An automated tool with human supervision to adapt difficult texts into Plain Language

Paul Poupet
U31
https://u31.io/
paul.poupet@u31.io

Morgane Hauguel
U31
https://u31.io/
morgane.hauguel@u31.io

Erwan Boehm
U31
https://u31.io/
erwan@u31.io

Charlotte Roze
U31
https://u31.io/
charlotte.roze@u31.io

Paul Tardy
U31
https://u31.io/
pltrdy@gmail.com

Abstract

In this paper, we present an automated tool with human supervision to write in plain language or to adapt difficult texts into plain language. It can be used on a web version and as a plugin for Word/Outlook plugins. At the publication date, it is only available in the French language. This tool has been developed for 3 years and has been used by 400 users from private companies and from public administrations. Text simplification is automatically performed with the manual approval of the user, at the lexical, syntactic, and discursive levels.

Screencast of the demo can be found at the following link: https://www.youtube.com/watch?v=wXVtjfKO9FI.

Keywords : text simplification, Plain language, French, automated tool, simplification tool

1 Introduction

Understanding textual information is a societal issue. The lack of clarity in textual content is an essential dimension of the accessibility issue, in both the physical and digital worlds. It is essential for everyone's access to goods and services, assistance and rights.

1.1 Reading difficulties

16% of the population encounters difficulties to read and understand common textual information of their daily life (INSEE 2012). The right to accessible information is a fundamental right that should be granted to all people (UN, 2020). It is the key factor of personal empowerment and social inclusion. Nevertheless, textual information found on the web, in the news, health leaflets, and other sources is often so linguistically complex it can impede their active participation in the society.

1.2 Plain language

Plain Language is one of the standard standards that aims at providing texts that can be more easily understood by people, especially those who experience difficulties to read and understand. In 2023, ISO-24495 established governing principles and guidelines for developing Plain Language. Plain Language is mainly about using reduced vocabulary, simple sentences and an easy-to-understand discursive organization.

1.3 Text simplification

In order to make texts more readable while preserving their original content, text simplification operates at different linguistic levels (lexical, morphosyntactic, and discursive). Syntactic simplification consists in reducing the complexity of syntactic structures by deleting or replacing complex constructions. Discursive simplifications address phenomena, such as paragraph splitting or reordering, explicitness of coreference chains, anaphora resolution, creation of titles.

2 An automated tool with human supervision

We have developed a tool to help people create plain language texts by using different text simplification techniques. A complexity score is computed for any given text input. The score ranges from 0 to 100, with 0 meaning the text is not clear at all, and 100 meaning it is very clear. Then, the user is presented with an experience/interface similar to a spell and grammar checker: difficult words, sentences or paragraphs are underlined and linked to a suggestion. The user can ignore or accept the suggestion in order to replace the difficult element by a simpler one. Figure 1 shows the interface.

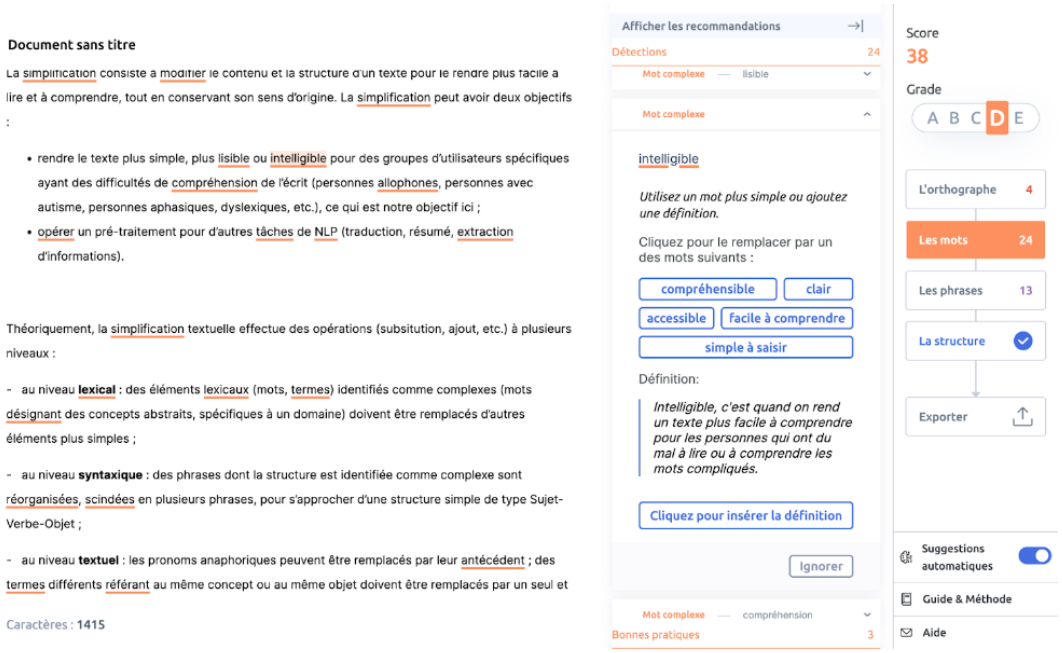

Figure 1: Tool interface

2.1 Scoring

The scoring part has been developed using two French pre-trained language models such as BARThez (Kamal Eddine et al., 2021) and CamemBERT (Martin et al., 2019). The model was then fine-tuned with a classification head as the output layer.

2.2 Complexity identification

The identification of complexity is handled at several levels. Lexical complexity takes into account several parameters: orthograph, appearance rate in a corpus, number of consonants, lists of domain-specific concepts, list of abbreviations and acronyms. At syntactic level, sentence structures are identified as complex with rule-based analysis using SpaCy pipeline based on both dependency and constituency parsing. At textual level, a ruled-base analysis is made to detect anaphora, coreference difficulties, discursive pronouns, length of paragraphs, presence of titles.

2.3 Substitution by simpler solutions

At lexical level, we use use databases of context-based synonyms with a lower difficulty score and context-based definitions with simple words. We also use a custom LLM approach for synonyms and definition generation.

At syntactic level, we both use a rule-based system and a custom LLM approach for simpler paraphrases generation. Simplification rules modify complex sentences by splitting them into several simpler ones and/or reorganizing them.

3 Conclusion and future work

An automated tool is effective to broadcast the use of Plain Language. Users of our tool are private companies and public administrations. More than 400,000 words have been analyzed at the publication date. Continuous improvements are made at each level. Moreover, federated learning allows the scoring, words difficulties and the LLM to improve themselves. At discursive level, logic or temporal reorganization will be tackled.

4 Lay Summary

We present an automated tool for writing in plain language and adapting difficult texts into plain language. It can be used on a web version and as a plugin for Word/Outlook. At the publication date, it is only available in the French language.

This tool has been developed for 3 years. It has been used by 400 users from private companies and from public administrations. Text simplification is automatic with the manual approval of the user, at word, sentence, and text levels.

Screencast of the demo can be found at the following link: https://www.youtube.com/watch?v=wXVtjfKO9FI.

References

INSEE 2012. 2012. Pour les générations les plus récentes, les difficultés des adultes diminuent à l'écrit, mais augmentent en calcul. Standard, INSEE.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *CoRR*, abs/1911.03894.