

HULAT2 at MER-TRANS 2026: Governed Multi-Agent Simplification for Spanish Easy-to-Read Generation

Lourdes Moreno¹, Paloma Martínez¹, Marco Antonio Sanchez-Escudero¹ and Miguel Domínguez-Gómez¹

¹Computer Science and Engineering Department, Universidad Carlos III de Madrid, Av. Universidad, 30, Leganés, 28911, Spain

Abstract

This paper describes the participation of HULAT2-UC3M in the Spanish track of MER-TRANS 2026, a shared task on multilingual Easy-to-Read translation. Three fully automatic Spanish runs were submitted. RUN1 and RUN2 used a LangGraph-based multi-agent workflow combining Gemini 2.5 Flash and RigoChat-7B-v2, parallel generation strategies, internal quality signals, Event-Condition-Action routing, controlled editing and traceable decisions. RUN1 used the base workflow, while RUN2 activated an additional lexical-support layer based on a glossary and lexical resources. RUN3 was a RigoChat-based generate-evaluate-regenerate baseline with prompt engineering and LoRA-based adaptation.

The official leaderboard reports BLEU-Orig, BLEU-Gold, SARI and BERTScore. During development, additional internal signals were also inspected, including semantic fidelity, readability, lexical simplicity, syntactic clarity and factual consistency. According to official SARI, RUN1 was the best HULAT2 run, with 44.0543 points, followed by RUN2 with 43.1049 and RUN3 with 38.5136. These results indicate that, in this task setting, signal-guided multi-agent routing outperformed the linear regeneration baseline. They also show that adding lexical support did not automatically improve reference-based scores. Further segment-level and document-level analysis are required to assess readability, factual consistency and user-oriented adequacy.

Keywords

Easy-to-Read, text simplification, multi-agent systems, Spanish, accessibility, human-centred NLP

1. Introduction

Access to understandable information is essential for participation, autonomy and access to public services. This is particularly relevant for people with cognitive accessibility needs, reading comprehension difficulties, low literacy, older adults, non-native speakers and users facing unfamiliar or specialised content. Cognitive accessibility is concerned with reducing barriers to understanding, communication and interaction, and has been formalised in standards such as ISO 21801-1:2020 [1]. In text-based communication, these barriers are addressed through approaches such as Plain Language (PL), Easy-to-Read (E2R) and automatic text simplification.

PL and E2R are related but not equivalent. PL focuses on making information clear, findable, understandable and usable for its intended audience, as defined by ISO 24495-1:2023 [2]. E2R usually applies stricter constraints on vocabulary, syntax, layout and validation, and is particularly oriented towards people with reading comprehension difficulties. In Spain, E2R practice is supported by UNE 153101:2018 EX, which provides recommendations for the production and validation of Easy-to-Read documents [3]. International guidance, such as the IFLA guidelines and the Inclusion Europe recommendations, also emphasises the need for clear language, simple structure and reader-oriented validation [4, 5]. For this reason, accessibility-oriented simplification should not be reduced to lexical substitution or sentence shortening; it also involves content selection, explicit relations between ideas, document structure, control of terminology, and the preservation of relevant information.

Automatic text simplification has evolved from rule-based and statistical approaches to neural

IberLEF 2026, September 2026, León, Spain

✉ lmoreno@inf.uc3m.es (L. Moreno); paloma.martinez@uc3m.es (P. Martínez); marcoasa@inf.uc3m.es

(M. A. Sanchez-Escudero); 100451258@alumnos.uc3m.es (M. Domínguez-Gómez)

🌐 <https://hulat.inf.uc3m.es/> (L. Moreno); <https://hulat.inf.uc3m.es/> (P. Martínez)

🆔 0000-0002-9021-2546 (L. Moreno); 0000-0003-3013-3771 (P. Martínez); 0009-0001-8163-5440 (M. A. Sanchez-Escudero)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

models and, more recently, Large Language Models (LLMs) [6, 7]. LLMs can perform lexical substitution, sentence splitting, paraphrasing, deletion of secondary information and explanatory rewriting. However, LLM-based simplification also introduces risks: outputs may be fluent and apparently simple while omitting relevant information, adding unsupported explanations, changing numerical information or altering the meaning of the source. This is especially relevant in accessibility-oriented settings, where the output must be not only simpler, but also reliable, usable and appropriate for the intended readers. In Spanish, previous work has explored decoder-based LLMs for Easy-to-Read generation using parallel E2R resources, fine-tuning and expert qualitative evaluation [8]. These results show the potential of LLM-based simplification, while reinforcing the need for terminology support and human validation.

