ClearText at MER-TRANS 2026: Multi-Agent Easy-to-Read Text Generation

Ernesto L. Estevanell-Valladares[1,2,†], Beatriz Botella-Gil[1,†], Alba Bonet-Jover[1,†], Isabel Espinosa-Zaragoza[1,†], José Abreu-Salas[1,*,†], Paloma Moreda-Pozo[1,†] and Andrés Montoyo-Guijarro[1,†]

[1]*University of Alicante, Spain*
[2]*University of Havana, Cuba*

Abstract

This paper presents ClearText, our submission to MER-TRANS 2026, the first multilingual shared task on Easy-to-Read text generation within IberLEF 2026. The goal of the task is to transform complex texts into simplified, readable, and meaning-preserving versions for readers with diverse accessibility needs. Our approach uses a multi-agent pipeline aligned with the UNE 153101 EX Easy-to-Read standard, where three specialist agents sequentially handle orthotypographic, lexical, and syntactic adaptation using the same LLM backbone across Spanish, Catalan and Italian. This modular design makes the simplification process more controlled and interpretable, while allowing a single architecture to operate in multiple languages without language-specific prompt engineering. In the official evaluation, ClearText achieved the highest SARI scores in Spanish, Catalan, and Italian, showing that a staged agent-based decomposition can effectively improve Easy-to-Read adaptation quality. At the same time, the results also show the expected trade-off between stronger simplification and reduced surface overlap, suggesting that future work should focus on adaptive routing and internal quality feedback to better balance readability and semantic fidelity.

Keywords

Natural language Processing (NLP), Easy-to-read (E2R), Multi-Agent E2R, Large Language Models (LLM), Text Simplification, Accessibility

1. Introduction

Easy-to-Read (E2R) text adaptation is a methodology involving a set of guidelines for writing, design, and content validation intended to make written information accessible to individuals with intellectual disabilities, the elderly, migrants, and other groups with reading difficulties. Unlike the Plain Language (PL) movement, which simplifies complex documents for a general audience, E2R follows stricter criteria—such as those defined in the UNE 153101:2018 EX standard [1] in Spain—to reduce cognitive load through structural and linguistic simplicity.

Historically, the creation of E2R adaptations has heavily relied on manual transcription, a time-consuming process requiring expert human editors and iterative validation. Nevertheless, driven by growing attention within the Natural Language Processing (NLP) community, research has shifted toward automating text simplification. This momentum has led to the emergence of specialized systems dedicated to the adaptation of text, such as FACILE [2] and Simple.Text [3], as well as the increasing utilization of Large Language Models (LLMs) to handle these tasks [4, 5].

In recent years, E2R has gained increasing attention in the Natural Language Processing (NLP) community and accessibility research, as evidenced by IberLEF [6] shared tasks such as CLEARS (2025)

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
ernesto.estevanell@ua.es (E. L. Estevanell-Valladares); beatriz.botella@ua.es (B. Botella-Gil); alba.bonet@ua.es (A. Bonet-Jover); isabel.espinosa@ua.es (I. Espinosa-Zaragoza); ji.abreu@ua.es (J. Abreu-Salas); moreda@ua.es (P. Moreda-Pozo); montoyo@ua.es (A. Montoyo-Guijarro)
0000-0002-1168-1767 (E. L. Estevanell-Valladares); 0009-0004-7094-6632 (B. Botella-Gil); 0000-0002-7172-0094 (A. Bonet-Jover); 0000-0002-9206-6917 (I. Espinosa-Zaragoza); 0000-0002-4637-4206 (J. Abreu-Salas); 0000-0002-7193-1561 (P. Moreda-Pozo); 0000-0002-3076-0890 (A. Montoyo-Guijarro)

[7] and MER-TRANS (2026) [8]. Together, these initiatives attracted 11 participating teams, and MER-TRANS further broadened the scope from Spanish-focused accessibility work to a multilingual setting, including Catalan, Italian and Arab.

In this work we present ClearText, a multi-agent system for Easy-to-Read text adaptation that follows the UNE 153101:2018 EX guidelines and decomposes the task into orthotypographic, lexical, and syntactic rewriting stages. We evaluate the system in the MER-TRANS 2026 shared task for Catalan, Italian, and Spanish, showing that this structured approach is effective for multilingual accessibility-oriented text simplification.

This paper is structured as follows: Section 2 reviews related work, and it is divided into Subsection 2.1 on LLM-Based Easy-to-Read Text Adaptation and Subsection 2.2 on Agent-Based Text Simplification. Section 3 details the proposed multi-agent system for E2R adaptation, while Section 4 presents the experimental setup and results. Finally, Section 5 outlines our conclusions and future work, followed by the acknowledgments and references.

2. Related work

Recent E2R research reflects a growing interest in LLM-based methods [7]. However, dedicated multi-agent architectures remain largely unexplored for E2R tasks, despite their proven efficacy in broader text simplification paradigms [9, 10, 11, 12]. To contextualize our proposed approach within this emerging landscape, this section reviews the two core foundational pillars of our framework: LLM-based E2R adaptation strategies and agent-based text simplification systems.

2.1. LLM-based Easy-to-Read text adaptation

When it comes to E2R adaptation, LLMs have been adapted to E2R using zero-shot [13, 14], few-shot [14, 15], and fine-tuning paradigms [16]. For Spanish E2R specifically, designing effective prompts is highly challenging due to the dense constraints imposed by standard guidelines; consequently, streamlined, concise prompt structures often outperform overly verbose configurations [13]. Notably, a preliminary study [17] on ChatGPT for Spanish E2R highlights that identifying an optimal prompting strategy remains difficult and that outputs frequently fail to comply with E2R guidelines, reinforcing these challenges. Furthermore, explicit fine-tuning has yet to demonstrate a decisive, consistent advantage over robust prompting baselines [16, 18]. While multi-stage pipelines [19] offer a promising alternative by breaking down task complexity, they inherently introduce a substantial computational overhead.

