Replace, Paraphrase or Fine-tune? Evaluating Automatic Simplification for Medical Texts in Spanish

Leonardo Campillos-Llanos,[1] Ana R. Terroba,[2] Rocío Bartolomé,[3] Ana Valverde,[4] Cristina González,[4] Adrián Capllonch,[5] Jónathan Heras[6]
[1]ILLA - CSIC; [2] F. Rioja Salud; [3] Fac. de Filosofía y Letras (UAM); [4] UTM RANME; [5] IPS Marañón & HGU Gregorio Marañón; [6] Fac. de Informática (Universidad de La Rioja)
leonardo.campillos@csic.es, arterroba@riojasalud.es, rocio.bartolome@uam.es, {avalverde,utm}@ranm.es, adrian.capllonch@salud.madrid.org, jonathan.heras@unirioja.es

Abstract

Patients can not always completely understand medical documents given the myriad of technical terms they contain. Automatic text simplification techniques can help, but they must guarantee that the content is transmitted rigorously and not creating wrong information. In this work, we tested: 1) lexicon-based simplification approaches, using a Spanish lexicon of technical and laymen terms collected for this task (SimpMedLexSp); 2) deep-learning (DL) based methods, with BART-based and prompt-learning-based models; and 3) a combination of both techniques. As a test set, we used 5000 parallel (technical and laymen) sentence pairs: 3800 manually aligned sentences from the CLARA-MeD corpus; and 1200 sentences from clinical trials simplified by linguists. We conducted a quantitative evaluation with standard measures (BLEU, ROUGE and SARI) and a human evaluation, in which eleven subjects scored the simplification output of several methods. In our experiments, the lexicon improved the quantitative results when combined with the DL models. The simplified sentences using only the lexicon were assessed with the highest scores regarding semantic adequacy; however, their fluency needs to be improved. The prompt-method had similar ratings in this aspect and in simplification. We make available the models and the data to reproduce our results.

Keywords: Automatic Text Simplification, Medical Language Processing, Medical Lexicon, Deep Learning, Clinical Trials

1. Introduction

The general public is becoming increasingly interested in learning about their own health conditions (Fox and Duggan, 2013). However, medical texts are written with technical terms and complex syntactic expressions that represent a barrier for the patient to understand them. Medical records, medication leaflets or clinical trial announcements often require medical professionals to explain details about procedures or pathological conditions. Unfortunately, healthcare professionals oftentimes lack the time to give explanations about the jargon during the consultation, which leaves the patient partly or totally uninformed. Automatic text simplification methods can alleviate this language gap and increase the accessibility of health information.

This work presents some experiments we conducted on automatic simplification of medical texts in Spanish. Using the CLARA-MeD corpus (Campillos-Llanos et al., 2022) collected by Campillos-Llanos et al. (2022) (§3.1), we tested a lexicon-based method using equivalences of technical and laymen terms (§3.2). We compared this method to using neural network-based approaches with large language models (LLMs) and combining both approaches—i.e., the lexicon and LLMs (§3.3). We report a quantitative evaluation with standard simplification metrics (§4.1) and a qualitative assessment by human evaluators (§4.2). We distribute the lexicon of equivalent terms and the trained models via the HuggingFace Hub. [1]

2. Background

Automatic text simplification methods (Shardlow, 2014; Saggion, 2017; Štajner, 2021; Al-Thanyyan and Azmi, 2021) may rely on lexicons with equivalent terms of technical and laymen words (Qenam et al., 2017; Koptient and Grabar, 2020b) and rule-based approaches (Suter et al., 2016; Wilkens et al., 2020). There are also data-driven techniques for automatic text simplification that, nowadays, are especially based on deep learning models (Van den Bercken et al., 2019; Shardlow and Nawaz, 2019; Sakakini et al., 2020), including methods based on prompt-learning in recent years (Wang et al., 2022). Hybrid approaches combine some of the above-mentioned methods (Bott et al., 2012; Cardon and Grabar, 2020; Todirascu et al., 2022).

Ideally, simplification strategies are applied at all linguistic levels (lexis, syntax, or discourse). *Lexical simplification* is the task of replacing a difficult-to-read word with an easier synonym or paraphrase (e.g. *autologous* → *from the same individual*). This involves sub-tasks such as complex word identifi-

[1] https://huggingface.co/CLARA-MeD

LREC-COLING 2024, pages 13929–13945

cation (CWI), substitute generation, substitute selection and substitute ranking (North et al., 2023).

In *syntactic simplification*, long sentences tend to be split in shorter clauses, passive voice is changed to active voice, double negation is restructured, or ambiguous anaphoric expressions are replaced (Peng et al., 2012; Collados, 2013; Mukherjee et al., 2017). These operations have generally been implemented with rules (Siddharthan, 2006; Brouwers et al., 2014; Scarton et al., 2017). For *discourse simplification*, these syntax-level operations are also applied, but the goal is simplifying phenomena beyond the sentence level. This also involves removing ambiguous or redundant co-reference chains and cohesion markers, and shortening long paragraphs (Wilkens and Todirascu, 2020).

All these strategies are evaluated using resources for simplification, such as Newsela (Newsela, 2016) introduced by Xu et al. (2015) or the Simple English Wikipedia dataset (Coster and Kauchak, 2011) collected by Coster and Kauchak (2011). Here, we will review only those for the medical domain. Several English corpora have been collected using medical articles from Wikipedia (Van den Bercken et al., 2019), journal articles (Guo et al., 2022; Attal et al., 2023), systematic reviews from the Cochrane library (Joseph et al., 2023), medical manuals (Basu et al., 2023) or heterogeneous sources, including speech corpora and newswire (Shardlow and Alva-Manchego, 2022). Other teams approached the task in real clinical data (Shardlow and Nawaz, 2019; Sakakini et al., 2020; Moramarco et al., 2021).

Simplification datasets or corpora in the medical domain for other languages are scarce. Some resources are the CLEAR corpus (Cardon and Grabar, 2018) collected by Grabar and Cardon (2018) for French, or the German corpora created by Seiffe et al. (2020) and Trienes et al. (2022). Specifically for Spanish, the ALEXSIS dataset (Ferrés and Saggion, 2022) was presented for lexical simplification (Ferrés and Saggion, 2022). However, we did not use it because it does not cover the medical domain. The EASIER Corpus (Rodrigo, Moreno and Martínez, 2022) created by Alarcon et al. (2023) contains few sentences from the medical domain, and we did not use it either. A contribution of this work is using data from clinical trial announcements in Spanish. This text genre that has been scarcely explored for simplification tasks (Fang et al., 2021), although eligibility criteria are semantically complex (Ross et al., 2010) and its readability has been assessed as difficult (Wu et al., 2016). Section §3.1 describes the data used in this work.