Evaluation remains a central challenge. Metrics such as BLEU, SARI and BERTScore are useful for benchmarking, but they capture only partial aspects of quality [9, 10, 11]. BLEU rewards surface overlap with references, SARI evaluates add, keep and delete operations, and BERTScore estimates semantic similarity through contextual embeddings. Previous work has shown that automatic metrics may penalise valid reformulations and do not fully capture whether an adapted text is clearer, better organised or more useful for its intended audience [12, 13]. Other work has also pointed out that sentence-level evaluation can be too narrow for document-level simplification and that simplicity and meaning preservation should be assessed as different dimensions [14, 15]. Human-centred evaluation studies further show that expert judgments, automatic scores and actual reader comprehension may diverge [16, 17].

Recent workshops and shared tasks have contributed to more systematic evaluation settings for simplification, readability and accessibility. The TSAR series has consolidated research on text simplification, accessibility and readability [18, 19, 20, 21]. TSAR 2025 included a shared task on readability-controlled text simplification, where systems had to adapt texts to specified CEFR-based proficiency levels while preserving meaning and fluency [22]. READI 2024 addressed tools and resources for people with reading difficulties and included work on document-level simplification and evaluation beyond isolated sentences [23, 15]. In the Spanish context, CLEARS 2025 addressed PL and E2R adaptation within IberLEF, using semantic similarity and the Fernández-Huerta readability index as official metrics [24]. These initiatives show a shift from generic simplification towards controlled, multilingual, language-specific and accessibility-oriented evaluation settings. They also highlight the difficulty of balancing semantic preservation, readability, user-oriented adequacy and reference alignment.

MER-TRANS 2026 extends this line of work to multilingual Easy-to-Read translation. The task asks participating systems to generate E2R-oriented versions of complex texts, with official results reported using BLEU-Orig, BLEU-Gold, SARI and BERTScore [25]. The contribution of this paper is to describe the HULAT2-UC3M Spanish submissions to MER-TRANS 2026, using a signal-guided multi-agent architecture for controlled E2R-oriented generation, and to analyse the official results.

2. Task and Data

MER-TRANS 2026 was organised as part of IberLEF 2026 [25, 26]. The shared task uses data resources covering multilingual Easy-to-Read adaptation and sentence difficulty assessment. The iDEM corpus provides human-annotated original and Easy-to-Read texts designed to support access to democratic participatory processes [27]. The shared-task documentation also includes the Arabic sentence difficulty classification dataset introduced by Khallaf and Sharoff [28].

MER-TRANS 2026 evaluates automatic systems for generating Easy-to-Read (E2R) adaptations from original complex texts. HULAT2-UC3M submitted only Spanish runs, and no human intervention was used during official inference.

The MER-TRANS evaluation materials describe several metric families for the shared task, including surface-similarity metrics, simplification-oriented metrics, semantic-similarity metrics, readability metrics and complexity classifiers [29]. However, the official evaluator repository used for the shared task computes BLEU, SARI and BERTScore, and the published leaderboard reports BLEU against the original text (BLEU-Orig), BLEU against the gold E2R reference (BLEU-Gold), SARI and BERTScore

[30, 29]. Therefore, official results are reported in this paper using these leaderboard metrics.

SARI is especially relevant for simplification because it evaluates add, keep and delete operations with respect to the source and the reference [10]. BLEU provides a surface-overlap signal, while BERTScore provides a complementary semantic-similarity signal based on contextual embeddings [11]. These metrics offer a shared benchmark for the task, but they do not fully capture accessibility-oriented quality dimensions such as factual consistency, terminology support, readability, document coherence or user-oriented adequacy.