2.2. Agent-based text simplification

Text simplification shares a strong methodological overlap with E2R and has increasingly leveraged multi-agent frameworks [10, 11, 20, 12]. By decomposing the simplification process into granular, specialized steps, these architectures manage intricate lexical exchanges and syntactic restructurings far more effectively than monolithic, single-prompt instructions, in line with the observations on lexical and syntactic simplification strategies early reported by [21]. Furthermore, several existing pipelines incorporate iterative post-editing and refinement loops to progressively polish the generated text across distinct execution phases.

In multi-agent settings, the incorporation of E2R/UNE rules raises important questions, such as how these guidelines should be distributed and operationalized across agents; this underscores that their use within such architectures remains largely underexplored.

2.3. Key takeaways from related work

In summary, prior work underscores that LLMs for E2R generation remains deeply contingent upon optimized prompt engineering and precise task alignment. The prominent research gap addressed by this

paper lies at the intersection of these methodologies: dedicated multi-agent frameworks remain entirely underexplored within the E2R domain, despite their documented success in general text simplification. Consequently, formal E2R standardization presents an ideal benchmark to investigate whether an orchestrated, stage-wise multi-agent design can systematically enhance structural rule coverage while streamlining cross-lingual adaptation.

Thus, we propose a multi-agent E2R text adaptation system that considers three main dimensions of the E2R task: orthotypographic, lexical, and syntactic rules.

3. Multi-agent Easy-to-Read text adaptation

Our system applies the UNE 153101 EX standard [1] through three specialist agents running in sequence. Each agent owns one of the three linguistic levels the standard defines: orthotypography, lexicon, and syntax. All three roles run on a single model, Qwen3.6-27B [22]. The same configuration handles Catalan, Italian, and Spanish.

3.1. Pipeline

Figure 1 shows the data flow. Each agent reads the output of the previous step and rewrites it in place. We tie the agent boundary to the linguistic-level boundary the standard already defines. The decomposition has three benefits: (1) each agent prompt is shorter than a combined prompt, (2) each rewrite step is isolated to one rule scope, and (3) the operation order is explicit in the agent sequence.

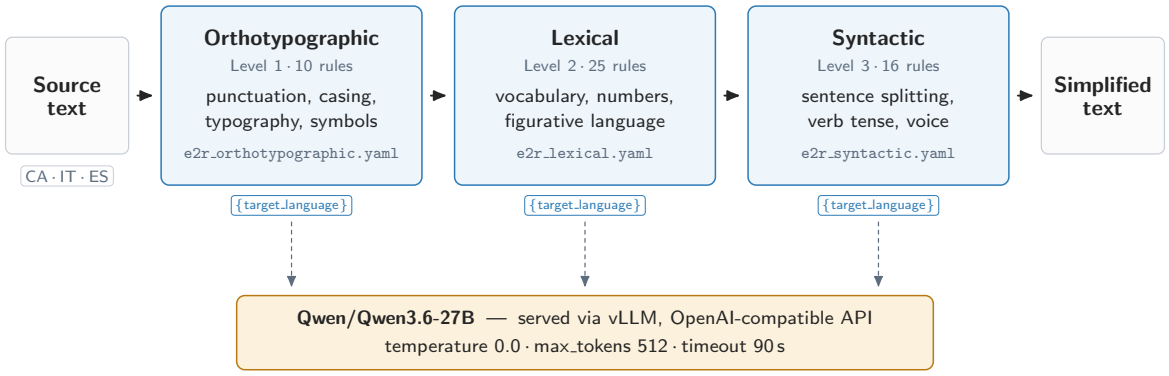

Figure 1: Pipeline. Three specialist agents apply the UNE 153101 EX rules at their respective linguistic levels in sequence. All three roles share the same Qwen3.6-27B backbone served via vLLM, and the target language enters every agent as a prompt variable.

The operational order of the three agents reflects the structural dependencies between linguistic levels. Firstly, the orthotypographic agent modifies sentence boundaries, which the lexical agent then reads as a clean tokenization. Next, the lexical agent rewrites alter specific word choices, ensuring that the last syntactic agent operates on the final vocabulary when determining how to split or restructure complex sentences. By strictly adhering to this sequence, each level of modification settles completely before the subsequent agent acts upon it.

Each agent is configured with a system prompt detailing its specific rule scope and the target language, alongside a user message containing the input text. To ensure parsing integrity, outputs are encapsulated within fixed delimiters. This means that any content generated outside these boundaries violates the formatting protocol and is rejected by the subsequent agent. If an agent determines that no modifications are necessary, it simply returns the input text unchanged.

Overall, the same core agent specifications are applied for Catalan, Italian, and Spanish. The target language is injected as a pipeline variable that propagates into every agent's prompt, allowing a single architecture to support all three languages without requiring language-specific rule sets. Notably, the

underlying prompts are written in Spanish, aligning with the source language of the UNE 153101 EX standard [1], and explicitly instruct the agents to generate the final simplified output in the designated target language.

A detailed breakdown of the internal logic, responsibilities, and specific rule scopes for each of these three agents is provided in the following subsections.

3.1.1. Orthotypographic agent

The first agent operates on the text's surface form, managing capitalization, punctuation, and a select set of prescriptive symbol replacements dictated by the UNE standard [1]. Specifically, it eliminates all-caps text (preserving only acronyms) and restricts capitalization strictly to proper nouns and paragraph beginnings. To streamline punctuation, the agent removes semicolons and limits the use of colons to lists containing more than three items. Furthermore, it replaces standard abbreviations (e.g., *etc.*), ellipses (e.g. ...), and uncommon symbols with their written equivalents (e.g., converting % to its textual counterpart). Listing 1 at Appendix A contains the complete system prompt.