3. Methods

Several experimental lines were tested. First, we tested a purely lexicon-driven simplification. To do so, we collected a dataset of 12 605 pairs of technical terms and simplified synonyms, definitions, explanations or paraphrases. Second, we tested purely deep learning-based simplification methods. Five state-of-the-art neural network models (including some trained in multilingual data) were applied, as well as a prompt-based neural model. Third, we combined the lexicon with the deep learning methods, either as data for fine-tuning the models or for pre- and post-processing the data. We evaluated the output with quantitative metrics and eleven human evaluators assessed qualitatively a subset of 500 sentences (out of the 5000 simplified ones). Figure 1 summarizes the methods.

3.1. Data

For the first experiment, we used 3800 parallel sentences (149 862 tokens), with aligned technical and laymen versions, extracted from the CLARA-MeD corpus (Campillos-Llanos et al., 2022). Sources of this dataset were summaries of product characteristics, cancer-related information summaries, and clinical trial announcements from the European Clinicial Trial Register (EudraCT).[2] These sentences were aligned semi-automatically and revised by linguists and experts in health communication, with a high inter-annotator agreement score (average Cohen's Kappa = 0.839 $\pm$ 0.076).

For the second experiment, three linguists created a new set of 1200 parallel sentences (144 019 tokens) by analysing 1040 Spanish texts from EudraCT. Sentences with co-reference ambiguities, digressions or needing syntactic simplification were selected. Then, a simplified version was manually prepared, following the criteria exposed in a companion guideline. More details are explained in (Campillos-Llanos et al., 2024). This dataset includes, for each technical (original) sentence, two versions of simplified sentences: a syntactically simplified sentence, and sentence with both syntactic and lexical simplification. These datasets are released publicly (Campillos-Llanos and Bartolomé and Terroba, 2024). [3] In the present work, we used the version with both syntactic and lexical simplification. The version with sentences only syntactically simplified is aimed at evaluating syntactic simplification methods, which is out of the scope in this work. Table 11 in the Appendix shows samples of the dataset.

[2] www.clinicaltrialsregister.eu

[3] Available at: https://digital.csic.es/handle/10261/346579

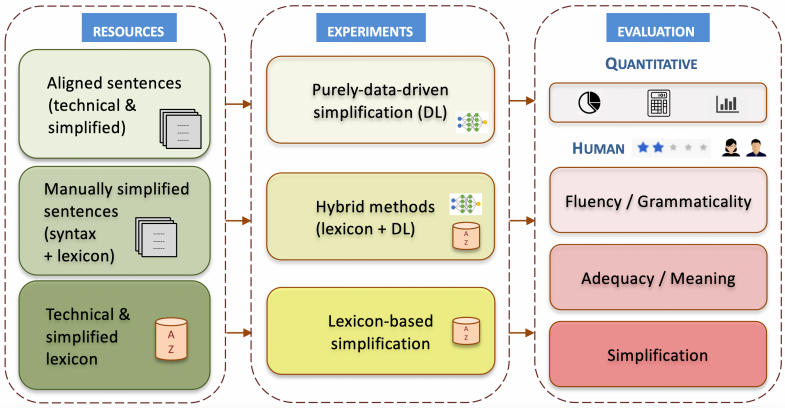

Figure 1: Summary of methods

3.2. SimpMedLexSp lexicon

We collected a first version of a simplified medical lexicon for Spanish (Figure 2 shows a sample). To the best of our knowledge, this is the first patient-oriented vocabulary in Spanish, with technical terms mapped to layperson variants, as in consumer health vocabularies (CHV) for other languages (Marshall, 2000; Keselman et al., 2008; Grabar and Hamon, 2016). The current version of SimpMedLexSp includes 14 013 pairs of technical terms and simplified synonyms, definitions or paraphrases (including variant forms of gender and number).[4] A subset of 4642 terms was also mapped to Concept Unique Identifiers from the Unified Medical Language System (Bodenreider, 2004). We applied a string matching technique to entry terms in the SimpMedLexSp lexicon and mapped them to CUIs using the MedLexSp lexicon (Campillos-Llanos, 2023). MedLexSp indexes terms based on UMLS CUIs and contains form variants including singular/plural, or masculine/feminine. The mapping of a subset of terms was manually revised, given that some terms are ambiguous: e.g. *aa* can be an abbreviation for *amino acid* (CUI: C0002520) or it can stand for *acute abdomen* (CUI: C0000727). We used the following sources for SimpMedLexSp:

- EUGLOSS (HIP, 1995): we used Spanish terms from this multilingual glossary of technical and popular medical terms.
- *Dictionary of Medical Terms* (RANME, 2011): we used pairs of technical and colloquial terms, which are encoded with a specific tag.
- Pairs of acronyms/abbreviations and full forms from the MedLexSp lexicon (Leonardo Campillos-Llanos, 2023).
- Pairs of terms extracted from the CLARA-MeD corpus (Campillos-Llanos et al., 2022), by applying paraphrase patterns adapted to Spanish from previous work (Vydiswaran et al., 2014). For example, a pair of technical and patient terms can be extracted using pattern *también conocido como* ('also known as'): *familial male precocious puberty, also known as testotoxicosis*.

To prepare the lexicon, we applied a frequency-based filter to discard widely-used medical terms (e.g. *tos*, 'coughing'). We used the frequency list from the CORPES Spanish corpus and excluded terms with a frequency over 100.[5] The motivation to avoid simplifying all medical terms is based on the outcomes of previous works (Leroy and Endicott, 2012), in which high-frequency words were judged easier. We also considered the qualitative analyses conducted with physicians to develop a system to link medical words with their lay definitions (Chen et al., 2018). The lexicon was split in two files, according to the type of lexical simplification:

- *Lexical substitution*: we mainly included synonym terms in a laymen register (e.g. *cephalea* → *headache*). This subset amounted to 6090 pairs. With these, we applied a substitution strategy so that technical terms were replaced with their laymen equivalent in the original text. Note that some terms may be replaced with different laymen terms: e.g. *abdomen* → *vientre* or *barriga*, 'belly'). To date, we manually selected only one candidate to be replaced. We did not conduct substitute generation, selection or ranking, which is left for future work.
- *Definition or paraphrase*: we gathered other types of medical terms that are not to be replaced in the original data. A common example is any active ingredient or medical drug (e.g. *abiraterone*), which lacks a laymen synonym; for these, we preserved the original term and appended a definition or paraphrase in

[4]The current released version has more entries than the one used for the experiments reported (12 605 pairs).

[5]Code at: https://shorturl.at/mqzFZ

```
CUI|TECHNICAL_TERM|PATIENT_TERM
C0004057|ácido acetilsalicílico|aspirina
C0006736|cálculo urinario|piedra en la orina
C0007137|ca epidermoide|carcinoma epidermoide
C0007137|cec|carcinoma escamocelular
C0040423|amigdalectomía|operación de anginas
C0231221|asintomático|sin síntomas
C0277797|apirético|sin fiebre
C0522523|percutáneo|a través de la piel
```

(a) Mapping to UMLS CUIs

```
p.a.|presión arterial
peg|polietilenglicol
per os|por la boca
percutánea|a través de la piel
percutáneas|a través de la piel
percutáneo|a través de la piel
percutáneos|a través de la piel
peribucales|alrededor de la boca
peribucal|alrededor de la boca
```

(b) Genre and plural variant forms

Figure 2: Samples of SimpMedLexSp

brackets (e.g. *cancer treatment*). This subset amounted to 6515 pairs.