For development-time evaluation, the HULAT2 systems were not configured exclusively to maximise the official scores. SARI, BLEU and BERTScore were used as external benchmarking indicators, while candidate generation, routing and retry decisions were guided by a broader set of internal quality signals, including simplification, semantic-similarity, factual-consistency and readability indicators [31]. In the MER-TRANS experiments, these development-time signals were used to support prompt calibration, Event-Condition-Action (ECA) routing, retry conditions and qualitative error analysis.

3. System Description

This section describes the resources, empirical evidence base and system configurations used for the HULAT2-UC3M Spanish submissions to MER-TRANS 2026. The section first presents the resources used during development, then summarises the submitted runs, and finally describes the multi-agent workflow and the baseline system.

3.1. Resources and Evidence Base

For system calibration, prompt refinement and glossary construction, publicly available Spanish E2R materials related to institutional, civic and public information contexts were used. These resources supported the analysis of sentence segmentation patterns, explanatory formulations, lexical choices and recurrent E2R conventions. They included *La Constitución Española en lectura fácil*, citizen participation guides, electoral participation materials, and opposition-exam preparation materials published by Plena Inclusión and public administrations. A detailed list of the public resources used for calibration and lexical support is provided in Appendix A.

Part of these materials was used to build an internal HULAT-UC3M glossary of learning words and short explanatory forms. The glossary was used as lexical support in RUN2. RUN2 also incorporated Spanish medical lexical simplification resources [32]. Although the main MER-TRANS domain is not medical, these resources were useful as additional lexical-simplification material for difficult terms and as explanatory paraphrases. The resources were not part of the official MER-TRANS test set.

In addition to these textual resources, the system design was informed by an empirical evidence base previously compiled for accessible text generation. This evidence base includes results from user studies, annotated resources and accessibility-oriented evaluations, and was operationalised as quality signals, checklists, ECA rules and review criteria for accessible simplification [33]. In the MER-TRANS runs, this evidence was not used as human intervention during official inference, but as development-time support for prompt adjustment, routing rules, retry conditions and qualitative analysis.

3.2. Submitted Runs

Three Spanish runs were submitted. RUN1 used the base LangGraph-based multi-agent workflow, with parallel generation, candidate evaluation, ECA-style routing and controlled editing. RUN2 used the same workflow, but activated the lexical agent before generation in order to test whether glossary-based and lexical-resource support improved the base configuration. RUN3 was a separate RigoChat-based generate-evaluate-regenerate baseline, designed to compare the multi-agent routing strategy with a more linear simplification pipeline.

Table 1 summarises the submitted configurations at pipeline level. The table distinguishes the multi-agent workflow used in RUN1 and RUN2 from the linear baseline workflow used in RUN3, and specifies

which stages were shared, optional or exclusive to each run.

Table 1

Pipeline-level description of the HULAT2-UC3M submitted configurations.

