A Preliminary Study of ChatGPT for Spanish E2R Text Adaptation

Margot Madina[1], Itziar Gonzalez-Dios[2], Melanie Siegel[1]
[1]Darmstadt University of Applied Sciences (Hochschule Darmstadt)
Hochschule Darmstadt, Max-Planck Str.2, 64807 Dieburg
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
Manuel Lardizabal, 1, 20018 Donostia-San Sebastian, Gipuzkoa
margot.madina-gonzalez@h-da.de, itziar.gonzalezd@ehu.eus, melanie.siegel@h-da.de

Abstract

The process of adapting and creating Easy-to-Read (E2R) texts is very expensive and time-consuming. Due to the success of Large Language Models (LLMs) such as ChatGPT and their ability to generate written language, it is likely to think that such models can help in the adaptation or creation of text in E2R. In this paper, we explore the concept of E2R, its underlying principles and applications, and provides a preliminary study on the usefulness of ChatGPT for E2R text adaptation. We focus on the Spanish language and its E2R variant, *Lectura Fácil* (LF). We consider a range of prompts that can be used and the differences in output that this produces. We then carry out a three-folded evaluation on 10 texts adapted by ChatGPT: (1) an automated evaluation to check values related to the readability of texts, (2) a checklist-based manual evaluation (for which we also propose three new capabilities) and (3) a users' evaluation with people with cognitive disabilities. We show that it is difficult to choose the best prompt to make ChatGPT adapt texts to LF. Furthermore, the generated output does not follow the E2R text rules, so it is often not suitable for the target audience.

Keywords: Easy-to-Read, readability, Large Language Models

1. Introduction

In today's fast-paced and information-rich world, effective communication is essential for successful interaction and knowledge transfer. However, a significant proportion of the population faces challenges in accessing and understanding written information due to various factors such as language barriers, cognitive limitations or learning disabilities. The United Nations reports that 16% of the world's population, which equates to 759 million adults, lack literacy skills[1]. Following the COVID-19 pandemic, the number of children struggling with reading has increased from 460 million to 584 million[2]. Furthermore, in the Organisation for Economic Co-operation and Development (OECD) countries, adult literacy proficiency at the most basic levels varies between 4.9% and 27.7% (OECD, 2013). Additionally, 10% of recent graduates in these countries demonstrate weak literacy abilities (OECD, 2015). Recognising the need for inclusive and accessible communication, Easy-to-Read language (E2R) emerged as a promising approach to bridging these gaps. E2R aims to make content accessible; it is a controlled language variant that employs the basic vocabulary and syntactic structures of a given language, and follows a set of rules regarding its writing, layout and visual aids (Nitzke et al., 2022; Hansen-Schirra and Maaß, 2020).

Even though E2R was primarily created, and it is aimed at people with cognitive disabilities, there are other groups that can also benefit from it. E2R's target groups include, but are not limited to, the following (Bredel and Maaß, 2016; Maaß, 2019; Hansen-Schirra and Maaß, 2020; Maaß and Garrido, 2020):

- People with intellectual or developmental disabilities.
- Groups with cognitive difficulties, such as the elderly, people with aphasia, people with dysphasia, or people with hyperactivity and attention disorders.
- People with auditory disabilities. People with prelocutive and perillocutive deafness (before and at the beginning of speech development).
- People with low literacy.
- Migrants who do not speak the language of the country they live in.
- Children in need of reading reinforcement.

E2R usually receives a different name depending on the standard language it is based on; in Spanish, it is known as *Lectura Fácil* (LF).
Spanish is the second mother tongue in the world in terms of number of speakers, after Mandarin Chinese, and the fourth language in terms of overall number of speakers (native speaker + limited proficiency + learners of Spanish), after English, Mandarin Chinese and Hindi (Instituto Cervantes,

[1] https://www.un.org/en/chronicle/article/education-all-rising-challenge (last accessed 04-10-2023)

[2] https://news.un.org/en/story/2021/03/1088392 (last accessed 04-10-2023)

LREC-COLING 2024, pages 1422–1434

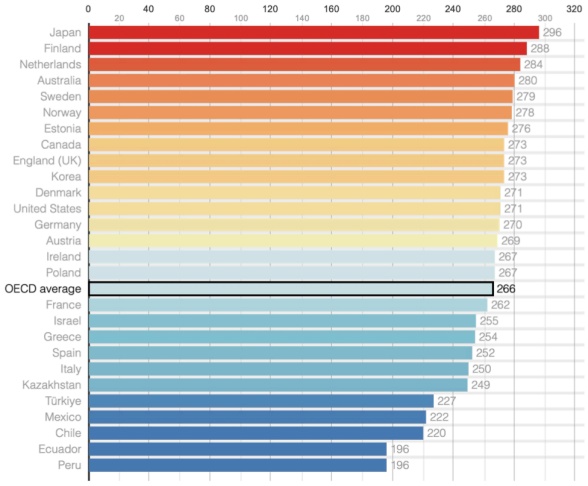

Figure 1: Mean literacy score in the Survey of Adult Skills (PIAAC) (aged 16-65)

2022). In spite of this, it is worth highlighting that the results obtained by the Programme for the International Assessment of Adult Competencies (PIAAC) showed that Spanish-speaking countries scored very low for literacy skills. Specifically, the average literacy score for the OECD countries participating in the assessment is 266 points, while the literacy score for Spain is 252 points, followed by Mexico (222), Chile (220), Ecuador (196) and Peru (196) (see Figure 1). Moreover, according to the *Imserso* (Institute for the Elderly and Social Services in Spain) data, in Spain there are 283,256 people with a recognised intellectual disability of 33% or more (Imserso, 2022). E2R language variants such as LF can be very valuable tools in addressing these issues, as they seek to ensure the participation of the whole population in society.

There is a very limited number of E2R texts, and the existing ones have been created manually; however, this approach is very costly, both in terms of time and money (Drndarević et al., 2013). These texts are either created from scratch or adapted from a given source text written in standard language. When we talk about E2R adaptation, we are referring to the processes that standard texts go through in order to be transformed into E2R texts. These processes may include the creation of auditory and/or visual aids and examples, elaboration (adding explanations to the concepts), syntactic simplification, lexical simplification and summarisation. In cases like news texts, which are only relevant for a limited period of time, it is very hard to keep pace with the upcoming content (Bott et al., 2012; Drndarević et al., 2013). The existence of an automatic adaptation tool would allow standard language texts to be adapted to E2R quickly and efficiently. This would promote accessibility, inclusivity and effective communication in a variety of settings.

