Integrating Linguistic Knowledge into Prompting Strategies for Spanish Text Simplification: Insights from the NIL-UCM participation

Daniel Fernández[1], Alberto Díaz[1]

[1]*Facultad de Informática and ITC, Universidad Complutense de Madrid, Spain*

Abstract

We present our participation in a text simplification shared task focused on Plain Language and Easy-to-Read. Our approach, based on explicit linguistic instructions, yielded good results in semantic fidelity and readability. However, the metrics used—such as the Fernández-Huerta index—are insufficient to capture the complexity of the task. Fine-tuning did not significantly outperform prompt-based generation. We highlight the need for more robust and multidimensional metrics to enable fairer and more accurate evaluation of text simplification.

Keywords

Automatic simplification, Plain Language, Easy-to-Read, Readability metrics, NLP

1. Introduction

This paper presents the participation of the NIL-UCM team in the CLEARS challenge [1], held as part of IberLEF [2]. Considering the growing demand for accessible texts through various methods, particularly Plain Language and Easy-to-Read, this challenge aimed to advance the automation of these processes using Natural Language Processing and Machine Learning techniques. Our approach leverages a Large Language Model (LLM) by providing explicit linguistic instructions within the prompt to incorporate linguistic knowledge.

1.1. Task Description

The competition included two distinct subtasks, allowing participants to take part in either one or both. The goal was the same in both cases: to automatically simplify a corpus of around six hundred texts [3], with the difference being the methodology applied in each. Thus, in Subtask 1, the aim was to ensure that administrative and news texts used clear and understandable language following the recommendations of Plain Language, while Subtask 2 required applying the specific criteria of Easy-to-Read, and more specifically, the guidelines established in the UNE 153101 EX standard [4].

1.2. Dataset Description

The dataset provided to participants [5] consisted of five CSV files. There is a clear division between training and test texts, as well as between Subtask 1 and Subtask 2.

The training files, which include both the original and the simplified versions of the texts according to each methodology, allow participants to train their models:

- `OriginalTrain.csv`. Contains the 2,400 original training texts, which are shared between both subtasks. As mentioned earlier, these are news articles of varying lengths, ranging from approximately 300 to 1300 words, mostly reporting on local events and occurrences in various municipalities of the province of Alicante, Spain.

IberLEF 2025, September 2025, Zaragoza, Spain

dferna17@ucm.es (D. Fernández); albertodiaz@fdi.ucm.es (A. Díaz)

0000-0003-1966-3421 (A. Díaz)

It is worth noting that the original texts are not particularly complex to begin with. In our view, this is especially problematic in the case of Subtask 1, considering that Plain Language, as we will see, is designed to adapt documents that are especially difficult for the average reader, often due to their legal nature.

- `Subtask1Train.csv`. Consists of the 2,400 training texts manually simplified according to Plain Language guidelines.
- `Subtask2Train.csv`. Comprises the same 2,400 training texts, this time manually simplified according to the UNE 153101 EX standard [4].

To better understand the nature of the original texts and those provided to participants as references for adaptation according to each methodology, we include below one of the texts in its three versions:

Text 25. Original

"Alicante, 25 de noviembre del 2022. El concejal de Deportes, José Luis Berenguer, ha presentado esta mañana el acto del XII Encuentro Club Esportiu Aquarium Alacant de natación adaptada junto al presidente de la entidad deportiva CE Aquarium, Jorge Chica, y el nadador Manuel Martínez. El evento que tiene lugar mañana sábado se celebrará en la Piscina Monte Tossal 'José Antonio Chicoy', a las 17 horas, y está organizado por el Club Aquarium, la Federación de Deportes Adaptados de la Comunitat Valenciana junto con la Concejalía de Deportes del Ayuntamiento de Alicante. El XII encuentro reunirá a numerosos deportistas de diferentes clubes de natación adaptada de ámbito nacional.El edil de Deportes José Luis Berenguer (Cs) ha comentado que 'siempre apoyamos este tipo de iniciativas porque pensamos que practicar deporte es la mejor terapia y medicina ya que es saludable y hace mucho bien a las personas'. El concejal ha destacado también que el Club Esportiu Aquarium Alacant 'hace una gran labor para con el deporte de nuestra ciudad y con los deportistas de su Club en particular, además de llevar a cabo este tipo de eventos'.Por otra parte, Berenguer ha señalado que el club 'cuando participa en competiciones siempre obtiene muy buenos resultados, lo que demuestra que el trabajo que hacen lo hacen muy bien'. Para el Ayuntamiento de Alicante y para la Concejalía de Deportes es 'muy satisfactorio que haya entidades como esta que se vuelcan tanto con el deporte y que llevan el nombre de Alicante con orgullo allá donde van'".