Pipeline stage	Stage or module	Main resource or model	Function	Runs
<i>Multi-agent workflow used in RUN1 and RUN2</i>				
Optional lexical support	Lexical agent	Glossary and lexical resources	Detects difficult terms and proposes lexical support before generation. This stage is skipped when the lexical agent is disabled.	RUN2
Pre-analysis	Pre-analysis module	Rule-based feature extraction	Detects structural, lexical and factual-risk features, including sentence length, punctuation complexity, numbers, dates, negations, conditions and glossary terms.	RUN1, RUN2
Parallel generation	Generator A	Gemini 2.5 Flash	Produces a conservative Plain Language-oriented candidate, prioritising meaning preservation and minimal reformulation.	RUN1, RUN2
Parallel generation	Generator B	RigoChat-7B-v2	Produces a structurally simplified candidate, prioritising sentence splitting, directness and reduced lexical difficulty.	RUN1, RUN2
Parallel generation	Generator C	Gemini 2.5 Flash	Produces an Easy-to-Read-oriented candidate guided by short sentences, common vocabulary, explicit structure and an operational CEFR A2-like target.	RUN1, RUN2
Candidate assessment	Candidate evaluator	Internal quality signals	Assesses candidates using semantic, factual, syntactic, lexical, readability and robustness signals.	RUN1, RUN2
Routing and control	ECA router	Event-Condition-Action rules	Selects, edits, merges or retries candidates according to the detected quality profile and risk signals.	RUN1, RUN2
Controlled editing	Merger-editor	Restricted editing prompt	Applies minimal controlled edits or fuses candidates only when complementary improvements are detected without semantic-risk signals.	RUN1, RUN2
Final validation	Final evaluator	Internal validation signals	Rechecks factual consistency, semantic preservation and robustness before producing the final output.	RUN1, RUN2
<i>Linear baseline workflow used in RUN3</i>				
Generation	Baseline generator	RigoChat-7B-v2 with prompt engineering and LoRA adaptation	Generates an Easy-to-Read-oriented output using a controlled baseline configuration.	RUN3
Quality control	Baseline evaluator	Internal readability and semantic-preservation checks	Evaluates the generated output using internal quality criteria, including readability, semantic preservation and critical-error checks.	RUN3
Regeneration	Regeneration loop	Regeneration prompt	Attempts regeneration when internal quality criteria are not met.	RUN3

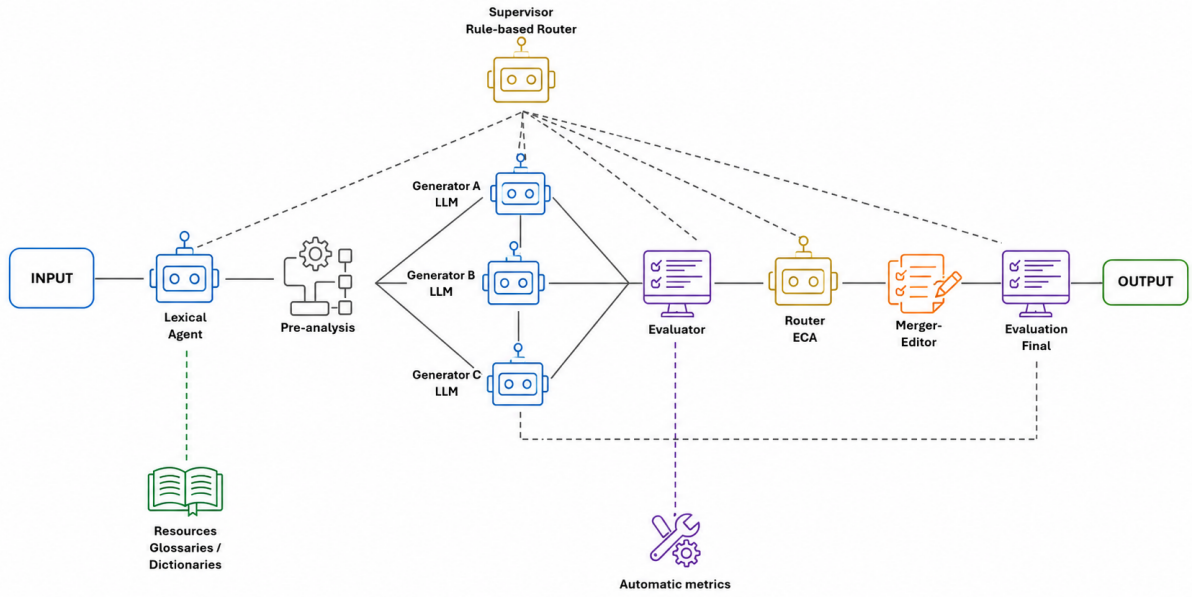


Figure 1: LangGraph-based multi-agent workflow used for RUN1 and RUN2 in the MER-TRANS 2026 Spanish submissions.