ChatGPT is an advanced Large Language Model (LLM) developed by OpenAI that is able to engage in dynamic and coherent conversations with users[3]. As conversations around ChatGPT continue to permeate different spheres, it is natural to consider its potential as a solution to the problem at hand. In this paper, we explore the usefulness of ChatGPT for the adaptation process of standard texts into LF. We adapt 10 standards news texts[4] into LF using ChatGPT, and we assess their suitability by means of three kinds of evaluations: we automatically analyse their linguistic aspects with *MultiAzterTest* (Bengoetxea and Gonzalez-Dios, 2021), we manually check whether they fulfil the capabilities for simplified texts proposed by Cumbicus-Pineda et al. (2021), and finally, we carry out an evaluation with a target group, namely people with cognitive disabilities. Our main objective is to shed light on ChatGPT's shortcomings, emphasizing that its outcomes, though seemingly trustworthy at first, might not consistently align with users' aims or anticipations. As technology continues to seep into our everyday lives, there is a vital necessity to assess and design these tools with not just an academic or engineering perspective, but also with a profound understanding of the end-user experience. While experts may evaluate technologies with great detail and precision, a significant portion of the population relies on them without sufficient knowledge or ability to determine their veracity. Therefore, in this study we place a deliberate emphasis on the user perspective, underlining the importance of making technologies both accessible and reliable for the general public.

The paper is structured as follows: in section 2, we focus on the LF text adaptation process. In section 3 we introduce our experimental setting and explain the evaluation methods chosen for this study. In section 4 we review the results we obtained. We discuss the results in section 5. Finally, we present our conclusions and future work in section 6.

2. E2R document adaptation

In this section we explain the process of adapting E2R documents and how ChatGPT can help.

The Spanish standard *Norma UNE 153101:2018 EX de Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos* (UNE 153101 EX of Lectura Fácil. Guidelines and recommendations for the elaboration of documents) (UNE,

[3] Introducing ChatGPT https://openai.com/blog/chatgpt (last accessed 12-10-2023)

[4] All three versions of the texts (Original, LF and ChatGPT) are available in https://github.com/margotmg/ChatGPT-for-Spanish-E2R-Text-Adaptation.git

2018) [5] aims to guarantee the understanding of written documents and the right of all people to access information, as it is a necessary means of exercising full citizenship. This standard also describes the process for adapting a document to LF. This process involves the figures of the adaptor (a person who makes a document LF), designer and/or layout designer (a person who designs and/or typesets the document in LF, bringing their creativity to the document and respecting the guidelines and recommendations for the drafting and design of documents in LF), facilitator (person managing the validation phase) and validators (end-users with reading comprehension difficulties who participate in the validation phase, who can read and have communication skills). According to the aforementioned UNE standard (UNE, 2018), the adaptation process consists of an adaptation phrase and a validation phase (see Figure 2):

- During the adaptation phase, the original document is adapted by the adaptor, with the participation of the designer and/or layout designer. They provide the facilitator with the LF draft and, if appropriate, also with the original document.
- During the validation phase, the facilitator organizes and leads the validation sessions with the validators. Validators who participate in these sessions should read the document and indicate difficulties of understanding, among others. The facilitator should collect this input and produce a report for the adaptor/designer/layout designer with the modification suggestions. The adaptor, designer and layout designer should incorporate the contributions mentioned in the report into the first draft. This should be repeated until validators confirm that the LF draft is comprehensible.

As it can be observed from the number of steps described above, the process of adapting a standard text to LF is very laborious and costly, both in terms of time and money[6]

2.1. ChatGPT for E2R text adaptation

ChatGPT is an intelligent chat machine developed by OpenAI and constructed upon the *InstructGPT* model, which is trained to follow an instruction in a prompt and provide a detailed response (Ouyang et al., 2022). ChatGPT is trained using Reinforcement Learning from Human Feedback (RLHF) and it is capable of answering follow-up questions, admitting its mistakes, challenging incorrect premises, and rejecting inappropriate requests[7]. ChatGPT is able to respond to a prompt in a matter of seconds. The latest version of ChatGPT is powered by GPT-4, the most recent LLM developed by OpenAI. GPT-4 outperforms both previous LLMs (namely GPT-3 and GPT-3.5) and most state-of-the-art systems on a suite of traditional NLP benchmarks (OpenAI, 2023). We opted for the ChatGPT version in our experiments[8] and decided to employ its web interface. This choice

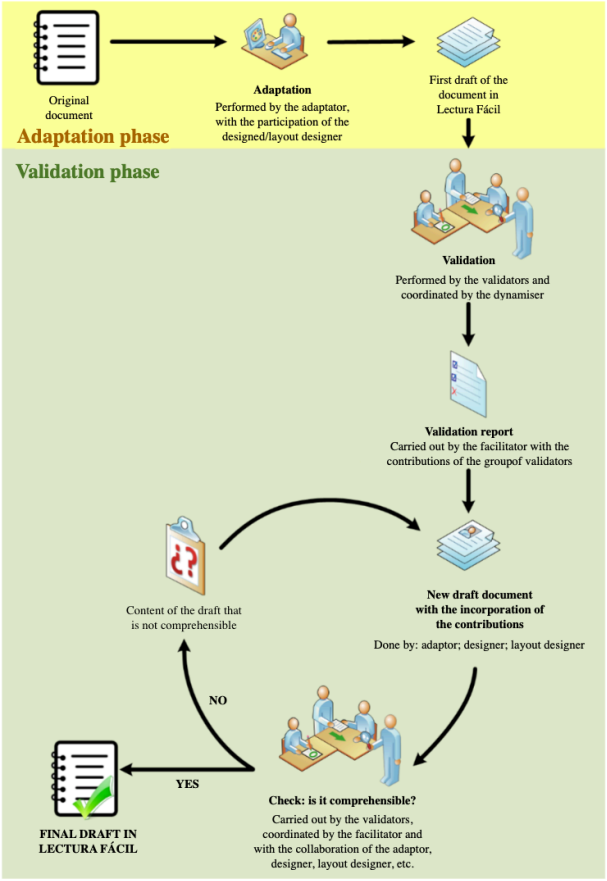

Figure 2: LF adaptation process according to the UNE standard ((UNE, 2018), own translation into English)

[5] UNE https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060036 (last accessed 17-10-2023)

[6] This UNE standard also describes the creation process of LF texts, which differs from the adaptation process we have described. This paper explores the usability of ChatGPT for the adaptation process of texts into LF; the creation process of LF texts is beyond the scope of this paper, and it is therefore not explained.

[7] Introducing ChatGPT https://openai.com/blog/chatgpt (last accessed 12-10-2023)

[8] August 3rd, 2023 version of ChatGPT-4 is used for the experiments. It should be noted that future updates might potentially impact the outcomes presented in this paper. Details of this version are provided here: https://help.openai.com/en/articles/6825453-chatgpt-release-notes (last accessed 12-10-2023)

was based on the observation that many individuals use ChatGPT without a full comprehension of its dependability, and we wanted to mimic the user perspective as much as possible. Moreover, it is apparent that some users lack proficiency with the API, resulting in their use of the website for various purposes, such as Machine Translation (MT), Question & Answer (Q&A) sessions, or E2R text adaptations, among others. Our primary goal was to illuminate ChatGPT's limitations, highlighting that its results, while initially appearing credible, may not always correspond with users' intentions or expectations.