The lexicon is freely available for research and educational purposes, and we applied it to simplify sentences in several ways:

1. Only the lexicon (Method A), either for lexical substitution or both lexical substitution and adding definitions/paraphrases.
2. The lexicon as training data along with the training subsets of sentences (Method D & E).
3. The lexicon for lexical substitution of medical terms in the training data (Methods F & G).
4. The lexicon for appending definitions/paraphrases to model predictions (Methods E and G).

3.3. Deep learning-based methods

We experimented with two strategies for text simplification: fine-tuning and prompt learning. In the fine-tuning strategy (Howard and Ruder, 2018), we took several language models pre-trained on large datasets. We replaced the head of the models with a new head adapted to simplify text, and finally trained the new model with a development set using the 3800 sentences from the CLARA-MeD corpus (Campillos-Llanos et al., 2022), and from the 1200 manually-simplified sentences (Campillos-Llanos et al., 2024). First of all, we tried to replicate former results (Shardlow and Nawaz, 2019) using the OpenNMT library (Klein et al., 2017). However, we did not pursue using this model owing to the results below the other models (Tables 1-7).

We tested the following Transformer-based models: (1) Multilingual BART (mBART) (Liu et al., 2020) Large; (2) Multilingual T5 (Xue et al., 2021), trained on the Simplext corpus (Saggion et al., 2015); (3) Pegasus XSUM (Zhang et al., 2020); (4) News Abstractive Summarization for Spanish (NASES), a BART model trained to summarize Spanish news articles (Ahuir et al., 2021); and (5) the Maria model trained on the MultiLingual SUMmarization (MLSUM) dataset (Fandiño et al., 2022), which we will call MariMari from here on.

These models were tested in several contexts:

1. Fine-tuning the model with the sentences training set, but without the lexicon (Method C).
2. Fine-tuning the model using both the sentences training set and medical lexicon as training data (Method D).
3. The previous approach and appending definitions or paraphrases to the models' output (Method E).
4. Fine-tuning the model on the training set where medical terms were replaced with synonyms (lexical substitution; Method F).
5. Fine-tuning the model on the sentences training set with lexical substitution of medical terms and post-processing with the lexicon (Method G).

When we applied lexical substitution, we did not keep both the original sentence as well as the modified one in the training set. We either used the original sentences without lexical substitution (Methods C, D and E) or sentences in which medical terms were substituted (Methods F and G).

The post-processing involves appending definitions or paraphrases to complex medical terms, after the sentence has already been simplified with a neural network-based model. This is typically the case for medical drug names, which tend to lack a lay synonym, and we concatenate an explanation about what the drug is used for. For example, if the mBART simplified sentence is *Ensayo clínico de seguridad de carfilzomib* ('Safety clinical trial of carfilzomib'), the post-processed version contains a definition of *carfilzomib*: *Ensayo clínico de seguridad de carfilzomib (medicamento para tratar el cáncer)* ('Safety clinical trial of carfilzomib (drug to treat cancer)').

The multilingual and Spanish models were trained using the HuggingFace libraries (Wolf et al., 2020) and a GPU Nvidia GeForce 3090. We trained all the HuggingFace models for 30 epochs with a batch size of 8, and a learning rate of 5.6e-5. The models are released in the HuggingFace Hub.

In the prompt learning strategy (Brown et al., 2020), a language model performs specific tasks by conditioning the model behavior through carefully

designed instructions and contexts. In our case, we took a BERTIN GPT-J-6B Spanish language model (BERTIN-project, 2023) fine-tuned on the Spanish Alpaca dataset (Taori et al., 2023) and provided by the HuggingFace library. We applied several prompting approaches. First, we used zero-shot prompting (Brown et al., 2020); that is, we just requested the model to simplify a given sentence. Second, we applied few-shot prompting by showing the model some examples of simplified sentences and then providing the model with the sentence to simplify (Method B in Table 2). Third, we fine-tuned the model using the training subset of technical and simplified sentence pairs. Finally, we fine-tuned the model both with the training subset and our lexicon using LoRA (Hu et al., 2021) and then applied zero-shot prompting (i.e. we asked the model to simplify sentences but prompts did not include simplification examples). Table 10 in the Appendix shows examples of the prompts.

3.4. Evaluation

Out of the 3800 aligned sentences from the CLARA-MeD corpus, we applied a 5-fold evaluation procedure using 80% (3040) of the sentences for training and 20% (760) for testing. We used the same proportion for the 1200 manually simplified sentences (960/240). We performed a quantitative evaluation with these metrics: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and SARI (Xu et al., 2016). BLEU and ROUGE take into account n-grams, and focus on precision and recall, respectively. SARI also considers n-grams but also simplification operations such as token addition or deletion. We did not use the Flesch-Kincaid score (Flesch, 1948) because it has been criticized for not correlating well with simplicity (Grabar and Saggion, 2022), and has limitations in assessing the readability of electronic health records (Zeng-Treitler et al., 2007).

We also conducted a human evaluation with eleven linguists and documentalists. We extracted 50 random sentences from the test set of the first experimental dataset (3800 manually aligned sentences) and 50 from the second experimental corpus (1200 manually simplified sentences). For each sentence, we obtained the output of five simplification methods, and eleven subjects assessed the simplification (in total, each evaluated 500 simplified sentences). We evaluated the prompt-learning method and the simplification using only the lexicon, and also most of the approaches that achieved the highest quantitative scores. Note that we did not evaluate the NASES or MariMari models, despite the high quantitative results (BLEU or SARI scores) they obtained. When we analyzed the output of NASES or MariMari, we found serious hallucinations and errors related to medical terminology. For example, NASES wrongly paraphrased 'prophylaxis' as 'serious allergic reaction' (*profilaxis (reacción alérgica grave)*).

In particular, the methods assessed in the qualitative evaluation were: (1) Using only the lexicon to replace synonyms and append paraphrases or definitions; (2) the multilingual BART model (mBART); (3) mBART fine-tuned with the lexicon; (4) mBART fine-tuned with the lexicon and post-processed (adding paraphrases or definitions); and (5) BERTIN-prompt-learning fine-tuned with the lexicon (zero-shot). The method applied to each sentence was blinded to the evaluators.

Similarly to previous works (Alva-Manchego et al., 2020; Maddela et al., 2021; Yamaguchi et al., 2023; Joseph et al., 2023), evaluators assessed: 1) grammar and fluency; 2) semantic coherence and adequacy; and 3) simplification. We explained to evaluators that the target reader of the simplified sentences should be a non-healthcare professional. Thus, any sentence with difficult-to-understand terms should receive poor scores. Table 12 in the Appendix includes the instructions given to conduct the task.