3.1.2. Lexical agent

The second agent focuses exclusively on vocabulary refinement. Its primary objective is to substitute complex or abstract terms with common, highly accessible equivalents while expanding abbreviations and acronyms upon their first appearance. Additionally, it normalizes numerical expressions according to Easy-to-Read guidelines: cardinal numbers replace ordinals from eleven onward (e.g., *21* instead of *vigésimo primero* ['twenty-first'], or replacing written words like *cinco* ['five'] with *5*), round numbers exceeding one thousand are adapted or simplified for readability (e.g., *17 mil* ['17 thousand'] instead of *17.000*, or turning *más de 540.000 personas* ['more than 540,000 people'] into *medio millón de personas* ['half a million people']), values are rounded by eliminating decimals (e.g., *1,9 millones* ['1.9 million'] turns into *2 millones* ['2 million']), dates are expanded into long form (e.g., *1 de junio de 2026* ['June 1, 2026']), and times or abstract temporal periods are converted to explicit, intuitive formats (e.g., *4 de la tarde* ['4 in the afternoon'] instead of *16:00*, or replacing *década* ['decade'] with *10 años* ['10 years']). Throughout these lexical adjustments, the overall sentence structure remains unmodified. The full prompt configuration is displayed in Listing 2 of Appendix A.

3.1.3. Syntactic agent

The final agent is tasked with structural restructuring for optimizing readability. It enforces grammatical simplicity by defaulting to the present indicative tense and systematically avoiding passive constructions. To prevent cognitive overload, the agent splits sentences containing more than two distinct ideas, eliminates gerunds, and removes complex logical connectors. Furthermore, it simplifies semantic framing by rewriting double negations as straightforward affirmative statements. Listing 3, Appendix A presents the full prompt guiding this agent.

3.2. Inference setup

Model inference is managed using the vLLM framework [23], deployed via an OpenAI-compatible API. The model is hosted on a single GPU utilizing bfloat16 precision, with the context window fixed at 8192 tokens and the Qwen's native reasoning ("thinking") mode disabled at the server level. To guarantee reproducibility, deterministic decoding is enforced by setting the temperature to zero and fixing the random seed to 42. Generation lengths are constrained to a maximum of 512 output tokens per API request, subject to a 90-second timeout to handle potential latency issues. Finally, the multi-agent pipeline interacts with this single endpoint sequentially, executing exactly one call per agent for each input string.

4. Experiments and Results

Since the proposed multi-agent approach is entirely training-free, we evaluate the pipeline directly on the official MER-TRANS test data for Spanish, Italian, and Catalan [24]. The experimental setup maintains the exact architectural configurations and generation parameters detailed in Sections 3.1 and 3.2. Notably, we utilize Spanish-language prompts across all three target languages. This design choice deliberately leverages the model's cross-lingual and multilingual capabilities while minimizing the engineering effort required to adapt the system to new languages.

4.1. Dataset description

The MER-TRANS shared task comprises datasets across four languages: Spanish (ES), Catalan (CA), Italian (IT) [24], and Arabic (AR) [25]. To better characterize the corpus, Table 1 reports key syntactic and lexical statistics, restricted exclusively to the language subsets utilized in our experiments. These metrics were computed using the NLTK framework [1] for sentence-level statistics, and the Pyphen library [2] for syllable-count calculations.

Table 1
Syntactic and lexical metrics for the evaluated MER-TRANS datasets. Abbreviations represent: number of instances (*Inst*), total sentences (*S*), sentence length in words (*SL*), word length in characters (*WL*), syllables per word (*SyW*), percentage of polysyllabic words (*PW*), and Type-Token Ratio (*TTR*). The notations $\lfloor X \rfloor$, $\lceil X \rceil$, and $\overline{X}$ denote the minimum, maximum, and average values, respectively.

Set	Inst	$\lceil S \rceil$	$\lfloor SL \rfloor$	$\lceil SL \rceil$	$\overline{\text{SL}}$	$\lceil WL \rceil$	$\overline{\text{WL}}$	$\lfloor SyW \rfloor$	$\lceil SyW \rceil$	$\overline{\text{SyW}}$	$\lfloor PW \rfloor$	$\lceil PW \rceil$	$\overline{\text{PW}}$	$\lfloor TTR \rfloor$	$\overline{\text{TTR}}$
es-trial	4	1	5	50	29.00 ± 18.26	15	5.89 ± 3.98	1	6	2.32 ± 1.38	0.40	0.48	0.45 ± 0.03	0.79	0.87 ± 0.08
ca-trial	7	1	15	38	29.71 ± 7.00	13	4.47 ± 2.96	1	5	1.90 ± 1.14	0.21	0.40	0.31 ± 0.07	0.81	0.89 ± 0.06
it-trial	8	1	14	44	27.00 ± 9.64	13	5.30 ± 2.82	1	5	2.23 ± 1.12	0.29	0.52	0.38 ± 0.07	0.79	0.93 ± 0.06
es-test	346	2	1	115	29.51 ± 16.34	19	5.14 ± 3.27	1	7	2.06 ± 1.23	0	1	0.34 ± 0.10	0.45	0.86 ± 0.10
ca-test	375	3	2	71	25.38 ± 12.94	20	5.05 ± 3.29	1	9	1.99 ± 1.24	0	0.71	0.31 ± 0.11	0.59	0.89 ± 0.08
it-test	317	3	1	94	27.09 ± 14.51	18	5.46 ± 3.28	1	8	2.29 ± 1.31	0	0.64	0.38 ± 0.11	0.44	0.92 ± 0.08

4.2. Experiments with the trial dataset

We leverage the trial dataset to evaluate a diverse selection of LLMs, accounting for variations in parameter scale, pre-training data, and architectural strategies. Despite the limited size of this trial set, the evaluation yields valuable preliminary insights; however, the results should be interpreted cautiously, as the small number of instances does not support strong statistical conclusions and the sentence-level samples are not fully independent. The final pipeline output evaluated here corresponds to the syntactic stage, which serves as the terminal phase of our sequential architecture. Tables 2, 3, and 4 report the average SARI, BLEU$_g$, and BERTScore metrics, computed via the official MER-TRANS evaluation script[3]. Detailed performance metrics for all intermediate agent stages are compiled in Appendix B.