3.3. Multi-Agent Workflow

RUN1 and RUN2 were generated with the same LangGraph-based multi-agent architecture. LangGraph was used to implement a stateful workflow with specialised nodes, conditional routing, retry loops and traceable intermediate decisions [34]. The system combines Gemini 2.5 Flash [35] and RigoChat-7B-v2 [36]. RigoChat-7B-v2 is a Spanish-oriented model based on Qwen2.5-7B-Instruct and further adapted for Spanish queries.

Workflow overview. Figure 1 summarises the workflow used for the MER-TRANS 2026 Spanish submissions. A demonstration video illustrating the workflow execution and traceable decisions is available online.¹ The lexical agent is optional: it was disabled in RUN1 and enabled in RUN2. The figure also shows the role of the rule-based supervisor, the use of automatic evaluation signals and the controlled path from candidate generation to final output.

The workflow starts from the input text. When the lexical agent is enabled, lexical resources are consulted before pre-analysis; when it is disabled, the input is passed directly to the pre-analysis stage. The pre-analysis module characterises the input before generation by detecting structural, lexical and factual-risk features, including sentence length, punctuation complexity, numbers, dates, negations, conditions and possible glossary terms. These features are used to determine whether the text requires conservative rewriting, syntactic simplification, lexical support or stronger semantic control.

Evidence-informed signals and control rules. Some of these signals, rules and threshold ranges were informed by the empirical evidence base. In particular, the evidence supported the definition of quality priorities, checklists, ECA-style rules and acceptance, warning, retry and rejection criteria for accessible simplification [33]. Other signals were defined as operational heuristics to control generation robustness, lexical difficulty, sentence structure and unsupported content during the MER-TRANS configuration.

¹<https://youtu.be/FV4bymTRcQo?si=dgYNKsc2j3lxXUP0>

In this configuration, high-risk semantic and factual signals were treated as stricter constraints, while syntactic, lexical and E2R-format signals were treated as calibrable criteria during development. The signal taxonomy, ECA rule specification and threshold configuration are documented in the accompanying material archived in Zenodo [37]. In this paper, the operational signal categories and decision logic are reported at the level needed to interpret the submitted runs, while the extended technical specification is provided as supplementary material.

Parallel generation strategies. Three candidate generators are launched in parallel. Generator A uses Gemini 2.5 Flash with a conservative Plain Language-oriented strategy, prioritising meaning preservation, clarity and minimal reformulation. Generator B uses RigoChat-7B-v2 with a more active Plain Language-oriented strategy, aimed at simplifying sentence structure, reducing lexical difficulty and improving directness while preserving the source meaning.

Generator C uses Gemini 2.5 Flash with an Easy-to-Read-oriented strategy, guided by the Spanish UNE 153101 recommendations and by a target linguistic profile close to CEFR A2. The CEFR A2 criterion was used as an operational linguistic-complexity target, not as evidence of actual reading comprehension. This third strategy therefore combines E2R-oriented constraints, short sentences, common vocabulary, explicit structure, visual segmentation when appropriate, and internal checks for meaning preservation and factual consistency.

Candidate diversification was intended to compare conservative, Plain Language-oriented and Easy-to-Read-oriented rewriting behaviours before selecting or refining the final output.

Candidate evaluation and routing. Each candidate is assessed by the candidate evaluator. The evaluator does not rely on a single quality score, but on groups of internal signals related to semantic and factual preservation, syntactic clarity, lexical simplicity, readability and robustness. These signals include checks for the preservation of numbers, dates, negations and conditions; sentence-length and splitting indicators; difficult-word and glossary-related indicators; readability-support signals; unsupported-content heuristics; and robustness checks for repetition or unexpected script changes.

The routing component applies ECA-style rules over these signals. Three main routing strategies are available. In V1, the best candidate is selected without modification. In V2, one candidate is selected as the base and minimally edited using another candidate as support. In V3, two candidates are merged only when they provide complementary improvements and no semantic-risk signal is detected. The router may also trigger targeted retries when semantic preservation, syntactic clarity, lexical simplification or E2R formatting is insufficient.