4. Results

4.1. Quantitative evaluation

Tables 1- 8 show the quantitative results of the different automatic simplification methods. We report the average score $\pm$ standard deviation of 5 experimental rounds, with the best results in bold (*par.* stands for 'paraphrase', and *def.*, for 'definition'); the name of the methods that the human evaluators assessed qualitatively appears in italics. When we used only the lexicon, the replacement-only strategy achieved higher scores (except for the SARI metric) than combining replacements and paraphrases. Results when using the lexicon outperformed those obtained with the mT5 and Pegasus XSUM models (except BLEU for Pegasus XSUM) with the subset of 3800 sentences. Therefore, we did not continue testing those models.

The mBART model achieved the highest scores across all experimental contexts (i.e. fine-tuning with and without the lexicon, and post-processing). With the 3800 sentences, fine-tuning with the lexicon yielded the highest ROUGE and BLEU scores; and pre- and post-processing with the lexicon combined with fine-tuning gave the highest SARI score. The fine-tuned prompt-based model (BERTIN) yielded ROUGE, BLEU and SARI scores just below the mBART model; on the contrary, the zero-shot and few-shot prompting approaches achieved worse results than the fine-tuned version.

On the 1200 manually-simplified sentences, BLEU, SARI and ROUGE were higher, but results were similar across models (Table 8). However,

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(A) Only lexicon	Replacements	**46.69** (±**1.06**)	**28.64** (±**0.82**)	**41.79** (±**1.00**)	45.85 (±0.5)	**19.67** (±**0.74**)
	Replacements + *par./def.*	46.24 (±0.8)	27.49 (±0.65)	40.29 (±0.79)	**47.41** (±**0.53**)	14.49 (±0.53)

Table 1: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method A)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(B) Only prompts	BERTIN (zero-shot prompting)	**41.89** (±**0.62**)	**23.19** (±**0.66**)	**35.60** (±**0.72**)	43.07 (±0.27)	**15.75** (±**0.78**)
	BERTIN (few-shot prompting)	39.27 (±0.59)	21.43 (±0.55)	34.08 (±0.60)	**43.29** (±**0.52**)	12.01 (±0.41)

Table 2: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method B)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(C) Fine-tuning without lexicon	BERTIN	46.90 (±1.95)	29.17 (±1.54)	41.73 (±1.79)	47.00 (±2.74)	22.03 (±0.74)
	MariMari	42.47 (±0.66)	24.28 (±0.78)	36.78 (±0.67)	47.08 (±0.72)	17.89 (±0.94)
	mBART	**48.82** (±**1.02**)	**31.04** (±**1.01**)	**43.96** (±**1.02**)	**50.93** (±**0.62**)	**22.90** (±**0.98**)
	mT5	34.20 (±0.79)	19.85 (±0.61)	31.56 (±0.66)	40.46 (±0.45)	6.58 (±0.39)
	NASES	44.33 (±1.18)	26.51 (±1.25)	38.54 (±1.23)	48.68 (±0.64)	20.47 (±1.16)
	Pegasus	43.78 (±0.7)	25.86 (±0.85)	39.74 (±0.76)	41.65 (±0.23)	14.62 (±0.43)
	XSUM					

Table 3: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method C)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(D) Fine-tuning with lexicon	*BERTIN (zero-shot)*	48.34 (±0.83)	29.27 (±0.85)	42.95 (±0.98)	49.09 (±0.52)	21.08 (±0.90)
	MariMari	43.35 (±0.92)	25.95 (±2.4)	37.75 (±1.06)	47.35 (±0.75)	18.30 (±1.07)
	mBART	**50.50** (±**0.98**)	**32.65** (±**1.00**)	**45.33** (±**1.03**)	**51.20** (±**0.63**)	**24.71** (±**1.34**)
	NASES	30.10 (±16.07)	14.74 (±12.63)	25.52 (±13.97)	42.67 (±7.57)	11.16 (±9.31)

Table 4: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method D)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(E) Fine-tuning with lexicon + post-processing (par./def.)	MariMari	41.88 (±0.99)	23.8 (±1.19)	35.83 (±1.04)	47.7 (±0.62)	14.27 (±1.16)
	mBART	**48.07** (±**0.82**)	**30.07** (±**0.83**)	**42.42** (±**0.82**)	**51.11** (±**0.43**)	**18.26** (±**0.83**)
	NASES	39.74 (±1.66)	22.2 (±1.84)	33.37 (±1.74)	48.17 (±0.44)	13.14 (±1.08)

Table 5: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method E)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(F) Lexical substitution with lexicon + fine-tuning	MariMari	42.73 (±0.94)	24.41 (±1.00)	36,99 (±1.11)	47.44 (±0.72)	17.63 (±0.97)
	mBART	**48.83 (±1.16)**	**30.69 (±1.27)**	**43.75 (±1.14)**	**50.28 (±0.79)**	**22.90 (±0.98)**
	NASES	44.19 (±0.70)	22.26 (±0.85)	38.31 (±0.74)	48.16 (±0.50)	19.15 (±1.28)

Table 6: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method F)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(G) Lexical substitution with lexicon + fine-tuning + post-processing	MariMari	44.0 (±0.82)	25.57 (±0.80)	38.28 (±1.03)	47.29 (±0.48)	17.9 (±1.00)
	mBART	**49.5 (±2.34)**	**31.42 (±2.34)**	**44.48 (±2.42)**	**50.12 (±1.01)**	**23.25 (±1.96)**
	NASES	45.32 (±1.02)	27.12 (±1.22)	39.2 (±1.12)	47.63 (±0.55)	19.73 (±0.89)

Table 7: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method G)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(A) Only lexicon	Replacements	73.95 (±0.86)	59.72 (±1.32)	68.77 (±0.98)	50.81 (±0.75)	41.11 (±1.13)
	Replacements + par./def.	69.29 (±1.0)	53.66 (±1.28)	62.71 (±0.94)	49.55 (±0.90)	36.41 (±1.98)
(C) Fine-tuning without lexicon	*BERTIN*	70.43 (±1.63)	54.21 (±1.71)	63.31 (±1.59)	51.87 (±1.73)	43.40 (±2.26)
	mBART	**76.96 (±0.82)**	**63.57 (±0.97)**	**71.37 (±0.77)**	**61.04 (±0.86)**	**52.49 (±1.06)**
(D) Fine-tuning with lexicon	*BERTIN (zero-shot)*	72.92 (±0.42)	59.77 (±0.70)	68.12 (±0.22)	55.66 (±0.41)	44.5 (±1.08)
	mBART	74.78 (±0.84)	61.21 (±1.21)	69.36 (±0.97)	57.81 (±0.87)	48.95 (±1.48)
(E) Fine-tuning with lexicon + post-processing (par./def.)	*mBART*	70.55 (±0.64)	55.75 (±1.1)	64.00 (±0.75)	55.71 (±0.47)	40.68 (±1.86)

Table 8: Results of the quantitative evaluation with the 1200 manually simplified sentences pairs (average score ± standard deviation of 5 experimental rounds); *par.*: 'paraphrase'; *def.*: 'definition'; the name of the methods that the human evaluators assessed qualitatively appears in italics; best results in bold

mBART fine-tuned without the lexicon achieved the highest scores.

