6 out of 20 pairs show the p value >0.05. Hence, the p is not statistically significant here, which indicates that some simplified versions of texts are not easier to understand than their original versions. This may be due to the complexity of information, even after simplification, or to the fact that information is added during the simplification, which may lead to sentences syntactically more complex.

Figure 2 shows the Pearson correlation test between the self-evaluation and the number of correct answers. The r-value is -0.4654, which indicates that participants tend to overestimate their ability to understand medical information in everyday life.

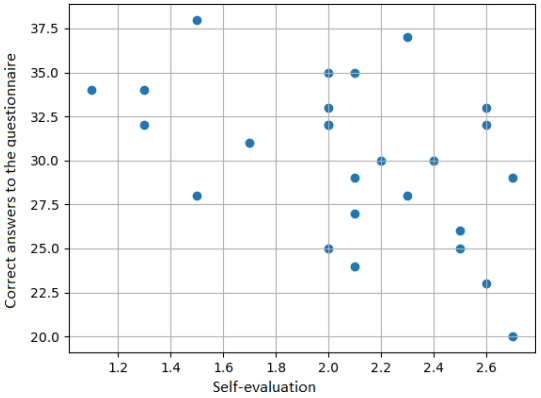

Figure 2. Correlation between self-evaluation and correct answers from participants

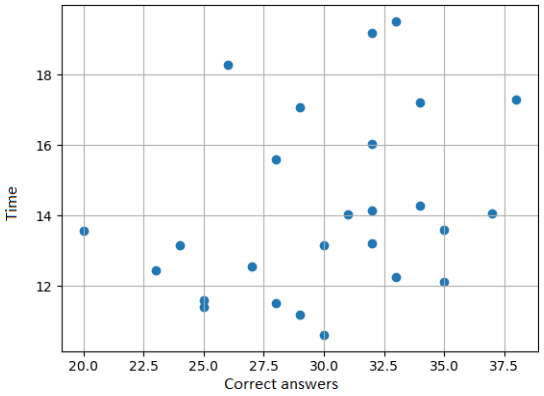

Figure 3. Correlation between correct answers and the time required for completion

Figure 3 shows a weak correlation (r = 0.3053) between the number of correct answers and the time taken to answer the questionnaire. At the same time, the p-value (0.12) does not confirm the statistical significance of this analysis. Therefore, we cannot say that participants who spent more time reading the texts and answering questions have a higher probability of providing correct answers.

We considered all the modifications at once (lexical, syntactic and semantic) and, for this reason, cannot indicate which simplification level is the most salient. Yet, we can state that modifications at several levels are required.

4. Conclusion and Perspectives

Our study shows that medical texts become more comprehensible after simplification, and this result is statistically significant. We have manually simplified the texts on three linguistic levels, and each text tended to be simplified on several levels at once. We performed manual simplification because the aim of the experiment was to determine which features in medical texts to consider when simplifying in order to optimize further work on automatic simplification. So we cannot say exactly which linguistic constructions (syntactic, lexical or semantic) make the text difficult to understand. Yet, we can state that the simplification should cover several levels of simplification. The results of the self-evaluation of comprehension of medical texts show that people tend to overestimate their knowledge. This prompts a search for more objective methods of identifying difficult segments in technical texts. For instance, we can use eye-tracking techniques paying particular attention to the use of large amounts of data and a more important number of participants.

Acknowledgements. This work was partly funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*), ANR-17-CE19-0016-01.

References

[1] Paetzold GH, Specia L. A Survey on Lexical Simplification. J Artif Int Res. 2017 Sep;60(1):549–593. https://doi.org/10.1613/jair.5526

[2] Kloehn N, Leroy G, Kauchak D, Gu Y, Colina S, Yuan NP, et al. Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish. J Med Internet Res. 2018.https://www.jmir.org/2018/8/e10779/

[3] Institute of Medicine. Health Literacy: A Prescription to End Confusion. Nielsen- Bohlman L, Panzer AM, Kindig DA, editors. Washington, DC: The National Academies Press; 2004.Availablefrom:https://www.nap.edu/catalog/10883/ health-literacy-a-prescription-to-end-confusion.

[4] Grabar N, Cardon R. CLEAR – Simple Corpus for Medical French. In: Workshop on Automatic Text Adaption (ATA); 2018. p. 1-11.

[5] Collet T. Obstacles lexico-sémantiques à la lecture réussie d'un texte de spécialité. TTR. 2014;27(1):123-48. https://doi.org/10.7202/1037121ar

[6] Gu Y, Leroy G, Kauchak D. When synonyms are not enough: Optimal parenthetical insertion for text simplification. AMIA Annual Symposium proceedings AMIA Symposium. 2017;2017:810-9. https://pubmed.ncbi.nlm.nih.gov/29854147/

[7] New B, Pallier C, Brysbaert M, Ferrand L. Lexique 2 : A new French lexical database. Behavior Research Methods, Instruments & Computers. 2004;36(3):516-24. https://pubmed.ncbi.nlm.nih.gov/15641440/

[8] Siddharthan A. Syntactic simplification and Text Cohesion. No. 597 in Technical Reports. University of Cambridge; 2004.http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-597.pdf

[9] Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics. 2016;4:401-15. https://aclanthology.org/Q16-1029.pdf

[10] Brouwers L, Bernhard D, Ligozat AL, François T. Simplification syntaxique de phrases pour le français (Syntactic Simplification for French Sentences) [in French]. In: JEP-TALN-RECITAL; 2012. https://aclanthology.org/F12-2016

[11] Rouquette, T N, P L, den Broucke S et al V. Validity and measurement invariance across sex, age, and education level of the French short versions of the European Health Literacy Survey Questionnaire.; 2018. https://pubmed.ncbi.nlm.nih.gov/30521552/