Text 25. Plain Language

"El concejal de Deportes de Alicante presentó el 12 Encuentro Club Esportiu Aquarium Alacant de natación adaptada que organiza el Club Aquarium, la Federación de Deportes Adaptados de la Comunitat Valenciana y la Concejalía de Deportes del Ayuntamiento de Alicante. El encuentro será en la Piscina Monte Tossal el sábado 26 de noviembre a las 5 de la tarde y participarán deportistas de diferentes clubes de natación adaptada. El concejal dijo que es muy importante apoyar estas actividades y el buen trabajo del Club Esportiu Aquarium Alacant para promocionar el deporte en Alicante. También dijo que el club tiene buenos resultados en las competiciones y que es bueno tener entidades que difunden una buena imagen de Alicante".

Text 25. Easy-to-Read

"Alicante presenta el encuentro número 12 de natación adaptada
que organiza el Club Esportiu Aquarium Alacant.

El concejal de Deportes de Alicante
presentó el 12 Encuentro Club Esportiu Aquarium Alacant de natación adaptada
que organiza el Club Aquarium,
la Federación de Deportes Adaptados de la Comunitat Valenciana
y la Concejalía de Deportes del Ayuntamiento de Alicante.

El encuentro será en la Piscina Monte Tossal
el sábado 26 de noviembre

a las 5 de la tarde.
Participarán deportistas de diferentes equipos de natación adaptada.

El concejal dijo que es muy importante apoyar estas actividades
para promocionar el deporte en Alicante
y que el Club Esportiu Aquarium está dando una buena imagen de la ciudad.
También dijo que el club está haciendo un buen trabajo a favor del deporte
y que tiene buenos resultados en las competiciones".

As for the test files, they contain the original texts that must be adapted automatically and are specifically intended for model evaluation. These texts belong to the same domain and are similar in length to those in the training set. As expected, they differ from the texts used during training.

- `Subtask1Test.csv`. Consists of 607 original texts that must be adapted to Plain Language.
- `SubTask2Test.csv`. Contains 600 original texts to be adapted to Easy-to-Read. These are the same texts as in the previous file, except for the last seven, which are not included for some reason.

2. Plain Language and Easy-to-Read Language

As we will explain in a later section, our approach to solving the two tasks in the competition has been based on integrating certain linguistic knowledge into our models. For this reason, we believe it is necessary to devote some attention to the sociolinguistic foundations behind both methodologies of text adaptation.

2.1. Plain Language

Plain Language is an initiative that emerged in the 1960s in the United States, the United Kingdom, Canada, and Australia, and since then it has spread to many countries, where in some cases legislation has even been enacted on the matter. As Petelin [6] points out, "the key principle of plain language is that the intended reader can use the document for its intended purpose." Unlike Easy-to-Read Language, it is not aimed at a specific group, but rather seeks to prevent texts from using overly convoluted and complex language—something that often occurs in legal and financial domains, the areas in which these recommendations have seen the most development precisely for that reason. So much so that Tartaglia [7] directly links the complexity of legal language and complaints about its dense and incomprehensible nature to the birth of the Plain Language Movement.

In her renowned book *Plain Language for Lawyers*, Asprey [8] perfectly summarizes the philosophy of this movement when she states: "Simple in this sense doesn't mean simplistic. It means straightforward, clear, precise. Writing in plain language is just writing in clear, straightforward language, with the needs of the reader foremost in mind (...). The main thing to remember is that if what you have written could be unclear or confusing for your reader, or difficult to read, you should rewrite it so that it becomes clear, unambiguous and easy to read."

It is recommended to consider the target user of the text and adapt its content, structure, and visual design to their needs, eliminating anything that could cause confusion or hinder readability.

Particularly relevant is the guide *How to Write Clearly*, published by the European Commission and primarily addressed to EU staff: "European Commission staff have to write many different types of documents. Whatever they type – legislation, a technical report, minutes, a press release or speech — a clear document will be more effective and more easily and quickly understood" [9]. We based our model on these recommendations, along with those found in the *Plain English Handbook* [10], published by the Office of Investor Education and Assistance of the SEC, the U.S. agency that regulates the stock market.

2.2. Easy-to-Read

Easy-to-Read is a cognitive accessibility tool "that brings together a set of guidelines and recommendations regarding text drafting, document design/layout, and the validation of their comprehensibility, aimed at making information accessible to people with reading comprehension difficulties" [4]. It is part of a broader strategy aimed at facilitating cognitive accessibility by removing barriers to the comprehension, interaction, and use of products and services. In this way, it helps to guarantee the right of access to information that people with cognitive disabilities are legally entitled to. Although Easy-to-Read documents can be useful for all users, they are specifically intended for individuals who experience reading comprehension difficulties.

The goal is to eliminate the barriers that people with reading comprehension difficulties may face in all aspects of life. As Hurtado and Reguera [11] note, public administrations are increasingly demanding Easy-to-Read texts, and interest in text adaptation has grown significantly.