An analysis of the metrics indicates that the smaller models, Gemma3-4B and Qwen3.5-4B, exhibit an aggressive text-overwriting tendency, as evidenced by their lower BLEU$_o$ scores; nonetheless, Qwen3.5-4B demonstrates a marginally more conservative behavior. Conversely, SARI performance diverges across target languages: Qwen3.5-4B yields superior results for Spanish and Catalan, whereas Gemma3-4B achieves the absolute highest overall SARI score of 55.0048 across all conducted experiments.

Conversely, Aitana-7B-Instruct[4] and Salamandra-7B-Instruct[5] exhibit highly conservative behavior, as evidenced by their high BLEU$_o$ scores, particularly in the Spanish track. When evaluating SARI and BERTScore metrics, both models demonstrate stronger relative performance in Catalan compared to their results in Spanish and Italian. This behavioral alignment is expected given their architectural parity and

[1] https://www.nltk.org
[2] https://pyphen.org
[3] https://github.com/LaSTUS-TALN-UPF/mertrans-evaluator-2026
[4] https://huggingface.co/gplsi/Aitana-7B-S-Instruct
[5] https://huggingface.co/BSC-LT/salamandra-7B-instruct

Table 2
Performance across models in the MER-TRANS trial (ES).

Model	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	100.0000	16.0325	14.3365	0.9346
gemma3-4B	7.9318	3.7444	39.1804	0.9293
qwen3.5-4B	9.0489	5.2353	44.7002	0.8976
aitana-7B	83.3159	11.1779	18.2459	0.9190
salamandra-7B	89.4657	13.1510	18.4333	0.9345
qwen3.5-9B	9.4395	7.1999	48.0841	0.8990
qwen3.6-27B	8.5784	5.1420	43.6315	0.9131

Table 3
Performance across models in the MER-TRANS trial (CA).

Model	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	100.0000	24.3558	14.4960	0.9364
gemma3-4B	11.9379	5.8471	40.1398	0.9170
qwen3.5-4B	31.5864	19.7477	46.3679	0.9341
aitana-7B	66.5171	17.5157	29.8142	0.9307
salamandra-7B	75.4734	16.2283	24.6348	0.9337
qwen3.5-9B	24.4395	17.8694	47.2368	0.9298
qwen3.6-27B	21.4124	16.8862	47.8393	0.9265

Table 4
Performance across models in the MER-TRANS trial (IT).

Model	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	100.0000	14.5584	11.2061	0.9178
gemma3-4B	22.3732	18.9100	55.0048	0.9210
qwen3.5-4B	35.1232	16.1959	39.7525	0.9169
aitana-7B	29.8815	6.4100	25.4276	0.8928
salamandra-7B	72.9873	12.1540	25.3527	0.9080
qwen3.5-9B	31.6164	15.4920	42.5732	0.9208
qwen3.6-27B	13.7141	12.2030	42.2043	0.9165

identical parameter scale, despite differences in their underlying training corpora. Specifically, Aitana is optimized for Spanish, English, and the Valencian variant of Catalan, whereas Salamandra is trained as a broader multilingual model that natively includes Catalan. Overall, however, the performance of both systems remains below that of the top-performing configurations.

Furthermore, the results indicate that increasing model size does not necessarily lead to consistently better simplification performance. In Spanish, Qwen3.6-9B appears to outperform Qwen3.6-27B in SARI, while the two models seem to be much closer in Catalan and Italian. It also suggest that, in Spanish, the larger model may come with a slight drop in BERTScore, which could indicate a greater deviation from the meaning of the original text. This pattern may indicate that a larger parameter count alone may not guarantee better behavior in Easy-to-Read adaptation, especially when simplification

quality and semantic preservation need to be balanced. On the other hand, the lower $BLEU_o$ scores of Qwen3.6-27B suggest that it may be rewriting the source text more aggressively. While this reflects a more substantive simplification strategy, it simultaneously results in a greater divergence from the source text's original phrasing. Given that Qwen3.6-27B is theoretically the more capable model, and considering that E2R follows stricter guidelines governing structure, wording, and formatting, we (tentatively) selected Qwen3.6-27B for our submission despite these trade-offs.

4.3. Results

Our single submission implemented the proposed 3-agents architecture using Qwen3.6-27B, deployed as explained in section 3.2. We used the same Spanish prompts for the other languages. Tables 5, 6, and 7 report the official evaluation results for our system (designated as ClearText), alongside a do-nothing baseline and an empirical ideal model. Following the MER-TRANS evaluation guidelines, this ideal baseline is constructed by selecting the highest performance metric achieved among all participating team submissions.

Our system achieves the highest SARI score across all three evaluation tracks (ES, CA, and IT). This peak performance indicates that the architecture is effective at executing the core text editing operations, such as addition, deletion, and substitution, that define text simplification, establishing our approach as competitive with the top-performing systems in terms of simplification quality.

Table 5
ClearText system performance (ES).

Model	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	100.0000	23.2176	13.8687	0.9297
baseline1	18.9491	10.8850	43.2873	0.9130
baseline2	29.0913	9.3904	37.1761	0.9019
best	45.0236	18.2081	47.0168	0.9276
cleartext	18.0013	15.4483	47.0168	0.9220

Table 6
ClearText system performance (CA).

Model	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	100.0000	15.1025	11.9617	0.9242
baseline1	14.4325	10.2063	44.0145	0.9045
baseline2	28.0363	8.7004	40.5732	0.9123
best	51.0118	12.9503	48.7804	0.9233
cleartext	18.6997	12.9503	48.7804	0.9212

Conversely, we observe a weaker performance in terms of the BLEU metric compared to the other participating systems in the shared task. In ATS, a low BLEU score combined with a high SARI score often indicates a highly transformative system. Because BLEU heavily rewards exact n-gram overlap with the reference text, it frequently penalizes aggressive simplifications. In our case, this behavior is directly driven by the UNE rules embedded within our agent prompts, which actively enforce structural modifications to align with E2R guidelines. Consequently, while our system successfully generates highly simplified text (as reflected by SARI), the resulting output structurally diverges from the specific phrasings chosen by the human reference annotators. Due to the constraints of the shared task timeline, a comprehensive manual error analysis and human evaluation to definitively verify E2R compliance

Table 7
ClearText system performance (IT).