Note that we also conducted a preliminary test of OpenNMT models without the lexicon on the 3800 sentences, but results were below the above-mentioned models. For example, OpenNMT (2 layers) yielded an average Rouge1 = 18.92 (±1.27), Rouge2 = 5.23 (±0.66), RougeL = 16.01 (±1.1), and SARI = 36.78 (±0.23) (5 evaluation rounds with 5-folds). For its part, OpenNMT BRNN yielded an average of Rouge1 = 18.71 (±0.99), Rouge2 = 4.00 (±1.2), RougeL = 15.56 (±0.97), and SARI = 35.43 (±0.36).

4.2. Human evaluation

Table 9 shows the results of the human evaluation of 500 simplified sentences (250 pairs from the 3800 sentences, and 250 from the 1200 sentences). Because evaluating all the methods is unfeasible, we only considered the five best performing models regarding the automatically computed metrics.

With the manually-aligned sentences (n=3800), both the lexicon approach and the prompt-learning-based method achieved the highest scores in semantic correctness and adequacy. The lexicon method was rated in second place for simplification, with lower scores in grammaticality and fluency. MBART-based methods (with and without

Method	Manually-aligned sentences (n=3800)				Manually-simplified sentences (n=1200)			
	G	M	S	Avg	G	M	S	Avg
Lexicon (replace + paraphrase/definition)	4.1	**4.3**	3.3	3.9	4.2	**4.5**	**3.3**	4.0
mBART	4.2	3.3	2.7	3.4	4.2	3.9	3.0	3.7
mBART + fine-tune with lexicon	4.3	3.7	3.1	3.7	4.2	3.9	3.0	3.7
mBART + fine-tune lexicon + post-proc.	4.0	3.7	3.1	3.6	3.9	3.9	3.0	3.6
Prompt learning fine-tuned with lexicon	**4.7**	**4.3**	**3.4**	**4.1**	**4.7**	**4.5**	3.1	**4.1**

Table 9: Results (average of 5-point Likert scale) of the human evaluation (n=11) of simplifying the 3800 sentence pairs (left) and 1200 sentences pairs (right) for Grammaticality / Fluency (G), Semantic adequacy / Meaning preservation (M), and Simplification (S); and average (Avg) of scores for the three aspects

the lexicon) achieved scores below the other methods. In general, mBART improves in semantic adequacy/correctness and simplication when the lexicon is used. The prompt-learning-based model achieved balanced scores in all aspects. It had the highest scores in grammaticality/fluency and simplification, with values closed to those obtained with the lexicon method. Lastly, all methods applied still need to improve the simplification aspect.

With the manually-simplified sentences (n=1200), the lexicon approach achieved the highest scores in simplification, and again lower scores in grammaticality/fluency. The lexicon approach and the prompt-learning strategy were rated similarly regarding regarding semantic adequacy. The prompt-learning strategy again received balanced scores in all aspects. Also, the mBART models were generally rated with lower scores compared to the lexicon-only or the prompt-learning methods.

Table 13 in the Appendix includes samples of sentences simplified by the evaluated models. In example 1, the lexicon produces a long sentence, but the result is more accurate semantically. The mBART model creates a hallucination (**tendinios* does not exist in Spanish) and wrong paraphrases for *arthritis*. For its part, the BERTIN (prompt-learning-based) model does not provide any simplification. In example 2, both the lexicon and the prompt-learning-based models simplified adequately, but the prompt method gave more information. The mBART models yielded wrong simplifications regarding the semantic content. In example 3, the term *rabdomiosarcoma* is wrongly paraphrased by mBART (it is not an autoimmune disorder); and applying the only lexicon or combined with mBART yields a long and unnatural paraphrase. The prompt-learning-based method fine-tuned with the lexicon provides a finer simplification.

5. Discussion

Large-language models (LLMs) undoubtedly bring a wide range of new applications for NLP tasks, including automatic text simplification for medical documents. This led us to assess the degree to which linguistically-motivated methods (e.g. lexicon-based approaches) are performant, given the time-consuming effort they require, compared to data-driven approaches. In this crossroad, the experiments here presented try to shed some light, although more research is needed to confirm our outcomes.

First, BLEU, ROUGE and SARI scores were higher with the manually simplified dataset (1200 sentences) compared to the manually aligned (3800 sentences). Although that subcorpus has a small size to generalize results (overfitting could have occurred), our outcomes seem to support the use of datasets manually prepared by professionals to achieve quality simplifications.

Second, using the lexicon tended to show higher quantitative and qualitative results. Compared to the models fine-tuned without the lexicon, models trained with the lexicon yielded higher BLEU, ROUGE and SARI scores (except the NASES model with the 3800 sentences). However, the best strategy in our experiments was using the lexicon for fine-tuning. When applying the lexicon for both lexical substitution and concatenation of paraphrases/definitions, models slightly decreased (exceptions are the mBART and NASES models for the SARI metric). Overall, the highest scores were obtained with the mBART model. Lastly, using only the lexicon to replace or paraphrase difficult-to-understand words (without any neural network-based model) showed middle-range quantitative results.

Third, quantitative metrics did not match well with human evaluation scores. The mBART model achieved the highest BLEU, ROUGE and SARI scores; nonetheless, it was not among the best evaluated models in the human evaluation (§4.2). For its part, the prompt-based (BERTIN) model achieved BLEU, ROUGE and SARI scores just below those of mBART, but received higher ratings and even the highest scores in grammaticality or fluency. In the human evaluation, the only-lexicon approach and the prompt-method were rated in first position regarding semantic adequacy and meaning preservation, and regarding simplification.

This implies that a such type of lexicon is preferable for a task where semantic correctness is to be maximized—and the medical domain demands such requirement. However, the size of the lexicon needs to be increased and updated regularly.

Nonetheless, grammaticality and fluency needs to be largely improved in the manner we implemented the lexicon-based approach. Sentences with replaced or paraphrase terms are too large or contain brackets inside other brackets (example 3 in Table 13), and a syntactic simplification is needed to split them in two clauses. Moreover, replacements may cause errors if the gender of the source word does not correspond to that of the target term (e.g. *cirugía del abdomen* → *cirugía del *barriga*, 'abdomen surgery'). Other teams also found the aforementioned problems (Siddharthan, 2006; Koptient and Grabar, 2020a). Altogether, in our experiments, the prompt-based strategy with the BERTIN model fine-tuned with the lexicon showed balanced results: it yielded fine scores across all quantitative metrics and achieved the highest average rating of all aspects assessed in the human evaluation.