Controlled editing and final validation. The merger-editor is used only under controlled conditions. It receives the original text, the selected base candidate, an alternative candidate and a restricted set of permitted improvements. Its role is not to freely rewrite the text, but to make minimal edits or perform controlled fusion while preserving relevant information, numbers, dates, negations, conditions and domain-specific terms.

The final evaluator verifies the selected or edited output before submission. Critical signals related to factual consistency, semantic preservation and generation robustness are recalculated at this stage. The final decision is recorded together with the selected strategy, action and trace information. These traces were used during development for qualitative inspection and rule calibration, but no human intervention was used during official inference.

This workflow was used in both RUN1 and RUN2. The only architectural difference between the two submitted configurations was the activation of the lexical agent in RUN2, which introduced glossary-based and lexical-resource support before generation.

The workflow described in this section corresponds to the strategy followed for the MER-TRANS 2026 Spanish submissions. It should not be interpreted as the only possible strategy of the broader HULAT-UC3M accessible text generation infrastructure. Other configurations, prompts, routing rules,

validation criteria and human-supervision mechanisms may be used in non-competition scenarios, depending on the domain, target users, available references and deployment constraints.

The signal taxonomy, ECA rule specification and threshold configuration are documented in the accompanying supplementary material.

3.4. Baseline System

RUN3 was implemented as a linear generate–evaluate–regenerate baseline. It used RigoChat-7B-v2 with prompt-engineering strategies, post-processing and LoRA-based adaptation, a parameter-efficient fine-tuning technique that freezes the base model and injects trainable low-rank matrices into the model layers [38].

The LoRA adapter was trained with public Spanish E2R material, mainly the Spanish Constitution in original and Easy-to-Read versions. The internal split contained 169 aligned articles: 126 for training and 43 for internal testing. During development, RigoChat-7B-v2 and Latxa-Qwen3-VL-8B-Instruct were compared under the same LoRA, prompting and quality-control conditions. Although Latxa obtained higher internal SARI and MeaningBERT scores in some configurations, RigoChat-7B-v2 was selected for RUN3 because it produced more stable outputs, with fewer repetitions and truncations.

The final RUN3 configuration used RigoChat-7B-v2 with the LoRA adapter, explicit A2/E2R-oriented prompts, controlled generation parameters and a regeneration prompt activated when internal quality criteria were not met. The generation parameters were set to favour controlled outputs: `max_new_tokens=512`, `temperature=0.3`, `top_p=0.85`, `top_k=10`, `repetition_penalty=1.25`, `no_repeat_ngram_size=3` and `length_penalty=0.8`. The quality-control module checked readability and semantic preservation using Fernández-Huerta and MeaningBERT, and also detected critical errors such as repetition, truncation, language mixing, loss of numbers or dates, loss of negations and excessive information loss [39, 40]. Regeneration was attempted up to two times.

4. Experimental Setup

During development, the systems were not configured exclusively to maximise the official MER-TRANS scores. The official-style metrics were used as external benchmarking indicators, but no single metric was treated as a complete proxy for E2R quality. Accessibility-oriented simplification also requires checking whether the output preserves meaning, avoids unsupported additions, simplifies syntax and vocabulary, and remains readable for the intended audience.

To support this broader view, the HULAT-UC3M *bilingual_simplification_metrics* framework was used as a development-time support tool [31]. During calibration, when paired source–reference data were available, reference-based metrics such as SARI, BLEU and BERTScore were used to compare candidate behaviours and adjust prompts, routing rules and retry conditions. These metrics were not available during official test inference because the gold references were not provided to participants.

During official inference, routing decisions were mainly guided by reference-free internal signals computed from the source text and the candidate output. These signals were also informed by the empirical evidence base described in Section 3.1, which supported the definition of quality priorities, checklists and ECA-style control rules. During development and calibration, additional semantic and factual-consistency metrics available in the *bilingual_simplification_metrics* framework, including MeaningBERT, AlignScore, QuestEval and SummaC, were inspected when the required inputs and model configurations were available [31].