In order to integrate this linguistic knowledge into our model, we have compiled the most relevant guidelines and recommendations on the subject. Our main reference has been the experimental standard UNE 153101:2018 EX [4]. This standard, which is the primary reference in Spain, includes both mandatory guidelines and recommendations, often illustrated with incorrect and correct examples. In fact, it is the document proposed as a reference for Subtask 2. García Muñoz [12] is also highly relevant, as he systematically presents drafting and evaluation proposals based on previous experiences.

Although both documents also include proposals related to design and layout, we have focused on those related to orthotypography, lexis, morphosyntax, style, and the organization of information.

2.3. Evaluation

The competition guidelines stated that the evaluation would consider both the lexical and semantic similarity between the original and adapted texts, as well as the readability of the latter:

- Cosine similarity (*Bag-of-Words*) to measure lexical overlap between the reference texts and the participants' submissions.
- Cosine similarity (*Embeddings*) to assess semantic similarity between the adapted and original texts.
- Fernández-Huerta readability index, aimed at measuring the readability of the adapted texts, in accordance with plain and accessible language guidelines.

We believe that the first two metrics—based on cosine similarity of sparse and dense vector representations of the original and adapted texts—are appropriate. We were more doubtful about the use of the Fernández-Huerta index. Fernández-Huerta [13] adapted Flesch's Reading Ease Score (RES) to Spanish, adjusting it to the linguistic characteristics of the language. Since its publication, it has become a pioneering reference for assessing the readability of texts in Spanish, particularly in educational contexts [14]. The formula is expressed as follows:

$$\text{Readability} = 206.84 - 0.60P - 1.02F$$

where P is the number of syllables per 100 words and F is the number of sentences per 100 words.

Based on this metric, Fernández-Huerta [13] established a classification into seven levels, each corresponding to an educational stage:

- **0–30**: very difficult – university level
- **30–50**: difficult – pre-university level
- **50–60**: fairly difficult – ages 13–16
- **60–70**: standard – ages 10–12
- **70–80**: fairly easy – age 9
- **80–90**: easy – age 6

- **90–100**: very easy – age 5

As can be seen, the Fernández-Huerta index, like Flesch's RES, is based on the assumption that the difficulty or readability of a text is determined by its length and by the number of words and sentences it contains [14]. One advantage of this type of formula is that its measurement can be easily automated, allowing for the evaluation of large amounts of text in a short time. However, this approach seems somewhat simplistic to us, and it is striking that the same metric is used to evaluate both subtasks, despite the fact that Plain Language and Easy-to-Read, while sharing some common elements, are based on different methodologies.

3. Methodology

3.1. Model Selection

To carry out the task of automatic text simplification, we employed Mistral-7B-Instruct-v0.3, a large language model (LLM) with 7 billion parameters, designed for automatic text generation. This model is a fine-tuned version of the base model Mistral-7B-v0.3, specifically trained to follow human instructions. This fine-tuning process, known as instruct fine-tuning, enables the model to respond more helpfully, coherently, and in a task-oriented manner. Thanks to its ability to interpret and execute instructions, the model is particularly suitable for tasks such as text simplification, which require transforming content while preserving its original meaning but reducing its linguistic complexity[1].

Several features proved decisive in selecting this model, as they make it especially suitable for the task of automatic text simplification. Its instruction-based fine-tuning allows the model to understand and execute specific directives, which is essential for transforming complex texts into more accessible versions without losing their original meaning. This fine-tuning equips the model to follow concrete linguistic instructions.

Mistral-7B-Instruct-v0.3 met all the requirements established by our methodology:

1. A model specialized in coherent and fluent text generation.
2. No need for additional fine-tuning with the competition's training data.
3. Ability to follow precise linguistic instructions.
4. Significantly lower cost compared to models from OpenAI, Meta, and Gemini.

3.2. Implementation

This model was loaded using Hugging Face's `transformers` library, which provides a modular and standardized interface for working with a wide variety of language models. In particular, we used the `AutoModelForCausalLM` and `AutoTokenizer` classes. The `AutoModelForCausalLM` class allows for loading pretrained autoregressive language models designed for causal text generation tasks, i.e., predicting the next token based on previous ones. This abstraction simplifies the loading of the specific model (in this case, Mistral-7B-Instruct-v0.3) without manually defining its architecture, as the class automatically adapts the appropriate structure and weights.

The `AutoTokenizer` class handles the tokenization and detokenization of text, converting text strings into numerical sequences (tokens) that the model can process and vice versa. This tokenizer also automatically adapts to the loaded model, ensuring consistency between the vocabulary and the encoding used.

The implementation was developed in Python, using PyTorch as the backend for processing, with GPU acceleration enabled. It is worth noting that the task was particularly demanding from a computational standpoint, requiring intensive use of GPU memory, reaching approximately 35 GB of VRAM to handle the model and generate text.