We would expect similar results on medical texts in similar languages. Indeed, Cardon and Grabar (2020) already reported that a medical lexicon of technical and simplified terms was efficient when used for training. Alarcón et al. (2023) also obtained fine results with BART-based models. However, a multilingual replication of our results is yet to be confirmed, and depends on the availability of patient-oriented lexicons for other languages, which are scarce.

6. Limitations

Among the weaknesses of our study, we did not apply any procedure to select the candidate synonyms to replace (e.g. by ranking or taking into account the target context). However, some medical terms may be replaced differently according to the context: e.g. *sintomático* ('symptomatic') in *tratamiento sintomático* → *tratamiento de los síntomas* ('treatment of symptoms') vs. *paciente sintomático* → *paciente con síntomas* ('patient with symptoms'). This is especially important for ambiguous acronyms and abbreviations: e.g. *fc* can stand for *farmacocinética* ('pharmacokinetics'), *fosfocreatina* ('phosphocreatine') or *frecuencia cardíaca* ('heart rate'). Besides, the coverage of the lexicon still needs to be improved with more equivalent terms from other term sources; and updated with new concepts. Corpus-based frequencies (from technical and laymen texts) could be added to the SimpMedLexSp lexicon. Finally, our results need to be explored with further experiments—especially, training with larger corpora—and, ideally, replicated with data from other languages.

7. Conclusions

We presented a set of experiments on automatic simplification of medical texts in Spanish using a lexicon of specialized and laymen terms (SimpMedLexSp), neural network-based methods, and combining both approaches. We evaluated the methods using standard quantitative metrics and conducting a human evaluation of the grammaticality/fluency, semantic adequacy and global simplification. Overall, our results showed that using such type of lexicon generally increased results when used as training data for fine-tuning deep learning models, compared to models without it. As well, using only the lexicon—for lexical substitution and appending paraphrases or explanations—, or using it in combination with a prompt-based method, obtained simplified sentences that were evaluated with the highest scores, either for semantic adequacy or simplification. Although our findings need more experimental evidence, they tend to show that using such type of lexicon benefits to text simplification in the medical domain, which requires high semantic correctness. Other contributions of this work are the trained models (available at the HuggingFace Hub)[6] and the current version of the lexicon, which is freely distributed for research and educational purposes.[7]

8. Ethics Statement

The models developed in this work should not be used for medical decision making without human assistance and supervision, nor for self-diagnosis by patients. Deep learning models can generate erroneous content and hallucinations that may contradict the medical knowledge, so health experts should check the generated output. Third parties who deploy or provide systems/services using these models should note that it is their responsibility to mitigate the risks arising from their use.

9. Acknowledgements

We greatly thank the human evaluators who assessed the simplification output, and also Marisol and Yara (Information Processing Unit at CCSH) for their revision of the lexicon. This work was done in project CLARA-MED (PID2020-116001RA-C33) funded by MCIN/AEI/10.13039/501100011033/, in call: "Proyectos I+D+i Retos Investigación"; and also partially supported by Grant PID2020-115225RB-I00 funded by MCIN/AEI/10.13039/501100011033.

[6] https://huggingface.co/CLARA-MeD

[7] https://digital.csic.es/handle/10261/349662

10. Bibliographical References

Vicent Ahuir, Lluís-F Hurtado, José Ángel González, and Encarna Segarra. 2021. NASca and NASes: Two monolingual pre-trained models for abstractive summarization in Catalan and Spanish. *Applied Sciences*, 11(21):9872.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Rodrigo Alarcón, Paloma Martínez, and Lourdes Moreno. 2023. Tuning BART models to simplify Spanish health-related content. *Procesamiento del Lenguaje Natural*, 70:111–122.

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. EASIER corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proc. of the 58th Annual Meeting of the ACL*, pages 4668–4679, Online. Association for Computational Linguistics.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-EASi: Finely annotated dataset and models for controllable simplification of medical texts. *Proc. of AAAI 2023*, pages 14093–14101.

BERTIN-project. 2023. BERTIN-GPT-J-6B Alpaca. `https://huggingface.co/bertin-project/bertin-gpt-j-6B-alpaca` Accessed 26 June 2023.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Stefan Bott, Horacio Saggion, and David Figueroa. 2012. A hybrid system for Spanish text simplification. In *Proc. of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 75–84.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for French. In *Proc. of the 3rd PITR Workshop@ EACL 2014*, pages 47–56.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leonardo Campillos-Llanos. 2023. MedLexSp–A medical lexicon for Spanish medical natural language processing. *Journal of Biomedical Semantics*, 14(1):1–23.

Leonardo Campillos-Llanos, Rocío Bartolomé Rodríguez, and Ana R Terroba Reinares. 2024. Enhancing the understanding of clinical trials with a sentence-level simplification dataset. *Procesamiento del lenguaje natural*, 72.

Leonardo Campillos-Llanos, Ana R Terroba Reinares, Sofía Zakhir Puig, Ana Valverde, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del lenguaje natural*, 69:189–196.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, Hong Yu, et al. 2018. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research*, 20(1):e8669.

José Camacho Collados. 2013. Splitting complex sentences for natural language processing applications: Building a simplified Spanish corpus. *Procedia-Social and Behavioral Sciences*, 95:464–472.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 665–669.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira

Ocampo, Casimiro Pio Carriño, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Yilu Fang, Jae Hyun Kim, Betina Ross S Idnay, Rebeca Aragon Garcia, Carmen E Castillo, Yingcheng Sun, Hao Liu, Cong Liu, Chi Yuan, and Chunhua Weng. 2021. Participatory design of a clinical trial eligibility criteria simplification method. In *Proc. of MIE*, pages 984–988.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: a dataset for lexical simplification in Spanish. In *Proceedings of LREC 2022*, pages 3582–94, Marseille, France.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Health*, 2013:1–55.

Natalia Grabar and Rémi Cardon. 2018. CLEAR - Simple corpus for medical French. In *Proc. of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9.

Natalia Grabar and Thierry Hamon. 2016. A large rated lexicon with French medical words. In *Proc. of LREC 2016*, pages 2643–2648, Portorož, Slovenia.

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*, pages 453–463.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J Ramanathan, Wei Xu, Byron C Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. *arXiv preprint arXiv:2305.12532*.

Alla Keselman, Robert Logan, Catherine Arnott Smith, Gondy Leroy, and Qing Zeng-Treitler. 2008. Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association*, 15(4):473–483.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Anaïs Koptient and Natalia Grabar. 2020a. Fine-grained text simplification in French: steps towards a better grammaticality. In *Proc. of Int. Symp. on Health Information Management Research*.

Anaïs Koptient and Natalia Grabar. 2020b. Rated lexicon for the simplification of medical texts. In *Proc. of HEALTHINFO 2020*, Porto, Portugal.

Gondy Leroy and James E Endicott. 2012. Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization of ACL*, pages 74–81, Barcelona, Spain.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proc.of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3536–3553, Online. Association for Computational Linguistics.