Some experiments, contemporaneous with our research, have been undertaken for German E2R, *Leichte Sprache* (Deilen et al., 2023), and showed that that ChatGPT generated texts are easier than standard ones, but do not meet the E2R criteria. With regard to LF, ChatGPT is currently being used to aid in the creation of the CLEARSIM Corpus, as part of the ClearText project (Espinosa-Zaragoza et al., 2023).

3. Experimental setting

In this section, we delve into the description of the experiments conducted in this study and the evaluation methods followed in our experiments.

3.1. Data

Irekia[9] is the open-government communication channel of the Basque Government. This website is based on the three main axes of open government: transparency, participation and collaboration. The portal has CC-BY licence, which has been chosen to provide legal coverage for this principle of open government whereby citizens can access the wealth of information, audiovisual and multimedia material generated in this space, which can be used freely for any purpose as long as the source and author are cited. The webpage contains news regarding the Government, among others. They are written in a non-administrative language, both in Spanish and Basque, as they are the official languages of the Basque Country. Some of these texts have also been adapted into LF[10]. Having their LF enables us to compare them against the ChatGPT output. In order to avoid data contamination from earlier years, we randomly selected 10 texts from 2023 to serve as our test texts, all in their original Spanish version. Each of them deals with different topics, but all of them belong to the journalism genre and offer an account of events.

[9] *Irekia* https://www.irekia.euskadi.eus (last accessed 04-10-2023)

[10] *Irekia* LF https://www.irekia.euskadi.eus/lf?locale=es (last accessed 04-10-2023)

3.2. Prompts

The design of appropriate prompts is pivotal when using LLMs for specific tasks. In fact, it has been shown that different prompt designs can lead to large variance in performance (Lu et al., 2022; Perez et al., 2021). However, LLMs are non-deterministic in nature. This means that semantically similar or even identical inputs may lead to different outputs. This makes it difficult to choose a prompt that performs well in all situations. Non-determinism can be an advantage, as it offers creative and diverse outputs, but for specific tasks such as the one at hand, this can also cause inconsistency. In an aim to get the best result as possible, we tried different prompts into ChatGPT (a list of prompts is provided in the Appendix A). The steps followed with each prompt were the following:

1. Design prompt.
2. Try the prompt out 3 times with different texts. This allowed us to see whether the output generated by ChatGPT was consistent and followed the LF rules (at first glance). In an aim to avoid possible biasing of the output, we conducted these trials with other texts from *Irekia* that were not part of the selected 10 test texts.
3. Use prompt with test texts. The texts that are included in this study were only tested once, and we took the first output generated by ChatGPT.

We finally opted for the following prompt:

prompt: *Las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil son las siguientes: (lista de normas)*
The rules for writing in Lectura Fácil according to the UNE 153101:2018 EX Lectura Fácil Standard are as follows: (list of rules)
output: ...
prompt: *Adapta esta noticia a Lectura Fácil siguiendo las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil: (noticia)*
Adapt this news to Lectura Fácil by following the UNE 153101:2018 EX Lectura Fácil Standard: (news)

The rules used in our prompt were adapted from the UNE 153101:2018 EX Lectura Fácil Standard. As it has been mentioned, E2R and LF rules cover text, layout and pictures. For our experiments, we focused on the (1) guidelines and recommendations related to orthotypography, (2) guidelines

and recommendations related to vocabulary and expressions, (3) guidelines and recommendations related to phrases and sentences, (4) guidelines and recommendations related to text organization and style, and (5) guidelines and recommendations related to the presentation of the document. This made a total of 85 rules out of the 130 that compose the UNE Standard, leaving out those rules related to pictures and paratextual complements such as graphics and glossaries. On the other hand, most rules tend to use modal verbs (for example, "The use of abstract, technical or complex terms should be avoided"); we changed the phrases so that the imperative was used instead ("Avoid using abstract, technical or complex terms"). We used this prompt and took the first output generated by ChatGPT.

3.3. Evaluation

There are three main criteria that are taken into account in the Automatic Text Simplification (ATS) evaluation process: grammar, meaning preservation and simplicity. The most commonly used automatic evaluation methods for ATS are BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) and SARI (System output Against References and Input sentence) (Xu et al., 2016). Although these automatic methods can be more objective and faster than human evaluation, they are both flawed (Alva-Manchego et al., 2020). BLEU is reliable for Machine Translation, but not for other tasks such as ATS, due to its negative correlation with simplicity (Sulem et al., 2018). SARI has limited usefulness as it can only serve as an approximation for assessing the gain in simplicity, focusing mainly on lexical simplifications and minor reordering while disregarding other potential text transformations (Alva-Manchego et al., 2021). On the other hand, E2R adaptation also implies syntactic simplification; this might increase the distance between the source and the adapted texts, leading to lower scores (Grabar and Saggion, 2022). We should bear in mind that the E2R adaptation process implies not only simplification, but also summarization, among others. ROUGE (Lin and Hovy, 2003; Lin, 2004) is a common evaluation metric for summarization, but it is not used in ATS. There is currently no automatic evaluation metric that covers all LF or E2R criteria; therefore, it is usually necessary to combine different metrics (Grabar and Saggion, 2022).

The manual evaluation procedure usually assesses grammar, meaning preservation and simplicity on a Likert scale from 1 to 5. However, this is very costly and it may result in subjective evaluations. Another important aspect to take into account is that the evaluation must be user-dependent and consider the needs of final users (Grabar and Saggion, 2022). In the case of E2R, it is important to include validators in the text creation or adaptation process.

Due to the lack of a standard evaluation method and in an aim to provide a more rounded perspective on our results, we decided to conduct a three-folded evaluation. We have employed 2 new forms of evaluation that have not been used so far to evaluate LF, as well as a user evaluation:

- **Automatic evaluation:** we used *MultiAzterTest* (Bengoetxea and Gonzalez-Dios, 2021), a multilingual NLP tool that analyses texts on over 125 measures of cohesion, language, and readability, and it can also be used for text analysis, profiling or stylometrics. It obtains 90% in accuracy when classifying into two reading levels (simple and complex) in Spanish. *MultiAzterTest* is also available as web service[11]. Among the values it analyses, we focused on the number of words, the number of sentences, the number of rare lexical words, the number of subordinate clauses, and the number of propositions.