[1]The model description is available at https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

We explored two different approaches to applying the model to the simplification task: on the one hand, we tested its performance without additional fine-tuning, relying solely on its ability to follow explicit linguistic instructions via prompts. On the other hand, we also evaluated the model's performance after fine-tuning it with the competition's training texts.

The simplification procedure consists of constructing a prompt composed of externally defined linguistic instructions (contained in a text file) followed by the original text, in the following structure:

`{instructions}\nOriginal text: {text}\nRewritten text:`

This *prompt* is tokenized using the model's tokenizer, generating the input tensors required for inference. Text generation was carried out using the model's `generate` method, with parameters configured to optimize output quality and diversity, such as:

- `max_length=10000`
- `do_sample=True`
- `top_p=0.9`
- `temperature=0.7`

In addition, the padding token was set to the end-of-sequence token (`pad_token_id=tokenizer.eos_token_id`) to ensure proper padding management.

The generation process was executed without gradient computation (`torch.no_grad()`), reducing computational resource consumption. After generation, the portion corresponding to the simplified text was extracted based on the textual marker "`Rewritten text:`", ensuring the retrieval of relevant content.

To avoid input size limitations, it was verified that the number of tokens in the *prompt* did not exceed a maximum threshold (`max_tokens=10000`). If this limit was exceeded, the text was marked as unprocessed. Finally, the simplified texts were stored in a `pandas` `DataFrame` and exported to `CSV` format for subsequent analysis and evaluation.

3.3. Fine-tuning

As previously noted, the Mistral-7B-Instruct-v0.3 model was adapted to two specific tasks: Plain Language simplification and Easy-to-Read simplification. For this purpose, an independent fine-tuning process was carried out for each task, employing efficient training techniques designed for computationally constrained environments. This procedure was repeated separately in both cases, thereby producing two models each tailored to a specific simplification objective.

Specifically, a parameter-efficient adaptation strategy known as Low-Rank Adaptation (LoRA) was applied to a 4-bit quantized version of the base model, significantly reducing computational requirements without substantially compromising performance.

The data used for training, validation, and testing were organized into three datasets (`train.csv`, `val.csv`, `test.csv`), composed of input-output pairs (`text`, `expected`), where each input corresponds to an original text and its respective simplified version. These data were converted into instances of the Hugging Face `datasets.Dataset` class, enabling seamless integration into the tokenization and training pipeline.

Each dataset instance was transformed into an instructive prompt following the format used by instruction-tuned models:

```
<s>[INST] {instructions}:
{text} [/INST] {expected}</s>
```

where `{instructions}` corresponds to a set of general guidelines for the text simplification task, loaded from an external file (`prompt.md`). For supervised learning purposes, tokens corresponding to the prompt were masked in the labels (`labels`) using the value `-100`, so that the loss function is applied exclusively to the tokens generated by the model's response.

To optimize memory usage during training, the `bitsandbytes` library was employed to load the Mistral-7B-Instruct model in 4-bit quantization mode. This configuration uses the `nf4` quantization technique with `float16` computation and double quantization enabled, providing a balanced trade-off between efficiency and precision.

The LoRA technique was used to insert trainable adapters into a specific subset of model layers: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The LoRA hyperparameters were set to `r=16` (decomposition rank) and `lora_alpha=32`, with a dropout rate (`dropout`) of 0.05. Adaptation was performed using the `peft` library, allowing only the LoRA-added layers to be updated during training while keeping the original base model weights frozen.

Training was conducted using the Hugging Face `Trainer` class with a configuration adapted for resource-limited environments. An effective batch size of 4 was achieved through gradient accumulation (`gradient_accumulation_steps=4`), and a learning rate of `2e-4` was used. To reduce computational cost, training was limited to a maximum of 3 epochs, which allowed the model to be fine-tuned without placing excessive demands on the available resources. Training was capped at a maximum of 1000 steps, with evaluation and model checkpointing performed every 200 steps. The optimizer employed was `paged_adamw_8bit`, specifically designed for quantized models. Finally, the fine-tuned model was saved to disk along with its corresponding tokenizer.

3.4. Prompt Design for Text Simplification

As previously indicated, the adopted strategy was based on the use of prompts with explicit and structured instructions for each subtask. This formulation was designed by leveraging the linguistic knowledge acquired throughout the research on Plain Language and Easy-to-Read guidelines, with the aim of optimizing interaction with an instruction-tuned model, thereby maximizing its ability to generate outputs aligned with the defined simplification criteria.

Rather than modifying the model's parameters, prompts allow the integration of pretrained models into specific tasks simply through the appropriate formulation of instructions. These instructions, which can be expressed in natural language or as learned vector representations, guide the model to produce the desired behavior by activating the relevant knowledge according to the provided context [15].