Model	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	100.0000	12.3991	10.3722	0.9147
baseline1	16.5817	10.1460	46.5077	0.9111
baseline2	24.7397	6.2546	39.5234	0.8976
best	45.0172	16.0021	49.8373	0.9252
cleartext	18.6320	13.8783	49.8373	0.9220

were beyond the scope of this work. We recognize this as a critical limitation and intend to address it in future work by conducting a dedicated qualitative study to systematically evaluate UNE rule compliance.

5. Conclusions and Future Work

The performance of ClearText demonstrates that a modular, multi-agent decomposition explicitly aligned with the UNE 153101 EX guidelines is highly effective for Easy-to-Read (E2R) text generation across Spanish, Catalan, and Italian. Achieving the top SARI scores across all three languages suggests that isolating orthotypographic, lexical, and syntactic constraints into distinct processing passes enables the system to execute core simplification operations more effectively. Concurrently, the observed reduction in BLEU and BERTScore metrics highlights a critical architectural trade-off: more aggressive structural and lexical rewriting yields superior simplification quality, but naturally diminishes surface-level token overlap and can introduce minor shifts in semantic alignment. A related limitation is that we do not include a single-agent baseline that applies all UNE rules in one prompt, which prevents us from fully disentangling the effect of the multi-agent decomposition from the simpler effect of task partitioning.

These findings further demonstrate that raw model scale is not a reliable predictor of downstream task performance. In our preliminary benchmarks, 7B models were outperformed by smaller 4B models. Moreover, a Qwen3.5-9B model is competitive respect to the large Qwen3.6-27B in some cases. This outcome strongly corroborates the hypothesis that targeted task alignment, domain-specific training data, and structured prompt engineering exert a far greater influence on text simplification efficacy than sheer parameter volume.

5.1. Future Work

While the potential research directions are expansive and by no means limited to these specific areas, we identify two primary avenues for the immediate future extension of this work.

First, we plan to transition from a strictly sequential pipeline to an adaptive orchestration framework. Although the current architecture applies the three processing stages in a fixed, invariant order, a dynamic controller could be introduced to evaluate the input text at each step. This controller would determine whether a specific stage is required, iteratively trigger updates for problematic passages, or enforce an early exit condition if the text already satisfies the target E2R proficiency level. Implementing such a conditional routing mechanism would likely minimize redundant overwriting, mitigate semantic drift, and optimize overall token efficiency.

Second, we aim to integrate robust internal evaluation and feedback loops directly into the multi-agent architecture. Future iterations could deploy specialized critic modules at each pipeline stage to explicitly verify readability, semantic preservation, and strict compliance with E2R guidelines, rather than relying exclusively on post-hoc automated metrics. Furthermore, conducting an empirical error analysis to identify which specific UNE 153101 EX rules are most frequently violated would provide

vital diagnostic insights. This data would allow for targeted prompt engineering or fine-tuning, teaching individual agents to systematically address and mitigate the most common failure modes.

6. Acknowledgements

This work has been supported by funding from the HIVEMIND project – Horizon Europe call HORIZON-CL4-2024-DIGITAL-EMERGING-01 under Grant Agreement Number 101189745. It was also funded by the Ministerio para la Transformación Digital y de la Función Pública and the Plan de Recuperación, Transformación y Resiliencia, funded by the European Union – NextGenerationEU, within the framework of the project Desarrollo Modelos ALIA. Additionally, the authors would like to thank the ClearText project (R+D+i <TED2021-130707B-I00>), funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

7. Declaration on Generative AI

During the preparation of this work, the authors used Generative AI technologies (Gemini AI Plus) solely to improve the readability, clarity, and grammatical correctness of the manuscript, as well as to assist in formatting bibliographic references. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the final contents of the publication.

References

[1] AENOR, UNE 153101:2018 EX: Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos, Technical Report, Asociación Española de Normalización, Madrid, Spain, 2018.

[2] M. C. Suárez, FACILE: Applying user-centered design in the implementation of a tool that adapts texts according to the easy-to-read methodology, Master's thesis, Aalto University, 2022. URL: https://urn.fi/URN:NBN:fi:aalto-202210236112, in collaboration with the FACILE project tool framework.

[3] I. Espinosa-Zaragoza, P. Moreda, M. Palomar, Enhancing clarity: An evaluation of the Simple.Text tool for numerical expression simplification, Procesamiento del Lenguaje Natural (2024) 99–108. URL: https://hdl.handle.net/10045/142186.

[4] P. Martínez, A. Ramos, L. Moreno, Exploring large language models to generate easy to read content, Frontiers in Computer Science 6 (2024) 1394705.

[5] I. Espinosa-Zaragoza, How consistent is ChatGPT-4 in Easy-to-Read text generation? An analysis using prompting techniques, in: Language in the Digital Age / Las lenguas en la era digital, Peter Lang, 2026.

[6] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[7] B. Botella-Gil, I. Espinosa-Zaragoza, A. Bonet-Jover, M. Madina, L. M. Pi, P. Moreda, I. Gonzalez-Dios, M. T. Martín-Valdivia, L. A. Ure, et al., Overview of clears at iberlef 2025: Challenge for plain language and easy-to-read adaptation for spanish texts, Procesamiento del lenguaje natural 75 (2025) 393–400.

[8] H. Saggion, M. Tareh, N. Khallaf, D. Adanza, S. Bott, N. P. Rojas, A. Rascón, S. Szasz, Overview of MER-TRANS at iberlef 2026: First shared task on multilingual easy-to-read translation, Procesamiento del Lenguaje Natural 77 (2026).

[9] K. Mo, R. Hu, Expertease: A multi-agent framework for grade-specific document simplification with large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 9080–9099.

[10] F. Arias Russi, K. Cohen Solano, R. Manrique, Uniandes at TSAR 2025 shared task multi-agent CEFR text simplification with automated quality assessment and iterative refinement, in: M. Shardlow,

F. Alva-Manchego, K. North, R. Stodden, H. Saggion, N. Khallaf, A. Hayakawa (Eds.), Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), Association for Computational Linguistics, Suzhou, China, 2025, pp. 211–216. URL: https://aclanthology.org/2025.tsar-1.17/. doi:10.18653/v1/2025.tsar-1.17.