- **Checklist-based manual evaluation:** we performed a checklist-based evaluation method following the guidelines provided by Cumbicus-Pineda et al. (2021). This checklist-based method evaluates the linguistic capabilities ATS systems need to meet. The capabilities they cover are grammar/fluency, meaning preservation, simplicity, prerequisites (a set of basic checklists that any simplification should comply with) and ethical aspects (they measure ethical issues that may arise due to the simplifications). For each sentence, the evaluator indicates whether a capability has been fulfilled with a binary score (1:yes, 0:no). Then, the percentage of the positive scores from each capability is calculated; the higher the percentage, the better the results: 96-100 % perfect, 81-95 % substantial, 61-80 % moderate and < 60 % low. The evaluation was performed on a document level by two native Spanish speakers with knowledge on LF. We also calculated the Cohen's kappa to measure the inter-annotator agreement. We decided to conduct this evaluation on a text level because (1) the adaptation process is performed on a text level and (2) there is not always a 1-to-1 sentence alignment in standard and LF texts. On the other hand, we noticed that there were no capabilities that took the layout of the text into account;

[11] *MultiAzterTest* http://ixa2.si.ehu.eus/aztertest/ (last accessed 02-10-2023)

therefore decided to include 3 new capabilities in relation to the text: (1) Format is easier to read, (2) Line spacing is bigger and (3) Phrases are not split into lines (a list of all the capabilities is provided in Appendix B).

- **Users' evaluation:** as previously mentioned, the evaluation with the target audience is key in E2R and LF text creation and adaptation processes. We conducted users' evaluations with 10 people with a disability percentage ranging between 33% and 79%. Each participant was randomly given a text in the ChatGPT and LF versions. They were asked to retell what they read and answer some questions about the content. Finally, they were asked to pick the version that they preferred. The procedure was inspired by Reichrath and Moonen (2022).

The automatic evaluation with *MultiAzterTest* enables a comparison between the values mentioned in the original, the LF and the ChatGPT versions of the texts. The checklist-based evaluation allows us to obtain quite detailed information on how the simplification process has been carried out and whether all essential information has been kept in the ChatGPT version. Finally, the users' evaluation allows us to know what potential future users prefer.

4. Results

In this section, we present the results drawn from our experiments.

4.1. Automatic evaluation

The results of the automatic evaluation are presented in Table 1. The key findings of this analysis are the following:

- The ChatGPT generated texts are the shortest ones (except for Text 6, which is only one word longer than the LF version).
- LF versions contain the highest number of sentences and also the shortest ones.
- Regarding rare words, the ChatGPT generated texts contain the lowest mean distinct rare lexical words (except for Text 6, in which the LF version exceeds in this aspect).
- The subordinate clauses per 1000 words is always the lowest in the LF versions.
- The mean of the number of propositions per sentence is always the lowers in the LF versions.

With this information at hand, we can say that ChatGPT performs a good lexical simplification and a deep summarisation of the texts. However, the sentences are longer and with more subordinate clauses and propositions; this can make reading and understanding difficult.

4.2. Checklist-based manual evaluation

The results for the checklist-based manual evaluation are shown in Table 2. On average, the ChatGPT generated texts obtained a 66.83%, which is a moderate result. The average agreement between the 2 evaluators was 0.85, which is an almost perfect agreement; the capabilities on which the evaluators disagreed most were those related to simplicity. Through this analysis, we observed that although the used prompt was always the same, the way ChatGPT adapted the texts was different. For 7 of the texts, it created subsections with subtitles. This would make sense for Text 2, as it deals with 3 different topics. However, it is neither necessary nor helpful for the other texts. When creating these subtitles, relevant information was sometimes included in the subtitles, but not in the text. For example, in Text 3, one of the subtitles reads *¿Por qué On-giz ha ganado el premio?* (Why did On-giz get the prize?), but no reference to the prize was made before. In this case, an introductory phrase like "The On-giz association got a prize" is necessary. For Text 5, ChatGPT enumerated all the ideas, instead of writing a coherent text.
Regarding the grammar, the texts were well-written. Nonetheless, the content was very summarised, and a lot of important information was lost in the adaptation process, making the ChatGPT versions very vague. On the other hand, the register was altered, making the texts more informal. Another important aspect to take into account was the splitting of sentences; ChatGPT split phrases in different lines on several occasions, which makes the reading process more difficult.

4.3. Users' evaluation

The results of the users' evaluation are shown in Table 3. As it can be observed, four participants were able to retell both the ChatGPT and the LF versions and answer the questions. One participant was able to retell the LF version but not the ChatGPT version, but was not able to answer the questions. One participant was able to answer the questions about the texts, but not retell them. Finally, there were four participants who could neither retell what they had read nor answer the questions. All participants claimed that they preferred the LF version of the text they were given. The reasons they gave were the following:

Text	Text 1			Text 2			Text 3			Text 4			Text 5		
Text version	Original	LF	ChatGPT	Original	LF	ChatGPT	Original	LF	ChatGPT	Original	LF	ChatGPT	Original	LF	ChatGPT
Words (total)	811.0	425.0	**175.0**	721.0	454.0	**179.0**	867.0	621.0	**237.0**	263.0	199.0	**156.0**	265.0	251.0	**218.0**
Distinct words	275.0	142.0	**85.0**	306.0	192.0	**94.0**	307.0	205.0	**98.0**	147.0	105.0	**81.0**	159.0	133.0	**110.0**
Distinct words (per 1000 words)	339.0875	**334.1176**	485.7143	424.4105	**422.9075**	525.1397	354.0946	**330.1127**	413.5021	558.9354	527.6382	**519.2308**	600.0	529.8805	**504.5872**
Sentences (total)	21.0	**60.0**	14.0	21.0	**60.0**	20.0	20.0	**88.0**	27.0	8.0	**28.0**	16.0	21.0	**33.0**	24.0
Sentences (per 1000 words)	25.894	**141.1765**	80.0	29.1262	**132.1586**	111.7318	23.0681	**141.7069**	113.9241	30.4183	**140.7035**	102.5641	79.2453	**131.4741**	110.0917
Words (length) in sentences (mean)	38.619	**7.0833**	12.5	34.3333	**7.5667**	8.95	43.35	**7.0568**	8.7778	32.875	**7.1071**	9.75	12.619	**7.6061**	9.0833
Rare nouns	21.0	4.0	**0.0**	11.0	3.0	**0.0**	44.0	11.0	**10.0**	3.0	1.0	**0.0**	8.0	5.0	**0.0**
Rare nouns (per 1000 words)	25.894	9.4118	**0.0**	15.2566	6.6079	**0.0**	50.7497	**17.7134**	42.1941	11.4068	5.0251	**0.0**	30.1887	19.9203	**0.0**
Rare adjectives	30.0	**2.0**	3.0	22.0	8.0	**1.0**	25.0	4.0	**0.0**	5.0	3.0	**0.0**	4.0	3.0	**1.0**
Rare adjectives (per 1000 words)	36.9914	**4.7059**	17.1429	30.5132	17.6211	**5.5866**	28.8351	6.4412	**0.0**	19.0114	15.0754	**0.0**	15.0943	11.9522	**4.5872**
Rare verbs	14.0	9.0	**0.0**	21.0	12.0	**5.0**	19.0	17.0	**2.0**	8.0	6.0	**1.0**	9.0	**7.0**	**7.0**
Rare verbs (per 1000 words)	17.2626	21.1765	**0.0**	29.1262	**26.4317**	27.933	21.9146	27.3752	**8.4388**	30.4183	30.1508	**6.4103**	33.9623	**27.8884**	32.1101
Rare adverbs	1.0	**0.0**	**0.0**	1.0	**0.0**	**0.0**	1.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0	**0.0**	**0.0**
Rare adverbs (per 1000 words)	1.233	**0.0**	**0.0**	1.387	**0.0**	**0.0**	1.1534	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	3.7736	**0.0**	**0.0**
Rare content words	66.0	15.0	**3.0**	55.0	23.0	**6.0**	89.0	32.0	**12.0**	16.0	10.0	**1.0**	22.0	15.0	**8.0**
Rare content words (per 1000 words)	81.381	35.2941	**17.1429**	76.2829	50.6608	**33.5196**	102.6528	51.5298	**50.6329**	60.8365	50.2513	**6.4103**	83.0189	59.761	**36.6972**
Distinct rare content words	53.0	11.0	**3.0**	51.0	21.0	6.0	61.0	25.0	4.0	16.0	10.0	1.0	21.0	13.0	7.0
Distinct rare content words (per 1000 words)	65.3514	25.8824	**17.1429**	70.7351	46.2555	**33.5196**	70.3576	40.2576	**16.8776**	60.8365	50.2513	**6.4103**	79.2453	51.7928	**32.1101**
Mean of rare lexical words	19.5846	8.2873	**3.1579**	20.5224	13.6905	**7.6923**	23.2984	11.8081	**10.7143**	14.9533	12.1951	**1.3514**	17.0543	15.4639	**7.6923**
Mean of distinct rare lexical words	26.5	11.9565	**5.8824**	23.1818	16.5354	**9.8361**	25.8475	17.6056	**7.4074**	17.5824	15.625	**2.0408**	20.0	16.4557	**0.6061**
Subordinate clauses	28.0	**8.0**	16.0	32.0	9.0	**8.0**	39.0	26.0	**17.0**	14.0	**9.0**	15.0	15.0	**12.0**	17.0
Subordinate clauses (per 1000 words)	34.5253	**18.8235**	91.4286	44.3828	**19.8238**	50.2793	44.9827	**41.868**	71.73	53.2319	**45.2261**	96.1538	56.6038	**47.8088**	77.9817
Mean of the number of propositions per sentence	3.5238	**1.2833**	2.3571	3.7619	**1.3167**	1.6	5.85	**1.5682**	2.0741	4.875	**1.5357**	2.375	2.2857	**1.6061**	1.9167

Text	Text 6			Text 7			Text 8			Text 9			Text 10		
Text version	Original	LF	ChatGPT	Original	LF	ChatGPT	Original	LF	ChatGPT	Original	LF	ChatGPT	Original	LF	ChatGPT
Words (total)	391.0	**200.0**	201.0	872.0	506.0	**238.0**	1141.0	534.0	**250.0**	547.0	297.0	**157.0**	795.0	440.0	**155.0**
Distinct words	160.0	101.0	**85.0**	387.0	223.0	**120.0**	425.0	218.0	**126.0**	227.0	143.0	**80.0**	312.0	194.0	80.0
Distinct words (per 1000 words)	**409.2072**	505.0	422.8856	443.8073	**440.7115**	504.2017	**372.4803**	408.2397	504.0	**414.9909**	481.4815	509.5541	**392.4528**	440.9091	516.129
Sentences (total)	12.0	**29.0**	22.0	30.0	**74.0**	29.0	33.0	**61.0**	23.0	16.0	**38.0**	13.0	23.0	**64.0**	16.0
Sentences (per 1000 words)	30.6905	**145.0**	109.4527	30.9633	**146.2451**	121.8487	28.922	**114.2322**	92.0	29.2505	**127.9461**	82.8025	28.9308	**145.4545**	103.2258
Words (length) in sentences (mean)	32.5833	**6.8966**	9.1364	32.2963	**6.8378**	8.2069	34.5758	**8.7541**	10.8696	34.1875	**7.8158**	12.0769	34.5652	**6.875**	9.6875
Rare nouns	10.0	3.0	**0.0**	24.0	8.0	**1.0**	25.0	7.0	**2.0**	14.0	3.0	**0.0**	36.0	13.0	**3.0**
Rare nouns (per 1000 words)	25.5754	15.0	**0.0**	27.5229	15.8103	**4.2017**	21.9106	13.1086	**8.0**	25.5941	10.101	**0.0**	45.283	29.5455	**19.3548**
Rare adjectives	14.0	4.0	**1.0**	17.0	9.0	**5.0**	55.0	24.0	**12.0**	18.0	7.0	**0.0**	44.0	10.0	**2.0**
Rare adjectives (per 1000 words)	35.8056	20.0	**4.9751**	19.4954	**17.7866**	21.0084	48.2033	**44.9438**	48.0	32.9068	23.569	**0.0**	55.3459	22.7273	**12.9032**
Rare verbs	6.0	**1.0**	6.0	22.0	8.0	**5.0**	37.0	10.0	**3.0**	17.0	8.0	**5.0**	12.0	7.0	**2.0**
Rare verbs (per 1000 words)	15.3453	**5.0**	29.8507	25.2294	**15.8103**	21.0084	32.4277	18.7266	**12.0**	31.0786	**26.936**	31.8471	15.0943	15.9091	**12.9032**
Rare adverbs	3.0	**0.0**	**0.0**	1.0	**0.0**	**0.0**	2.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Rare adverbs (per 1000 words)	7.6726	**0.0**	**0.0**	1.1468	**0.0**	**0.0**	1.7528	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Rare content words	33.0	8.0	**7.0**	64.0	25.0	**11.0**	119.0	41.0	**17.0**	49.0	18.0	**5.0**	92.0	30.0	**7.0**
Rare content words (per 1000 words)	84.399	40.0	**34.8259**	73.3945	49.4071	**46.2185**	104.2945	76.779	**68.0**	89.5795	60.6061	**31.8471**	115.7233	68.1818	**45.1613**
Distinct rare content words	26.0	8.0	**7.0**	58.0	20.0	**9.0**	83.0	23.0	**9.0**	45.0	18.0	**5.0**	74.0	24.0	**6.0**
Distinct rare content words (per 1000 words)	66.4962	40.0	**34.8259**	66.5138	39.5257	**37.8151**	72.7432	43.0712	**36.0**	82.2669	60.6061	**31.8471**	93.0818	54.5455	**38.7097**
Mean of rare lexical words	17.6471	**8.4211**	8.046	20.6452	12.8866	**10.1852**	23.1969	17.5966	**13.0769**	26.0638	16.6667	**7.9365**	23.4694	13.6986	**8.8608**
Mean of distinct rare lexical words	21.4876	**12.1212**	13.4615	23.4818	15.1515	**12.5**	26.0188	15.6463	**11.1111**	28.125	20.0	**11.1111**	29.8387	16.6667	**11.7647**
Subordinate clauses	8.0	**4.0**	18.0	45.0	17.0	**12.0**	66.0	21.0	**17.0**	22.0	**8.0**	11.0	32.0	12.0	**8.0**
Subordinate clauses (per 1000 words)	20.4604	**20.0**	89.5522	51.6055	**33.5968**	50.4202	57.844	**39.3258**	68.0	40.2194	**26.936**	70.0637	40.2516	**27.2727**	51.6129
Mean of the number of propositions per sentence	4.3333	**1.6552**	2.0455	4.4074	**1.473**	1.6897	4.7576	**1.623**	2.0435	3.75	**1.4474**	2.1538	5.1739	**1.6562**	2.125