Our prompt-based approach offers a significant advantage in terms of computational efficiency, and it also enables us to direct text simplification following a specific methodology, which, in the case of Easy-to-Read, is particularly thorough. Nonetheless, our intuition was that the combined use of a detailed prompt with linguistic instructions and fine-tuning using the original and manually adapted texts would yield the best results. However, it was necessary to verify this and assess whether the difference compared to the non-finetuned model justified the computational cost of fine-tuning.

3.4.1. Prompt Design for Plain Language

The prompt used in Subtask 1 was designed in accordance with the guidelines and recommendations of Plain Language outlined in a previous section. Its elaboration was highly detailed and included multiple examples, both correct and incorrect, for each instruction, since—as already mentioned—this strategy has proven to be highly effective. As a result, the prompt reached a considerable length, with a total of 19,617 characters.

It is divided into two main sections: "Mission" and "Instructions". The first section, in which we provided the necessary context and introduced general instructions, can in turn be divided into three parts:

1. Description of the task and the general behavior expected of the model.
2. Prohibition against including explanations of the changes applied in the simplifications.
3. General example of simplification.

The prohibition against including explanations was necessary because in preliminary tests we observed that the model had included a list of the changes made in some of the texts. It is worth noting

that even this was not sufficient to prevent such output entirely, and the changes had to be removed from the final result using regular expressions.

In the "Instructions" section, we introduced specific guidelines in a highly detailed manner, including several correct and incorrect examples for each one. When multiple guidelines were closely related, we included them under the same entry. We based these on the linguistic and methodological knowledge of Plain Language discussed in a previous section, and the instructions covered lexical, syntactic, morphological, and information structure domains.

3.4.2. Prompt Design for Easy-to-Read

For the development of the prompt focused on Easy-to-Read, we adopted a similar approach, this time following the guidelines and recommendations detailed in the corresponding section. Our prompt aimed to reflect the higher level of precision and thoroughness characteristic of Easy-to-Read, as evidenced by its even greater length of 21,726 characters. The lexical, syntactic, morphological, and information structure dimensions were maintained, albeit with the specific features of this methodology. In fact, some instructions—although worded differently—were essentially the same as those in the prompt for Subtask 1. As we have previously explained, the two methodologies share many points in common, despite their notable differences. Additionally, we included some considerations related to formatting and the need for glosses, which are highly significant aspects in Easy-to-Read practices.

The prompt is divided into eight parts, such that general instructions are followed by seven sections grouping a substantial number of guidelines along with their examples: general objective, punctuation marks and symbols, line formatting, lexis, numbers, morphosyntax, and information structure.

4. Results

4.1. Preliminary Evaluation with the Test Set

Before proceeding with the automatic simplification of the more than six hundred texts included in the competition—a task with high computational cost—we chose to validate our approach in advance. To this end, we applied the three metrics described above to the test subset of the training texts, which we split into train, validation, and test. Our main interest was to compare the results obtained with the manually adapted texts.

We carried out this validation both with the texts generated using only the prompt and with those simplified through a combination of prompt and fine-tuning, although we limited this analysis to Subtask 2. This decision was based on the fact that, at this stage, our goal was merely to assess the feasibility of the approach. In any case, Subtask 1 would later be evaluated on the full set of the six hundred and seven competition texts.

In order to assess the quality of the texts generated by the automatic simplification system, we applied three complementary metrics: lexical similarity, semantic similarity, and readability score.

1. **Lexical Similarity (Bag-of-Words)**: We used the Bag-of-Words approach with `CountVectorizer` to represent texts as frequency vectors. Then, cosine similarity was calculated between the original texts, the reference texts, and the simplified outputs. This metric estimates the degree of lexical overlap without taking word order or meaning into account.
2. **Semantic Similarity (embeddings)**: To capture meaning relations beyond surface-level lexicon, we employed the multilingual model `paraphrase-multilingual-MiniLM-L12-v2` from *Sentence Transformers*, which generates dense sentence embeddings. We again computed cosine similarity between various pairs of texts, allowing us to evaluate semantic content preservation in the simplified versions.
3. **Readability (Fernández-Huerta Index)**: Finally, we applied the Fernández-Huerta index to compare the readability of the original, reference, and simplified texts. It is particularly interesting to compare not only the reference and generated texts, but also the originals, in order to assess whether there are significant differences in readability.

Taken together, these three metrics provide a multidimensional evaluation encompassing lexical fidelity, semantic equivalence, and textual accessibility of the generated simplifications.

In the case of simplification using only the prompt, the results of this preliminary evaluation were positive, as the generated texts outperformed the manually simplified references in both lexical and semantic similarity, while scoring only slightly lower in readability.