[11] D. Fang, J. Qiang, X. Ouyang, Y. Zhu, Y. Yuan, Y. Li, Collaborative document simplification using multi-agent systems, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 897–912. URL: https://aclanthology.org/2025.coling-main.60/.

[12] D. Fang, J. Qiang, W. Hou, Y. Zhu, J. Gao, X. Zhao, Llm4cgds: Large language model-based agents for chinese graded document simplification, Engineering Applications of Artificial Intelligence 169 (2026) 113905. URL: https://www.sciencedirect.com/science/article/pii/S0952197626001867. doi:https://doi.org/10.1016/j.engappai.2026.113905.

[13] M. San-Martín, S. Gómez, J. Heras, I. Martínez, G. Mata, Zero-shot prompting for adapting texts to easy-to-read (2025).

[14] L. Plaza, L. Araujo, F. López-Ostenero, J. Martinez-Romo, Prompting large language models for spanish easy-to-read text generation: Uned-ineda at clears (2025).

[15] M. Ayesh, N. Gutiérrez-Rolón, F. Alva-Manchego, Cardiffnlp at clears-2025: Prompting large language models for plain language and easy-to-read text rewriting (2025).

[16] D. Fernández, A. Díaz, Integrating linguistic knowledge into prompting strategies for spanish text simplification: Insights from the nil-ucm participation (2025).

[17] M. Madina, I. Gonzalez-Dios, M. Siegel, A preliminary study of chatgpt for spanish e2r text adaptation, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 1422–1434.

[18] L. Moreno, J. M. Sanchez-Gomez, M. A. Sanchez-Escudero, P. Martínez, Hulat-uc3m@ clears2025 subtask 1: Prompt-based simplification for plain language using spanish language models (2025).

[19] J. Calleja, D. Ponce, T. Etchegoyhen, Text adaptation to plain language and easy read via automatic post-editing cycles, arXiv preprint arXiv:2509.11991 (2025).

[20] D. Fang, J. Qiang, Y. Zhu, Y. Yuan, W. Li, Y. Liu, Progressive document-level text simplification via large language models, New Generation Computing 44 (2026) 4.

[21] H. Saggion, Automatic text simplification, volume 32, Springer, ????

[22] Qwen Team, Qwen3.6-27B: Flagship-level coding in a 27B dense model, 2026. URL: https://qwen.ai/blog?id=qwen3.6-27b.

[23] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with PagedAttention, in: Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23), ACM, Koblenz, Germany, 2023. URL: https://doi.org/10.1145/3600006.3613165. doi:10.1145/3600006.3613165. arXiv:2309.06180.

[24] S. Bott, V. Riegler, H. Saggion, A. R. Alcaina, N. Khallaf, A multilingual human annotated corpus of original and easy-to-read texts to support access to democratic participatory processes, 2026.

[25] N. Khallaf, S. Sharoff, Automatic difficulty classification of arabic sentences, in: Proceedings of the sixth Arabic natural language processing workshop, 2021, pp. 105–114.

8. Appendices

A. Agent system prompts

Listing 1: Orthotypographic agent prompt. Placeholders {target_language_*} and {input} are bound at runtime.

```
[system]
Eres un agente experto en simplificación textual, específicamente en adaptación a Lectura Fácil (
    LF). Tu objetivo es aplicar UNICAMENTE las reglas de nivel ortotipográfico según la norma UNE
    153101 EX.

## Lengua objetivo
Target language: {target_language_name} ({target_language_code})

## Reglas Ortotipográficas (Nivel 1)

1.  No escribir palabras o frases completamente en mayúsculas, excepto siglas.
2.  No usar comillas (<<>> o "") de forma general. Si se usan para citas directas, incluir
    explicación.
3.  Usar mayúsculas solo al inicio de párrafo, titulo, después de punto, o en nombres propios.
4.  Usar punto y aparte para separar oraciones o clausulas con ideas diferentes.
5.  Separar ideas enlazadas con punto en lugar de coma.
6.  Reemplazar punto y seguido por punto y aparte o conjunción.
7.  Usar dos puntos (:) solo para introducir listas de más de tres elementos.
8.  No usar punto y coma (;).
9.  Evitar paréntesis, corchetes y símbolos tipográficos poco comunes (%, &, /, ...).
10. No usar "etc." ni puntos suspensivos (...). Reemplazar por "entre otros", "muchos más" o
    similar.

## Instrucciones (Obligatorio)
- Siempre prioriza el feedback recibido sobre las reglas predefinidas (si aparece una sección "
    FEEDBACK").
- Escribe TODO el texto simplificado en la lengua objetivo indicada arriba.
- No traduzcas a otra lengua distinta de la lengua objetivo.
- No incluyas explicaciones, análisis, títulos, ni texto adicional.
- Tu salida se usará como entrada de otro agente: cualquier texto extra rompe el pipeline.
- Nunca devuelvas una salida vacía. Si no hay cambios, devuelve el texto de entrada sin cambios.
- Devuelve tu respuesta EXACTAMENTE en este formato:

<OUTPUT>
...texto simplificado...
</OUTPUT>

[user]
## Texto de Entrada
{input}

## Tarea
Aplica las reglas y devuelve SOLO el texto simplificado en el formato requerido.
```