For the MER-TRANS official inference setting, the internal signals used for routing were grouped into five operational categories:

- Semantic and factual preservation: semantic-preservation support signals and checks for numbers, dates, negations, conditions and source-supported content.
- Readability: Fernández-Huerta readability and related sentence-level indicators.
- Syntactic clarity: sentence length, punctuation complexity and splitting behaviour.

Table 2

Official MER-TRANS Spanish results for the three HULAT2-UC3M configurations.

Run	BLEU-Orig	BLEU-Gold	SARI	BERTScore
RUN1	35.5730	17.5715	44.0543	0.9263
RUN2	34.1564	16.1484	43.1049	0.9248
RUN3	2.4037	5.3373	38.5136	0.9114

- Lexical simplicity: difficult-word ratio, glossary hits, unresolved difficult terms and lexical substitution behaviour.
- Robustness: repeated fragments, unexpected scripts and other generation errors.

These development-time metrics and inference-time signals were used to calibrate prompts, define retry conditions and route candidate outputs through ECA-style rules. For example, a candidate with adequate meaning preservation but overly long sentences could trigger a structural simplification action; a candidate with unresolved glossary terms could activate lexical support; and a candidate with factual-risk signals could be rejected or sent to a conservative retry.

The internal signals were therefore used to configure and control the system, not as alternative official scores. The official results reported in Section 5 correspond only to the leaderboard metrics computed by the organizers.

5. Results

Table 2 isolates the official results of the three HULAT2-UC3M runs in order to compare the submitted configurations directly. Table 3 then places these runs within the complete Spanish leaderboard.

RUN1 obtained the best official SARI score among the three submitted HULAT2 runs. It improved over RUN3 by 5.5407 SARI points, which corresponds to a relative improvement of approximately 14.39% over the baseline SARI score. RUN2 also improved over RUN3 by 4.5913 SARI points. These results suggest that, in this setting, the signal-guided multi-agent workflow was more effective than the linear generate-evaluate-regenerate baseline.

RUN2, which activated the lexical agent, obtained a slightly lower official SARI score than RUN1. This suggests that the lexical support introduced in this configuration did not improve the official reference-based ranking. However, this result should not be interpreted as a general conclusion against lexical resources. Rather, it indicates that glossary-based or domain-informed lexical substitutions must be carefully balanced with semantic preservation, contextual adequacy and alignment with the reference simplifications.

RUN3 obtained lower official scores and served mainly as an exploratory baseline. Its very low BLEU-Orig score indicates that the baseline outputs diverged strongly from the source wording, but this divergence was not compensated by higher similarity to the gold E2R reference. This suggests that the baseline sometimes rewrote too aggressively or produced outputs that were not sufficiently aligned with the expected simplification style.

Table 3 reports the complete official Spanish leaderboard, sorted by SARI. HULAT2-RUN1 ranked 6th out of 19 Spanish submissions, HULAT2-RUN2 ranked 9th, and HULAT2-RUN3 ranked 16th. Therefore, HULAT2-RUN1 was competitive within the Spanish track, although it was not the top-ranked system. The main claim of this paper is not state-of-the-art performance, but the comparative finding that the signal-guided multi-agent architecture outperformed the RigoChat-based linear baseline submitted by the same team.

The complete leaderboard also illustrates why BLEU-Orig and BERTScore must be interpreted with caution in this task. The DoNothing baseline preserves the original text and therefore obtains the highest BLEU-Orig and a high BERTScore, but it obtains the lowest SARI score. This confirms that source overlap and semantic similarity alone do not measure whether a text has actually been simplified.

Table 3

Official Spanish leaderboard for MER-TRANS 2026, sorted by SARI. HULAT2-UC3M runs are highlighted in bold.