1. **Lexical Similarity**
    - Original – Reference: 0.8408
    - Original – Generated: 0.8963
2. **Semantic Similarity**
    - Original – Reference: 0.8193
    - Original – Generated: 0.8498
3. **Fernández-Huerta**
    - Original: 51.402 (somewhat difficult)
    - Reference: 76.3211 (somewhat easy)
    - Generated: 76.0391 (somewhat easy)

For the fine-tuned model, the results were similar, although it is worth noting that the readability index obtained in this case was slightly lower than that of the other model.

1. **Lexical Similarity**
    - Original – Reference: 0.8408
    - Original – Generated: 0.9061
2. **Semantic Similarity**
    - Original – Reference: 0.8193
    - Original – Generated: 0.8593
3. **Fernández-Huerta**
    - Original: 51.402 (somewhat difficult)
    - Reference: 76.3211 (somewhat easy)
    - Generated: 75.3677 (somewhat easy)

In short, the results show that the texts generated by both the base and the fine-tuned models exhibit high similarity to the original texts, particularly in lexical and semantic terms. In both cases, lexical similarity was even higher between the original and the generated texts (0.8963 and 0.9061) than between the original and the reference texts (0.8408), suggesting a considerable preservation of original vocabulary. Similarly, semantic similarity was also higher in the original–generated comparison (0.8498 and 0.8593) than in original–reference (0.8193), indicating that the simplified texts maintain the original meaning effectively. Finally, the Fernández-Huerta readability index reveals a notable increase in reading ease: while the original texts are classified as "somewhat difficult" (51.4), both the reference and generated texts reach "somewhat easy" levels (between 75 and 76). Although the reference texts—i.e., those manually simplified—achieved a slightly higher readability score, the difference compared to the automatically generated texts is minimal, further supporting the effectiveness of the model. Moreover, the fact that the base model achieved such competitive results suggests that the additional computational cost of fine-tuning may not be justified, especially given the relatively small improvements in similarity. Nevertheless, this hypothesis will be tested on the texts to be simplified for the competition.

4.2. Results of Subtask 1

For the analysis of the texts generated in Subtask 1, we present below both the results obtained using our implementation of the three metrics—focused on lexical similarity, semantic similarity, and the Fernández-Huerta readability index—and the official results published at the end of the competition. This comparison enables a more comprehensive assessment of the models' performance by contrasting internal automatic evaluation with the external evaluation provided by the competition organizers. Unlike the latter, the internal evaluation analyzed the performance of both the fine-tuned and non-fine-tuned models, offering the advantage of directly comparing the impact of fine-tuning on simplification metrics.

4.2.1. Internal Evaluation

1. **Lexical similarity**
   - Prompt only: 0.8615
   - Fine-tuning: 0.8671
2. **Semantic similarity**
   - Prompt only: 0.8388
   - Fine-tuning: 0.8339
3. **Fernández-Huerta index**
   - Original: 75.5377 (somewhat easy)
   - Prompt only: 80.8373 (easy)
   - Fine-tuning: 83.3521 (easy)

As can be observed, the differences between the instruction-only model (prompt only) and the fine-tuned model are minimal in terms of both lexical similarity (0.8615 vs. 0.8671) and semantic similarity (0.8388 vs. 0.8339), suggesting that fine-tuning does not lead to substantial improvements. However, the difference is somewhat more pronounced in the case of the Fernández-Huerta readability index, which increases from 80.8373 with the non-fine-tuned model to 83.3521 with the fine-tuned one.

4.2.2. External Evaluation

For the external evaluation, the organizers employed a lexical similarity approach based on a bag-of-words model using TF-IDF vectors, as opposed to our internal evaluation, which relied on raw frequency vectors through `CountVectorizer`. Two separate rankings were published: one based on the average of lexical and semantic similarity (cosine similarity with TF-IDF and embeddings), and another based on the Fernández-Huerta readability index. Our team, NIL-UCM, achieved competitive results in the first ranking, securing second place with an average cosine similarity of 0.71, just behind HULAT-UC3M (0.75), and ahead of CARDIFFNLP and VICOMTECH (both 0.70). In terms of individual metrics, NIL-UCM obtained a lexical similarity score of 0.67 and a semantic similarity score of 0.75. However, in the ranking based on readability, our system ranked third, with an average Fernández-Huerta index of 70.42, behind VICOMTECH (82.98) and CARDIFFNLP (78.81), but ahead of HULAT-UC3M (69.72). This discrepancy suggests that while our approach preserved lexical and semantic content effectively, there may still be room for improvement in optimizing textual accessibility.

Ranking based on the average of cosine similarities (TF-IDF / Embeddings):

1. HULAT-UC3M: 0.71 / 0.78 (average: 0.75)
2. NIL-UCM: 0.67 / 0.75 (average: 0.71)
3. CARDIFFNLP and VICOMTECH: 0.63 / 0.77 (average: 0.70)

Ranking based on Fernández-Huerta readability index:

1. VICOMTECH: 82.98

2. CARDIFFNLP: 78.81
3. NIL-UCM: 70.42
4. HULAT-UC3M: 69.72

4.3. Results of Subtask 2

For Subtask 2, we present an analysis similar to that of Subtask 1, reporting internal results based on the same three metrics and comparing them with the official evaluation. Again, we distinguish between the fine-tuned model and the base model in order to assess the impact of fine-tuning on this subtask.