Philip D Marshall. 2000. Bridging the terminology gap between health care professionals and patients with the Consumer Health Terminology (CHT). In *Proceedings of the AMIA Symposium*, page 1082. American Medical Informatics Association.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, et al. 2021. Towards more patient friendly clinical notes through

language models and ontologies. In *Proc. of the AMIA Annual Symposium*, pages 881–890.

Partha Mukherjee, Gondy Leroy, David Kauchak, Srinidhi Rajanarayanan, Damian Y Romero Diaz, Nicole P Yuan, T Gail Pritchard, and Sonia Colina. 2017. NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of biomedical informatics*, 69:55–62.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

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

Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu, and K Vijay-Shanker. 2012. iSimp: A sentence simplification system for biomedical text. In *2012 IEEE Int. Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE.

Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12):e417.

RANME. 2011. *Diccionario de términos médicos*. Madrid: Panamericana.

Jessica Ross, Samson Tu, Simona Carini, and Ida Sim. 2010. Analysis of eligibility criteria complexity in clinical trials. *Proc. of the Summit on Translational Bioinformatics*, 2010:46.

Horacio Saggion. 2017. *Automatic text simplification*, volume 32. Synthesis Lectures on Human Language Technologies, Springer.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, et al. 2020. Context-aware automatic text simplification of health materials in low-resource domains. In *Proc. of the 11th LOUHI Workshop*, pages 115–126.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proc. of the IJCNLP 2017, System Demonstrations*, pages 25–28.

Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. 2020. From witch's shot to music making bones - resources for medical laymen to technical language and vice versa. In *Proc. of LREC 2020*, pages 6185–6192, Marseille, France.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple TICO-19: A dataset for joint translation and simplification of COVID-19 texts. In *Proceedings of LREC 2022*, pages 3093–3102.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proc. of the 57th ACL*, pages 380–389, Florence, Italy. Association for Computational Linguistics (ACL).

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based automatic text simplification for German. In *Proc. of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 280–287, Bochum, Germany.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca` Accessed 26 June 2023.

Amalia Todirascu, Rodrigo Wilkens, Eva Rolin, Thomas François, Delphine Bernhard, and Nuria Gala. 2022. Hector: A hybrid text simplification tool for raw texts in French. In *Proc. of the 12th Language Resources and Evaluation (LREC)*, pages 4620—4630.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: towards a new text simplification dataset. In *Proc. of the TSAR-2022 Workshop*, pages 19–27, Marseille, France.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *Proc. of the World Wide Web Conference*, pages 3286–3292.

Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. 2014. Mining consumer health vocabulary from community-generated text. In *AMIA Annual Symposium Proceedings*, volume 2014, pages 1150–1159. American Medical Informatics Association.

Haochun Wang, Chi Liu, Nuwa Xi, Sendong Zhao, Meizhi Ju, Shiwei Zhang, Ziheng Zhang, Yefeng Zheng, Bing Qin, and Ting Liu. 2022. Prompt combines paraphrase: Teaching pre-trained models to understand rare biomedical words. In *Proc. of the 29th International Conference on Computational Linguistics*, pages 1422–1431, Gyeongju, Republic of Korea.

Rodrigo Wilkens, Bruno Oberle, and Amalia Todirascu. 2020. Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 93–100.

Rodrigo Wilkens and Amalia Todirascu. 2020. Un corpus d'évaluation pour un système de simplification discursive. In *6e conférence conjointe JEP 33e éd., TALN, 27e éd., RÉCITAL, 22e éd.*, pages 361–369. ATALA; AFCP.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. *Proc. EMNLP 2020: system demonstrations*, pages 38–45. Association for Computational Linguistics.

Danny TY Wu, David A Hanauer, Qiaozhu Mei, Patricia M Clark, Lawrence C An, Joshua Proulx, Qing T Zeng, VG Vinod Vydiswaran, Kevyn Collins-Thompson, and Kai Zheng. 2016. Assessing the readability of ClinicalTrials.gov. *Journal of the American Medical Informatics Association*, 23(2):269–275.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL 2021*, pages 483–498, Online. Association for Computational Linguistics.

Daichi Yamaguchi, Rei Miyata, Sayuka Shimada, and Satoshi Sato. 2023. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 359–375.

Qing Zeng-Treitler, Hyeoneui Kim, Sergey Goryachev, Alla Keselman, Laura Slaughter, and Catherine-Arnott Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in health technology and informatics*, 129(2):1117.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *International Conference on Machine Learning*, pages 11328–11339.

11. Language Resource References

Leonardo Campillos-Llanos, Rocío Bartolomé and Ana R. Terroba. 2024. *CLARA-MeD simplified sentences*. CSIC. Distributed via Digital CSIC. PID https://doi.org/10.20350/digitalCSIC/16110.

Leonardo Campillos-Llanos, 2022. *CLARA-MeD corpus*. CSIC. Distributed via Digital CSIC. PID https://doi.org/10.20350/digitalCSIC/14644.

Daniel Ferrés and Horacio Saggion. 2022. *ALEXIS*. Universidad Pompeu Fabra. PID https://github.com/lastus-taln-upf/alexsis.

HIP. 1995. *EUGLOSS: Multilingual Glossary of technical and popular medical terms in nine European Languages*. Heymans Institute of Pharmacology.

Leonardo Campillos-Llanos. 2023. *Medical Lexicon for Spanish (MedLexSp)*. CSIC. Distributed via Digital CSIC, 1.0. PID https://digital.csic.es/handle/10261/270429.

Newsela. 2016. *Newsela Article Corpus*. PID https://newsela.com/data.

Rodrigo Alarcón, Lourdes Moreno and Paloma Martínez. 2022. *EASIER corpus*. Universidad Carlos III de Madrid. PID https://github.com/LURMORENO/EASIER_CORPUS.

Rémi Cardon and Natalia Grabar. 2018. *CLEAR corpus*. CNRS. PID http://natalia.grabar.free.fr/ress/encytotal-txt.tar.gz.

William Coster and David Kauchak. 2011. *Simple English Wikipedia*. PID https://cs.pomona.edu/ dkauchak/simplification/ data.v2/sentence-aligned.v2.tar.gz.