Table 1: Results for the automatic evaluation using MultiAzterTest.

Text	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10	Average
Evaluator 1	70.00	66.67	63.33	70.00	70.00	66.67	70.00	66.67	73.33	70.00	68.67
Evaluator 2	66.67	63.33	60.00	66.67	70.00	50.00	73.33	66.67	70.00	63.33	65.00
Average	68.33	65.00	61.67	68.33	70.00	58.33	71.57	66.67	71.67	66.67	66.83
Agreement	0.77	0.93	0.93	0.92	1	0.53	0.92	0.70	0.92	0.85	0.85

Table 2: Results for the checklist-based manual evaluation.

- They claimed to be able to understand more words in the LF version.
- The line spacing is bigger in the LF version.
- Shorter sentences.
- Better explained.
- The LF version has more content and more explanations.

Some participants read the texts aloud, while others decided to read them in silence. For those who read the texts aloud, we did notice that they could read them faster than the ChatGPT generated ones. Participants 4 and 5 mentioned that they would rather read about something they were familiar with and something they liked, such as recipes or football; they claimed that not only they find those topics more entertaining, but they are also able to understand them more, as they already know all the vocabulary. Participant 6 mentioned that they found the ChatGPT version easier, but they still preferred the LF version because of how the ideas were organised and explained. Participant 8 said that they did not have previous knowledge on what the texts were about and therefore needed a bit more time to understand them. Not being able to retell should not be taken as accurate evidence of their understanding of the text. Participant 2 was not able to retell the texts, but they were still able to answer the questions. Besides, Participant 10 was not able to retell, but commented on the content of the text and gave their opinion on what was being presented (difficulties for young people to find jobs and for companies to hire people).

They claimed that one of the reasons why they preferred the LF version was because they were able to understand more words. However, as mentioned in subsection 4.1, the ChatGPT generated texts contain less rare words than the LF versions overall. It is possible that other reasons why they found the LF versions easier are the length of the sentences, the number of subordinate clauses or the number of propositions.

5. Discussion

Our findings show that ChatGPT is able to perform well regarding lexical simplification; as it can be seen in the results for the automatic evaluation, the presence of rare words is very low in the ChatGPT outputs. However, the system fails to per-

Participant	1	2	3	4	5	6	7	8	9	10
Able to retell ChatGPT	yes	no	no	no	no	yes	no	yes	yes	no
Able to retell LF	yes	no	no	yes	no	yes	no	yes	yes	no
Able to answer questions ChatGPT	yes	yes	no	no	no	yes	no	yes	yes	no
Able to answer questions LF	yes	yes	no	no	no	yes	no	yes	yes	no
Preferred version	LF	LF	LF	LF	LF	LF	LF	LF	LF	LF

Table 3: Results for the users' evaluation.

form other important tasks, such as creating examples or explanations. This is one of the most important tasks in E2R adaptation; to anticipate the knowledge of users and adapt the texts according to this, adding examples and explanations when necessary. As highlighted by the evaluators in the users' evaluation, the explanations on the LF versions make them easier to understand. It is recommended to ensure that all significant information is retained within the ChatGPT generated output, as crucial details can be lost during summarisation with ChatGPT. We also observed that other LF rules were broken during the adaptation process, such as not splitting sentences in different lines. E2R and its language variants, including LF, have numerous specific rules. Achieving a certain standard of output is difficult due to ChatGPT's nondeterministic nature, which also makes it impossible to replicate results via the web interface. In addition, finding an adequate prompt is challenging. The prompt chosen yielded consistent outputs during initial trials, but final outputs varied. It might be possible to adapt texts to E2R or LF by pre- and post-editing them using ChatGPT's API. This would ensure the consistency of results and their replicability. However, professional adaptors must ensure that all information from the original text is included and that all rules are followed. Upon initial evaluation, the system's performance seemed satisfactory and consistent with baseline expectations. However, a comprehensive analysis exposed underlying complications and deficiencies that jeopardised its overall effectiveness and dependability.

With regard to the evaluation methods followed for this study, we opted for novel evaluation methods due to the limitations that BLEU, SARI and ROUGE present for E2R evaluation, and the lack of an established evaluation system. The automatic evaluation with *MultiAzterTest* can be helpful when analyzing certain features, such as the length of sentences and documents or the number of rare words. In contrast, we found that the checklist-based manual evaluation lacked capabilities related to the layout of texts, and many of the original capabilities are designed to be only used with sentences, and not texts. We therefore propose three new capabilities to be taken into account when assessing LF and E2R texts. We can conclude that new evaluation methods should be designed in order to be able to check the adequacy of E2R/LF texts before giving them to future users. Finally, conducting a users' evaluation is pivotal in the adaptation and creation processes of E2R/LF texts, as future users know what is best for them. Inclusive design is integral to the development of products and systems that cater to the diverse needs and capabilities of all users, thereby promoting a more holistic and representative approach to user-centered innovation.

6. Conclusion and future work

In this paper, we present a preliminary study of ChatGPT for E2R text adaptation. We focus on the Spanish language and its E2R version, LF. We conduct our experiments with 10 texts from the news domain, and we perform an automatic evaluation, a manual evaluation and a users' evaluation. Throughout our investigation, it became clear that while ChatGPT shows potential in lexical simplification for LF text adaptation, it has deficiencies in key areas like providing explanatory content and adhering to E2R/LF rules. Additionally, the variability in its outputs and difficulties in establishing consistent prompts emphasise the need for improvement. Overall, while promising, the model's current capabilities demand a combination of automation and human intervention for optimal E2R adaptation.