4.3.1. Internal Evaluation

1. **Lexical similarity**
   - Prompt only: 0.8481
   - Fine-tuning: 0.8787
2. **Semantic similarity**
   - Prompt only: 0.8280
   - Fine-tuning: 0.8396
3. **Fernández-Huerta Index**
   - Original: 75.6803 (somewhat easy)
   - Prompt only: 82.7251 (easy)
   - Fine-tuning: 82.1256 (easy)

The results obtained for Subtask 2 are generally positive. The fine-tuned model achieves slightly better performance than the base model on the lexical (0.8787 vs. 0.8481) and semantic (0.8396 vs. 0.8280) similarity metrics, particularly the former. In contrast, for the Fernández-Huerta readability index, the prompt-only model achieves a slightly higher score (82.7251) than the fine-tuned model (82.1256), although both fall within the range of texts considered easy to read.

4.3.2. External Evaluation

As in Subtask 1, Subtask 2 includes two different rankings: one based on the average of lexical similarity (measured with TF-IDF) and semantic similarity (measured with embeddings), and another based on the Fernández-Huerta readability index. Our team, NIL-UCM, achieved the highest score in the first ranking, with a global average similarity of 0.72, due to high values in both lexical similarity (0.68) and semantic similarity (0.75). This was followed by CARDIFFNLP (0.71) and UR (0.70), while UNED-INEDA and VICOMTECH ranked fourth and fifth with 0.68 and 0.66, respectively. However, the ranking based on the Fernández-Huerta index, which evaluates the readability of the generated texts, shows an inverse pattern: VICOMTECH leads with an average of 85.44, closely followed by UR (85.12). In this case, our team ranks fifth with a score of 69.40, indicating that although our simplifications exhibit a high degree of similarity with the references, there is still room for improvement in terms of readability—similarly to what was observed in Subtask 1.

Ranking based on the average of cosine similarities (TF-IDF / Embeddings):

1. NIL-UCM: 0.68 / 0.75 (average: 0.72)
2. CARDIFFNLP: 0.65 / 0.77 (average: 0.71)
3. UR: 0.64 / 0.76 (average: 0.70)
4. UNED-INEDA: 0.60 / 0.75 (average: 0.68)
5. VICOMTECH: 0.58 / 0.74 (average: 0.66)

Ranking based on Fernández-Huerta readability index:

1. VICOMTECH: 85.44
2. UR: 85.12
3. CARDIFFNLP: 77.85
4. UNED-INEDA: 72.39
5. NIL-UCM: 69.40

5. Discussion

As noted in a previous section, the Fernández-Huerta index is not an appropriate metric for evaluating readability in the context of automatic text simplification. This measure, which focuses exclusively on superficial aspects such as word and sentence length, fails to capture the complexity of the processes involved in tasks such as Plain Language and Easy-to-Read. Using the same metric for both subtasks is problematic, as it does not reflect the methodological differences or specific objectives of each. Although our results on this metric were outperformed by those of other teams in both tasks, we consider it necessary to adopt more comprehensive and multidimensional indicators that can adequately assess the various aspects involved in text adaptation. In this regard, the prompts employed in our proposals provide detailed guidance that addresses multiple dimensions of simplification (lexical, morphosyntactic, information organization, etc.); however, their effectiveness is only partially reflected through such a limited metric.

Another important aspect concerns the nature of the original texts used in the tasks. These texts already exhibit a high readability index, which limits the potential for improvement. As previously mentioned, this issue is particularly relevant in the subtask focused on Plain Language. It is worth recalling that, in the test subset corresponding to the training texts, the average Fernández-Huerta index of the original texts was significantly lower, which allowed the effectiveness of our simplification methodology to stand out more clearly.

As shown, our models perform particularly well in terms of lexical and semantic similarity between the original texts and their simplified versions, indicating a high level of fidelity to the source content. Moreover, the readability results are also satisfactory. In fact, for the test texts from the training set, the Fernández-Huerta scores often match or even exceed those obtained by the reference simplified texts. Regarding the competition texts, while the performance of other teams suggests there is room for improvement, it is important to emphasize that the limitations of the metric used hinder a fair and comprehensive evaluation of model performance.