Appendix

Zero-shot prompting

```
Below is an instruction that describes a task. Write a response that
appropriately completes the request.
### Instruction:
Write as easy-to-read text the following text: {asymptomatic}
### Response:
```

Few-shot prompting

```
Below there are some examples to simplify medical text. Write a response
following the examples:
### Complex text:
{asymptomatic}
### Simple text:
{no symptoms}
### Complex text:
{apyretic}
### Simple text:
```

Table 10: Samples of prompts used to train the BERTIN model

Original	*Se considera mujer en edad fértil como aquellas mujeres que no hayan sido sometidas a procedimientos de infertilidad permanente o que sean amenorreicas desde hace menos de 12 meses.* (2019-004871-38) 'Women of childbearing age are considered to be those women who have not undergone permanent infertility procedures or who have been amenorrheic for less than 12 months'
Syntactic simplification	*Se consideran mujeres en edad fértil aquellas mujeres que no han sido sometidas a procedimientos de infertilidad permanente.* ***También*** *se consideran en edad fértil a quienes sean* ***amenorreicas*** *desde hace menos de 12 meses.* 'Women of childbearing age are considered to be those women who have not undergone permanent infertility procedures. Women are **also** considered of childbearing age if they have been **amenorrheic** for less than 12 months'
Lexical and syntactic simplification	*Se consideran mujeres en edad fértil aquellas que no han sido sometidas a procedimientos de infertilidad permanente.* ***También*** *se consideran en edad fértil a quienes* ***no tienen menstruación*** *desde hace menos de 12 meses*. 'Women of childbearing age are considered those who have not undergone permanent infertility procedures. Women are **also** considered of childbearing age if they have **not been menstruating** for less than 12 months.'
Original	*Los participantes con radiación en el pecho o la pared torácica pueden estar permitidos* ***si*** *la radiación torácica se documenta >6 meses antes de iniciar el tratamiento del estudio.* (2022-000131-23) 'Participants with chest or chest wall radiation may be permitted if chest radiation is documented $>$ 6 months before starting study treatment.'
Syntactic simplification	***Se permiten*** *los participantes con radiación en el pecho o la pared* ***torácica****.* ***La condición es que*** *la radiación torácica se documente durante más de 6 meses antes de iniciar el tratamiento del estudio.* 'Participants with chest or chest wall radiation are allowed. **The condition is** that chest radiation is documented $>$ 6 months before starting study treatment.'
Lexical and syntactic simplification	*Se permiten los participantes con radiación en el pecho o la pared* ***del pecho****.* ***La condición es que*** *la radiación se documente durante más de 6 meses antes de iniciar el tratamiento del estudio.* 'Participants with chest or chest wall radiation are allowed. The condition is that chest radiation is documented $>$ 6 months before starting study treatment'

Table 11: Samples of sentences used in the experiments (EudraCT id in brackets)

Score	Grammaticality / Fluency
	Is the simplified sentence grammatically correct or sufficiently readable? Does it have errors in syntax, agreement or too many words that make it difficult to read?
5	The sentence is fluent (native speaker level) and grammatically correct.
4	The sentence is nearly fluent (non-native speaker level) and grammatically correct.
3	The sentence is less fluent, with some ungrammatical but understandable parts.
2	The sentence is less fluent and with fewer grammatical parts, but is partially understandable.
1	The sentence is completely unintelligible.
Score	Semantic adequacy / Meaning preservation
	Does the simplified sentence adequately preserve the original meaning, and is it meaningless or inadequate in relation to the original?
5	The sentence adequately conveys the core meaning of the original sentence.
4	The sentence mostly conveys the essential meaning of the original sentence.
3	The central meaning of the original text is not conveyed, but some information from the original text is partially preserved.
2	The meaning of the sentence is markedly different from the original sentence, even contrary.
1	It is impossible to evaluate the meaning of the target sentence due to its unintelligibility.
Score	Simplification
	Is the resulting text simpler than the original? Is the simplification of good quality?
5	The target sentence is much simpler than the original and no information is lost.
4	The target sentence is simpler than the original and not much (or any) information is lost.
3	The target sentence is as simple/difficult as the original.
2	The target sentence is slightly more difficult than the original and/or the resulting sentence loses some relevant information.
1	The target sentence is more difficult and/or loses important information compared to the original; or it is impossible to evaluate the simplicity of the target sentence due to its unintelligibility.

Table 12: Instructions for human evaluation, translated to English and adapted from (Yamaguchi et al., 2023)

(1) Original	*Muy raras: tendinitis, artritis* 'Very rare: tendonitis, arthritis'
Only lexicon	*Muy raras: dolor e inflamación de un tendón, inflamación de una o más articulaciones* 'Very rare: pain and swelling of a tendon, swelling of one or more joints'
mBART	*Muy raras: tendinitis (inflamación de los tendinios), artritis (inflamación de la artritis)* 'Very rare: tendonitis (swelling of tendinios), arthritis (swelling of arthritis)'
mBART + lexicon	*Muy raras: tendinitis, artritis (enfermedad de los huesos largos).* 'Very rare: tendonitis, arthritis (long bone disease)'
mBART + lexicon + postproces.	*Muy raras: tendinitis, artritis (enfermedad de los huesos largos).* 'Very rare: tendonitis, arthritis (long bone disease)'
BERTIN + lexicon	*Muy raras: tendinitis, artritis* (Not simplified) 'Very rare: tendonitis, arthritis'
(2) Original	*Disección radical de ganglios linfáticos.* 'Radical dissection of lymph nodes'
Only lexicon	*Disección radical de ganglios linfáticos (nódulos que forman parte del sistema inmunitario).* 'Radical dissection of lymph nodes (nodes that are part of the immune system)'
mBART	*Dificultad para abrir los ganglios linfáticos.* 'Difficulty opening the lymph nodes'
mBART + lexicon	*A veces se administran ganglios linfáticos a través de la piel.* 'Sometimes, lymph nodes are administered through the skin.'
mBART + lexicon + postproces.	*A veces se administran ganglios linfáticos (nódulos que forman parte del sistema inmunitario) a través de la piel.* 'Sometimes, lymph nodes (nodes that are part of the immune system) are administered through the skin.'
BERTIN + lexicon	*Disección radical de ganglios linfáticos (extirpación de todos los ganglios linfáticos de una zona del cuerpo)* 'Radical dissection of lymph nodes (removal of all lymph nodes from a body part)'
(3) Original	*Rabdomiosarcoma* 'Rhabdomyosarcoma'
Only lexicon	*Rabdomiosarcoma (cáncer de tejidos blandos (como el músculo), tejido conectivo (como el tendón o el cartílago) o el hueso).* 'Rhabdomyosarcoma (cancer of soft tissues (such as muscle), connective tissue (such as tendon or cartilage) or bone'
mBART	*Rabdomiosarcoma (enfermedad autoinmune del sistema inmunitario)* 'Rhabdomyosarcoma (autoimmune disease of the immune system)'
mBART + lexicon	*Rabdomiosarcoma (cáncer de tejidos blandos (como el músculo), tejido conectivo (como el tendón o el cartílago) o el hueso).* 'Rhabdomyosarcoma (cancer of soft tissues (such as muscle), connective tissue (such as tendon or cartilage) or bone'
mBART + lexicon + postproces.	*Rabdomiosarcoma (cáncer de tejidos blandos (como el músculo), tejido conectivo (como el tendón o el cartílago) o el hueso).* 'Rhabdomyosarcoma (cancer of soft tissues (such as muscle), connective tissue (such as tendon or cartilage) or bone'
BERTIN + lexicon	*Rabdomiosarcoma (cáncer de tejido blando que se origina en los músculos)* 'Rhabdomyosarcoma (soft tissue cancer originating in the muscles)'

Table 13: Samples of original and simplified excerpts used in the human evaluation