In spite of our findings, we must bear in mind that this study predominantly centres on the user perspective. It is vital to recognize that assessing new technologies should not be limited to an academic or engineering viewpoint. Many laypeople utilize these tools without a comprehensive understanding of their operation or the ability to critically evaluate their outputs. As such, ensuring technologies are intuitive and trustworthy for the broader public remains paramount.

Future work will include similar experiments with the ChatGPT API, other LLMs such as Bloom (Scao et al., 2022) or LLaMa [12] and other languages. We plan to include more texts in our experiments to see whether our findings can be gen-

[12] Introducing LLaMa https://ai.facebook.com/blog/large-language-model-llama-meta-ai/ (last accessed 10-10-2023)

eralised. To address the issue of ChatGPT, and other potential future models, not adhering to certain E2R or LF rules, we could explore developing a rule-based post-processing system that checks and corrects the model's output. We also aim to develop a standardized evaluation method to ensure consistent and objective assessments across studies, which will provide clear guidance for subsequent advancements in the field of E2R and LF text adaptation.

7. Ethics Statement

In developing and presenting this study, our primary aim has been to promote understanding and inclusivity. Our reference to Easy-to-Read language and *Lectura Fácil* is purely descriptive and is based on the aim of promoting accessibility and comprehension for a wide audience. This terminology is free of any value judgements about the intrinsic worth of different languages, dialects or linguistic constructs. We deeply respect all forms of linguistic expression and recognise the potential sensitivities around language discussions. If any aspect of our discussion or the terminology used appears to have unintended connotations, we sincerely apologise.

8. Acknowledgements

We would like to thank *Atzegi*, the Guipuzcoan Association in favour of people with cognitive disabilities, for letting us conduct the users' evaluations at their centres and all the participants who took part. This work has also been partially supported by the projects i) Ixa group A type research group (IT-1805-22) funded by the Basque Government ii) DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe and iii) AWARE Commonsense for a new generation of natural language understanding applications (TED2021-131617B-I00) funded by MCIN/AEI /10.13039/501100011033 by the European Union NextGenerationEU/ PRTR.

9. Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Stefan Bott, Horacio Saggion, and Simon Mille. 2012. Text Simplification Tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1665–1671.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*. Bibliographisches Institut GmbH.

Oscar M Cumbicus-Pineda, Itziar Gonzalez-Dios, and Aitor Soroa. 2021. Linguistic Capabilities for a Checklist-based evaluation in Automatic Text Simplification. In *CTTS@ SEPLN*.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in Easy Language translation. *TSAR 2023*.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic text simplification in Spanish: A comparative evaluation of complementing modules. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–500. Springer.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Paloma Moreda, and Manuel Palomar. 2023. Automatic Text Simplification for People with Cognitive Disabilities: Resource Creation within the ClearText Project. *TSAR 2023*.

Rudolf Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education*.

Rudolf Flesch. 1948. A readability formula in practice. *Elementary English*, 25(6):344–351.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale*, pages 453–463.

Silvia Hansen-Schirra and Christiane Maaß. 2020. Easy language, plain language, easy language plus: perspectives on comprehensibility and stigmatisation. *Easy Language Research: Text and User Perspectives*, 2:17.

Imserso. 2022. Base Estatal de Datos de Personas con Valoración del Grado de Discapacidad.

Instituto Cervantes. 2022. El español: una lengua viva. Informe 2022.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Christiane Maaß. 2019. Easy Language and Beyond: How to Maximize the Accessibility of Communication. Invited Plenary Speech at the Klaara 2019 Conference on Easy-to-Read Language Research (Helsinki, Finland. 19-20 September 2019).

Christiane Maaß and Sergio Hernández Garrido. 2020. Easy and Plain Language in Audiovisual Translation. *Easy language research: Text and user perspectives*, 2:131.

Jean Nitzke, Silvia Hansen-Schirra, Ann-Kathrin Habig, and Silke Gutermuth. 2022. Translating Subtitles into Easy Language: First Considerations and Empirical Investigations. *Translation, Mediation and Accessibility for Linguistic Minorities*, 128:127.

OECD. 2013. *OECD Skills Outlook 2013*.

OECD. 2015. *OECD Skills Outlook 2015*.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

Enid Reichrath and Xavier Moonen. 2022. Assessing the effects of Language for all. *Nordic Journal of Linguistics*, 45(2):232–248.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

UNE. 2018. UNE 153101:2018 EX Lectura fácil. Pautas y Recomendaciones para la Elaboración de Documentos. Madrid: Asociación Española de Normalización.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

A. Appendix

Here we include the prompts we tried with ChatGPT and the issues that arose with them.

PROMPT 1: *Redacta esta noticia en Lectura Fácil: (link a la noticia) Para ello, deberás respetar las normas de Lectura Fácil que aparecen aquí: (link a las normas de Inclusion Europe)*
Write this news item in Lectura Fácil: (link to the news) To do this, you must respect the Easy Read rules that appear here: (link to Inclusion Europe's rules)

Sometimes the output seemed to be following some of the LF rules. However, this is a pdf document (even if it is publicly available in the Internet) so ChatGPT would sometimes say it cannot access it or cannot interact with it.

PROMPT 2: *Redacta esta noticia en Lectura Fácil: (link) Para ello, deberás seguir las normas indicadas en Información para todos: Las reglas europeas para hacer información fácil de leer y comprender de aquí: (link a la web donde se pueden encontrar las normas de Inclusion Europe en pdf, pero no al pdf en sí)*
Write this news item in Lectura Fácil: (link to the news) To do this, you must follow the rules outlined in Information for all: Europe's rules for making information easy to read and understand here: (link to the website where you can find the Inclusion Europe rules in pdf, but not the pdf itself)

PROMPT 2 is the same as PROMPT 1 except that here we did not give the link to the pdf itself, but to the web page where all the pdf documents for all languages are. In this case, the output would normally contain a summary, but overall we got very inconsistent outputs.

PROMPT 3: *Aquí tienes un mensaje en Lectura Fácil que puedes utilizar como referencia: (link)*
¿Puedes crear un texto en Lectura Fácil basado en esta noticia? (link)
Here is an Lectura Fácil text that you can use as a reference: (link)
Can you create a Lectura Fácil text based on this news? (link)

For this prompt, ChatGPT would provide a summary, but never followed the rules for Lectura Fácil.