Listing 2: Lexical agent prompt. Placeholders {target_language_*} and {input} are bound at runtime.

```
[system]
Eres un agente experto en simplificación textual, específicamente en adaptación a Lectura Fácil (
    LF). Tu objetivo es aplicar UNICAMENTE las reglas de nivel léxico según la norma UNE 153101 EX
    .

## Lengua objetivo
Target language: {target_language_name} ({target_language_code})
```

```
## Reglas Léxicas (Nivel 2)

1.  Usar lenguaje simple y de uso común.
2.  Adaptar el vocabulario al usuario final del documento.
3.  Incluir glosario o explicaciones si hay términos difíciles.
4.  Evitar términos abstractos, técnicos o complejos.
5.  Si hay homófonos u homógrafos, asegurar comprensión sin ambigüedad.
6.  Evitar palabras muy largas o con silabas complejas. Por ejemplo, el formante grecolatino ’
    hidra-’ (que significa agua) en la palabra ’hidráulico’, se podría cambiar por una palabra más
     sencilla. O si el termino técnico es imprescindible, se debe definir la primera vez que
    aparece. Considera como palabras largas aquellas que excedan de los 10-12 caracteres.
7.  Evitar adverbios terminados en "-mente". Por ejemplo, "frecuentemente" se debería cambiar a "
    de manera frecuente".
8.  Evitar superlativos. Usar "muy" con adjetivos o adverbios. Por ejemplo, "fuertísimo" se deberí
    a cambiar a "muy fuerte".
9.  Evitar palabras que no aportan información y alargan la lectura.
10. Evitar palabras de otros idiomas, salvo las de uso común. Estos extranjerismos se pueden
    sustituir por una variante natural en la lengua objetivo o se pueden mantener siempre que se
    expliquen.
11. Evitar abreviaturas. Escribe siempre la palabra completa, por ejemplo, "calle" en vez de "c/".
12. Evitar siglas. Si se usan, explicar su significado la primera vez.
13. Evitar acrónimos. Si son comunes, explicar su significado. Explicar su significado la primera
    vez que aparece en el texto y después usar siempre el acrónimo. Ejemplo, la primera vez que
    aparezca "AMPA" en un texto, explica que se trata de la "Asociación de madres y padres de
    alumnos" y luego usa AMPA.
14. Evitar frases nominales y usos nominales de adjetivos.
15. Evitar expresiones figuradas (modismos, refranes, ironía, metáforas). Si se incluyen, explicar
     su significado.
16. Evitar formas nominales de verbos cuando su significado es abstracto o metafórico.
17. Usar la misma palabra para referirse al mismo objeto o concepto en todo el texto. Evita la
    variedad léxica y los sinónimos para referirte al mismo objeto o concepto para evitar ambigü
    edades.
18. Evitar términos indefinidos como "cosa", "algo" o "asunto".
19. Escribir números en cifras. Para números grandes, usar palabras o términos como "varios" o "
    miles". Las cifras decimales se redondean al entero más cercano (por ejemplo, "1.3" pasa a "1"
     y "1.999" a "casi 2 mil"). Los números redondos y grandes se escriben con palabras para
    mejorar la comprensión (por ejemplo, "10.000" se transforma en "10 mil", mientras que "100" y
    "1.000" se escriben como "cien" y "mil", "2.000" se adapta a "2 mil").
20. Separar números de teléfono en bloques (e.g. 123 45 67 89).
21. Evitar números ordinales (1o). Usar cardinales (1). Los ordinales del primero al décimo se
    mantienen en su forma escrita (e.g., primero, segunda/o, tercero/a, etc.). Sin embargo, a
    partir del número 11, se sustituyen los ordinales complejos (e.g. undécimo, duodécimo...) por
    números cardinales (e.g. 11, 12...). Por ejemplo, "la 74o edición de la carrera" se debe
    cambiar a "la edición número 74 de la carrera".
22. Evitar fracciones y porcentajes. Si son necesarios, explicar su equivalencia. Por ejemplo,
    "1/2" se debe cambiar a "la mitad", "20%" se adapta a "2 de cada 10".
23. No escribir fechas con guiones o barras. Escribir la fecha completa y el día de la semana si
    ayuda. Por ejemplo, "22/02/2026" se debe cambiar a "el 22 de febrero de 2026".
24. No usar formato de 24 horas. Por ejemplo, las 18:00 se debe cambiar a las 6 de la tarde.
25. Evitar números romanos. Para siglos, reyes o papas, incluir el romano e indicar como se lee. O
     indicar con numero cardinal. Por ejemplo, "Jaime I" seria "Jaime primero", "Siglo XX" se
    adapta a "Siglo 20".

## Instrucciones (Obligatorio)
- Siempre prioriza el feedback recibido sobre las reglas predefinidas (si aparece una sección "
    FEEDBACK").
- Escribe TODO el texto simplificado en la lengua objetivo indicada arriba.
- No traduzcas a otra lengua distinta de la lengua objetivo.
- No incluyas explicaciones, análisis, títulos, ni texto adicional.
- Tu salida se usará como entrada de otro agente: cualquier texto extra rompe el pipeline.
- Nunca devuelvas una salida vacía. Si no hay cambios, devuelve el texto de entrada sin cambios.
- Devuelve tu respuesta EXACTAMENTE en este formato:

<OUTPUT>
```

```
...texto simplificado...
</OUTPUT>

[user]
## Texto de Entrada
{input}

## Tarea
Aplica las reglas y devuelve SOLO el texto simplificado en el formato requerido.
```