With regard to further training via fine-tuning, the results are inconclusive. In Subtask 1, some improvement in readability is observed with the fine-tuned model (83.35 vs. 80.84), whereas in Subtask 2, the results are slightly lower than those of the base model (82.75 vs. 82.13). In light of these findings, one may question whether the computational cost of fine-tuning is justified, as its benefits do not appear to be consistent or significant. It should be noted, however, that fine-tuning was limited to only three epochs to avoid excessive computational costs, and further experimentation with longer training could potentially yield different results.

Finally, it is worth commenting on the impact of the prompts used. The prompt corresponding to Subtask 1 appears to be the most effective for guiding textual simplification, as suggested by the fact that the highest overall readability score was obtained with the fine-tuned model in that task. Nevertheless, this conclusion must be qualified due to the limitations of the metric employed. Furthermore, the difference between the best result for the Subtask 1 prompt (fine-tuned model: 83.96) and the best result for the Subtask 2 prompt (non-fine-tuned model: 82.73) is not statistically significant. Similarly, in terms of lexical and semantic similarity, no substantial differences are observed either between subtasks or between base and fine-tuned models, which further highlights the need for a more in-depth analysis using a combination of metrics for a more robust evaluation. As noted previously, since fine-tuning was limited to only three epochs to manage computational costs, further investigation with extended training is necessary to fully assess its potential impact.

6. Conclusion

This challenge represents a significant step forward in the field of automatic text simplification, a task with important social implications, particularly regarding the right to access information for individuals with cognitive disabilities. Such technologies can contribute to a more inclusive society, in which all citizens are able to exercise their rights on equal terms.

Our approach, based on the use of explicit linguistic instructions, has proven effective in terms of semantic fidelity and readability. Nevertheless, while the results are encouraging, we believe there is still room for improvement. To continue making progress, it is essential to adopt more precise and comprehensive evaluation metrics that better reflect the complexity of the processes involved in text simplification.

In conclusion, we consider it a priority to move towards more accurate and multidimensional evaluations, which allow for a fairer and more comprehensive assessment of the different approaches to automatic simplification.

7. Acknowledgements

This publication is part of the R&D&I project HumanAI-UI, Grant PID2023-148577OB-C22 (Human-Centered AI: User-Driven Adaptative Interfaces-HumanAI-UI) funded by MICI-U/AEI/10.13039/501100011033 and by FEDER/UE.

8. Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] B. Botella-Gil, I. Espinosa-Zaragoza, A. Bonet-Jover, M. Madina, L. Molino Piñar, P. Moreda, I. Gonzalez-Dios, M. T. Martín Valdivia, Ureña, Overview of clears at iberlef 2025: Challenge for plain language and easy-to-read adaptation for spanish texts, Procesamiento del Lenguaje Natural 75 (2025).

[2] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[3] I. Espinosa-Zaragoza, J. Abreu-Salas, P. Moreda, M. Palomar, Automatic text simplification for people with cognitive disabilities: Resource creation within the ClearText project, in: S. Štajner, H. Saggio, M. Shardlow, F. Alva-Manchego (Eds.), Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 68–77. URL: https://aclanthology.org/2023.tsar-1.7/.

[4] AENOR, UNE 153101 EX. Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos, AENOR, 2018.

[5] B. Botella-Gil, I. Espinosa-Zaragoza, P. Moreda, M. Palomar, Corpus ClearSim, 2024. URL: http://hdl.handle.net/10045/151688.

[6] R. Petelin, Considering plain language: issues and initiatives, Corporate Communications: An International Journal 15 (2010) 205–216.

[7] M. Tartaglia, Getting a movement to move: the plain language movement, ICADE. Revista de la Facultad de Derecho 94 (2015) 177–208. URL: https://revistas.comillas.edu/index.php/revistaicade/article/view/5433. doi:10.14422/icade.i94.y2015.008.

[8] M. M. Asprey, Plain language for lawyers, The Federation Press, 1996.

[9] E. Commission, D.-G. for Translation, Z. Field, How to write clearly, Publications Office of the European Union, 2015. doi:doi/10.2782/022405.
[10] U. S. Securities, E. C. O. of Investor Education, Assistance, A Plain English Handbook: How to Create Clear SEC Disclosure Documents, The Office, 1998.
[11] C. J. Hurtado, A. M. Reguera, Metodología de la traducción a lectura fácil: Retos de investigación, in: Translation, Mediation and Accessibility for Linguistic Minorities, volume 128, 2022, p. 205.
[12] O. García Muñoz, Lectura fácil: métodos de redacción y evaluación, Real Patronato sobre Discapacidad, 2012.
[13] J. Fernández Huerta, Medidas sencillas de lecturabilidad, Consigna 214 (1959) 29–32.
[14] J. M. Porras-Garzón, R. Estopà, Escalas de legibilidad aplicadas a informes médicos: límites de un análisis cuantitativo formal, Círculo de Lingüística Aplicada a la Comunicación 83 (2020) 205–216. doi:10.5209/clac.70574.
[15] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, arXiv preprint arXiv:2402.07927 (2024).