PROMPT 4: *¿Sabes lo que es la Lectura Fácil?*
Do you know what Lectura Fácil is?
OUTPUT: ...
prompt 4: *Redacta esta noticia en Lectura Fácil: (link)*
Write this news in Lectura Fácil: (link)

In this case, we first asked ChatGPT whether it knew the rules for Lectura Fácil. The output for this question was always different, but it would always say yes and then provide a summary of the rules. When telling ChatGPT to write the news in Lectura Fácil, it would provide a summary, but never followed the rules for Lectura Fácil.

PROMPT 5: *Redacta esta noticia en Lectura Fácil: (link) Para ello, deberás seguir las normas para hacer información fácil de leer y comprender de aquí: (link a una página web que creamos en la que escribimos las normas de Lectura Fácil)*
Write this news in Lectura Fácil: (link) To do this, you must follow the guidelines for making information easy to read and understand from here: (link to a web page we created where we wrote the guidelines for Lectura Fácil).

As ChatGPT is not able to access pdf documents, we decided to write the rules in a publicly availble web page. ChatGPT provided summaries but not texts that followed LF rules.

PROMPT 6: *Las normas de redacción de Lectura Fácil son estas: (link a una página web que creamos en la que escribimos las normas de Lectura Fácil) Redacta esta noticia en Lectura Fácil siguiendo las normas que te he dado: (link)*
The Lectura Fácil writing rules are as follows: (link to a web page we created where we write the guidelines for Lectura Fácil) Write this news in Lectura Fácil following the rules I gave you: (link)

We decided to write the same prompt, but here we changed the order of the information we provided in the prompt. We first provided the rules and then the link to the news. ChatGPT provided summaries and inconsistent answers.

PROMPT 7: *Redacta esta noticia siguiendo las normas de Lectura Fácil europeas para hacer información fácil de leer y comprender: (link)*
Write this news following the European Easy to Read standards to make information easy to read and understand (link)

PROMPT 8: *Tienes que redactar la siguiente noticia en Lectura Fácil: (link)*
You must write the following news in Lectura Fácil: (link)

For PROMPT 7 and PROMPT 8, we decided not to give any guidelines or rules tu ChatGPT. In PROMPT 7 we told the system to follow the Inclusion Europe rules, while in PROMPT 8 we simply told the system to write the text in LF. Both prompts provided summaries, but the outputs did not follow any of the LF rules.

PROMPT 9: *Esta es la versión original de una noticia: (link a la noticia original) Esta es la misma noticia redactada en Lectura Fácil: (link a la misma noticia en versión Lectura Fácil) Basándote en este ejemplo, redacta esta noticia en Lectura Fácil: (link a otra noticia en versión original)*
This is the original version of a news text: (link to the original news) This is the same news written in Lectura Fácil: (link to the same news in Lectura Fácil version) Based on this example, write this news item in Lectura Fácil: (link to other news in original version)

We decided to provide ChatGPT with an example of an original text and its LF version, and then ask it to adapt another text into LF. In this case, ChatGPT would take more time to provide an output, as it would have to visit all links, and it would sometimes start looping. It would then provide generic summaries of the text or provide a summary of any of the given texts, but not necessarily of the one we wanted. ChatGPT works better if you only provide it a single internet link.

PROMPT 10: *Las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil son las siguientes: (lista de normas) Redacta esta noticia en Lectura Fácil siguiendo las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil: (texto original)*
The Lectura Fácil writing rules according to the UNE 153101:2018 Lectura Fácil Standard are as follows: (list of rules) Write this news in Lectura Fácil according to the Lectura Fácil writing rules in accordance with the UNE 153101:2018 EX Easy-to-Read Standard: (original text)

We thought it would be better to write the rules directly in the chat. We also wrote the text in the chat, but this would usually lead to prompts that were too long for ChatGPT to process.

PROMPT 11: *Las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil son las siguientes: (lista de normas) Redacta esta noticia en Lectura Fácil siguiendo las normas de redacción de Lectura Fácil Según la Norma UNE 153101:2018 EX Lectura Fácil: (link a la noticia)*
The Lectura Fácil writing rules according to the UNE 153101:2018 Lectura Fácil Standard are as follows: (list of rules) Write this news in Lectura Fácil according to the Lectura Fácil writing rules in accordance with the UNE 153101:2018 EX Easy-to-Read Standard: (link to the news)

This is the prompt that we ended up using. However, PROMPT 11 shows a single prompt, and in the last version we used, we decided to split this prompt into two separate prompts (see the prompt in subsection 3.2). When using this prompt as a single one, ChatGPT would sometimes reply in English, instead of Spanish or would get issues with the ”click” option.
Apart from the prompts presented here, we also asked ChatGPT to provide us with ten concise prompts that we should use with it to make it adapt a text to LF. We tried the prompts proposed by it, but they all resulted to be too general and did not provide us with the results we were looking for.

PROMPT: *Proporciona diez promts concisas que debería decirte para hacerte adaptar un texto a la lectura fácil*
OUTPUT:

B. Appendix

In this appendix, a list of all the capabilities used is provided:

Prerequisites

- **(P0)**
- **(P1)** No UKN tokens
- **(P2)** No non required quotation marks or strange characters
- **(P3)** No relation orr alignment problems

Grammar/Fluency

Word level:

- **(G1)** No repeated words, e.g. repeated determinants, adjectives, nouns, prepositions, or conjugated verbs
- **(G2)** No tense change, unless necessary for a certain target audience or modality in the verbs

Morpho-syntactic level:

- **(G3)** Correct and finished phrases or sentences
- **(G4)** Correct agreement of subject and verb, grammatical gender of determinants, nouns and adjectives, contractions...
- **(G5)** Correct phraseology: verb + preposition, collocations...
- **(G6)** No repeated arguments: double subject or verb

Cohesion level (both at inter- and intra-sentence levels:

- **(G7)**Correct punctuation
- **(G8)** Correct grammatical order: phrase order, sentence order
- **(G9)** Correct coreference
- **(G10)** No definiteness change

Meaning preservation

- **(M1)** Important information kept
- **(M2)** Register kept, unless required by the target audience
- **(M3)** No meaning change or only subtle nuances changges e.g. deleting or adding emphasizers

Simplicity

- **(S1)** Shorter sentences
- **(S2)** Same term for same concept
- **(S3)** Logical or temporal ordering of relations
- **(S4)** Active voice (instead of passive)
- **(S5)** Simple, frequent words
- **(S6)** Same term consistently used
- **(S7)** Only one main idea per sentence covered
- **(S8)** Only one finite verb for sentence
- **(S9)** Simple punctuation

Ethical aspects

- **(E1)** No wrong information or misinformation
- **(E2)** No non-present stereotypes or unnecessary mentions to discriminate/minoritary groups

Layout

- **(L1)** Format is easier to read
- **(L2)** Line spacing is bigger
- **(L3)** Phrases are not split into lines