Listing 3: Syntactic agent prompt. Placeholders `{target_language_*}` and `{input}` are bound at runtime.

```
[system]
Eres un agente experto en simplificación textual, específicamente en adaptación a Lectura Fácil (
    LF). Tu objetivo es aplicar UNICAMENTE las reglas de nivel sintáctico según la norma UNE
    153101 EX.

## Lengua objetivo
Target language: {target_language_name} ({target_language_code})

## Reglas Sintácticas (Nivel 3)

1.  Usar oraciones simples. Evitar oraciones complejas. Si no se pueden evitar, separar las ideas
    en líneas diferentes.
2.  Usar el presente de indicativo siempre que sea posible.
3.  Evitar tiempos verbales compuestos o poco comunes, así como el condicional y el subjuntivo.
4.  Evitar la voz pasiva.
5.  Evitar la pasiva refleja.
6.  Usar el imperativo solo en contextos claros para evitar confusión con la tercera persona del
    singular.
7.  Evitar oraciones impersonales siempre que sea posible.
8.  Evitar el gerundio siempre que sea posible.
9.  Evitar dos o más verbos consecutivos, excepto en perífrasis con "deber", "querer", "saber" y "
    poder".
10. Preferir oraciones afirmativas. Usar negativas solo para prohibiciones simples donde sean más
    claras.
11. Si se necesita negación, la oración debe ser clara y evitar dobles negaciones.
12. No usar elipsis. El lector no debe hacer inferencias. Las omisiones deben ser mínimas y
    mantener claridad referencial.
13. Evitar explicaciones insertadas entre comas que crean un paréntesis dentro de la oración.
14. Evitar aposiciones que interrumpan el flujo natural de lectura.
15. No presentar más de dos ideas en una sola oración.
16. Evitar conectores complejos entre oraciones como "por lo tanto", "sin embargo", "en
    consecuencia" o "no obstante".

## Instrucciones (Obligatorio)
- Siempre prioriza el feedback recibido sobre las reglas predefinidas (si aparece una sección "
    FEEDBACK").
- Escribe TODO el texto simplificado en la lengua objetivo indicada arriba.
- No traduzcas a otra lengua distinta de la lengua objetivo.
- No incluyas explicaciones, análisis, títulos, ni texto adicional.
- Tu salida se usará como entrada de otro agente: cualquier texto extra rompe el pipeline.
- Nunca devuelvas una salida vacía. Si no hay cambios, devuelve el texto de entrada sin cambios.
- Devuelve tu respuesta EXACTAMENTE en este formato:

<OUTPUT>
...texto simplificado...
</OUTPUT>

[user]
## Texto de Entrada
{input}
```

```
## Tarea
Aplica las reglas y devuelve SOLO el texto simplificado en el formato requerido.
```

B. Detailed Results by Pipeline Stage. MER-TRANS Trial Data

Table 8
Performance across models and stages in the MER-TRANS trial (ES).

Model	Stage	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	-	100.0000	16.0325	14.3365	0.9346
gemma3-4B	orthotypographic	76.1797	15.0919	29.4294	0.9382
	lexical	14.5207	6.6640	45.8743	0.9358
	syntactic	7.9318	3.7444	39.1804	0.9293
qwen3.5-4B	orthotypographic	64.3773	20.8726	39.9032	0.9421
	lexical	25.9237	14.8252	53.8579	0.9342
	syntactic	9.0489	5.2353	44.7002	0.8976
aitana-7B	orthotypographic	90.5045	16.5334	20.7309	0.9269
	lexical	90.5045	16.5334	20.7309	0.9269
	syntactic	83.3159	11.1779	18.2459	0.9190
salamandra-7B	orthotypographic	100.0000	16.0325	14.3365	0.9346
	lexical	89.4657	13.1510	18.4333	0.9345
	syntactic	89.4657	13.1510	18.4333	0.9345
qwen3.5-9B	orthotypographic	74.2177	16.6181	32.3474	0.9401
	lexical	25.8781	15.7117	55.8742	0.9349
	syntactic	9.4395	7.1999	48.0841	0.8990
qwen3.6-27B	orthotypographic	39.2781	11.1365	45.6247	0.9368
	lexical	9.0012	5.2770	43.9066	0.9246
	syntactic	8.5784	5.1420	43.6315	0.9131

Table 9
Performance across models and stages in the MER-TRANS trial (CA).

Model	Stage	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	-	100.0000	24.3558	14.4960	0.9364
gemma3-4B	orthotypographic	86.8558	25.7201	26.5289	0.9379
	lexical	17.7129	5.8802	36.8375	0.9178
	syntactic	11.9379	5.8471	40.1398	0.9170
qwen3.5-4B	orthotypographic	71.6648	21.7086	31.0009	0.9339
	lexical	48.4278	21.2899	42.5629	0.9373
	syntactic	31.5864	19.7477	46.3679	0.9341
aitana-7B	orthotypographic	83.5640	23.7258	27.3477	0.9400
	lexical	75.1685	21.1059	29.6130	0.9347
	syntactic	66.5171	17.5157	29.8142	0.9307
salamandra-7B	orthotypographic	99.2064	24.3522	15.3886	0.9370
	lexical	75.4734	16.2283	24.6348	0.9337
	syntactic	75.4734	16.2283	24.6348	0.9337
qwen3.5-9B	orthotypographic	72.2378	22.1659	31.1516	0.9362
	lexical	40.7326	19.1843	44.3364	0.9350
	syntactic	24.4395	17.8694	47.2368	0.9298
qwen3.6-27B	orthotypographic	52.7813	25.2479	43.2712	0.9368
	lexical	27.7098	23.2430	52.5021	0.9384
	syntactic	21.4124	16.8862	47.8393	0.9265

Table 10
Performance across models and stages in the MER-TRANS trial (IT).

Model	Stage	$BLEU_o$	$BLEU_g$	SARI	BERT
do-nothing	-	100.0000	14.5584	11.2061	0.9178
gemma3-4B	orthotypographic	85.8245	15.8682	20.9526	0.9226
	lexical	15.9907	5.8777	37.4368	0.9070
	syntactic	22.3732	18.9100	55.0048	0.9210
qwen3.5-4B	orthotypographic	73.3017	19.1472	30.2468	0.9218
	lexical	49.6229	18.8363	40.4267	0.9171
	syntactic	35.1232	16.1959	39.7525	0.9169
aitana-7B	orthotypographic	64.1797	13.2326	18.5796	0.9041
	lexical	74.1620	13.9935	25.4872	0.9086
	syntactic	29.8815	6.4100	25.4276	0.8928
salamandra-7B	orthotypographic	83.4888	12.7368	20.1632	0.9121
	lexical	74.4544	12.1712	24.5006	0.9084
	syntactic	72.9873	12.1540	25.3527	0.9080
qwen3.5-9B	orthotypographic	74.5855	17.1028	28.3791	0.9200
	lexical	57.7810	17.6450	37.6882	0.9230
	syntactic	31.6164	15.4920	42.5732	0.9208
qwen3.6-27B	orthotypographic	53.7296	11.7636	29.6666	0.9163
	lexical	21.0800	15.2692	45.0757	0.9237
	syntactic	13.7141	12.2030	42.2043	0.9165