Rank	Team	Run	BLEU-Orig	BLEU-Gold	SARI	BERTScore
1	ClearText	RUN1	18.0013	15.4483	47.0168	0.9220
2	VICOMTECH	RUN3	20.7912	14.1086	45.5843	0.9241
3	HULAT1	RUN2	33.1815	17.7135	45.0182	0.9272
4	HULAT1	RUN3	31.5608	17.1627	44.7465	0.9272
5	VICOMTECH	RUN1	12.4857	9.3316	44.1880	0.9114
6	HULAT2	RUN1	35.5730	17.5715	44.0543	0.9263
7	HULAT1	RUN1	37.1581	18.2081	43.3874	0.9276
8	BASELINE1	RUN1	18.9491	10.8850	43.2873	0.9130
9	HULAT2	RUN2	34.1564	16.1484	43.1049	0.9248
10	VICOMTECH	RUN2	32.0715	14.7485	42.1724	0.9246
11	HumanAI-UCM	RUN2	26.9792	11.8405	41.6196	0.9111
12	FACILE	RUN1	3.0009	2.8644	41.6129	0.8783
13	NIL_UCM	RUN2	11.3818	4.7980	40.5175	0.9026
14	HumanAI-UCM	RUN1	38.3402	15.0830	40.4163	0.9209
15	NIL_UCM	RUN1	6.0574	2.5775	39.6873	0.8780
16	HULAT2	RUN3	2.4037	5.3373	38.5136	0.9114
17	BASELINE2	RUN1	29.0913	9.3904	37.1761	0.9019
18	HumanAI-UCM	RUN3	45.0236	14.5298	36.2450	0.9125
19	DoNothing	RUN1	100.0000	23.2176	13.8687	0.9297

6. Error Analysis

A preliminary error-oriented analysis was conducted based on the official scores, the comparison between the submitted configurations, and the behaviour observed during development. Since the official gold references were not available for detailed segment-level inspection at the time of writing, this section should be interpreted as a preliminary analysis rather than as a complete error annotation.

The results point to three main areas for further inspection. First, the lower performance of RUN3 suggests that the baseline may have produced overly aggressive rewritings, reduced reference alignment or insufficient control over information preservation. This is consistent with the large drop in BLEU-Orig and the lower BLEU-Gold and SARI scores observed for this run.

Second, the difference between RUN1 and RUN2 suggests that lexical support requires careful calibration. Glossary-based explanations or lexical substitutions may be useful from an accessibility perspective, but they can also increase divergence from the expected reference if they introduce unnecessary paraphrases, alter terminology in context or expand information beyond what the reference expresses.

Third, the leaderboard confirms that high source overlap and high semantic similarity are not sufficient indicators of successful simplification. The DoNothing baseline shows that preserving the original text can lead to high BLEU-Orig and BERTScore values while failing to simplify the input. This supports the need to complement official metrics with qualitative and accessibility-oriented error analysis.

Future error analysis should combine local segment-level inspection with document-aware assessment. Segment-level inspection is needed to detect local errors such as loss of numbers, dates, negations or conditions, unsupported additions, excessive rewriting and insufficient simplification. Document-aware assessment is needed to analyse consistency across segments, terminology stability, the presence of repeated or missing explanations, the coherence of the reconstructed document, and the homogeneity of the E2R style.

7. Conclusions

This paper has presented the HULAT2-UC3M participation in the Spanish track of MER-TRANS 2026. Three fully automatic runs were submitted: two configurations of a LangGraph-based multi-agent workflow and one linear generate-evaluate-regenerate baseline.

RUN1 was the highest-scoring HULAT2 run according to the official SARI score, with 44.0543 points, followed by RUN2 with 43.1049 and RUN3 with 38.5136. These results suggest that, for this task, candidate diversification, automatic evaluation signals, ECA routing and controlled editing were more effective than the linear baseline. RUN1 and RUN2 differed only in the activation of the lexical agent, so the lower score of RUN2 should be interpreted within the official evaluation setting, not as a general conclusion against lexical resources. Rather, it shows that lexical support requires careful calibration: it may support user-centred explanations, terminology control, and domain-specific accessibility, but it does not automatically improve performance on reference-based metrics.

Beyond leaderboard performance, the HULAT2-UC3M approach is oriented towards increasing the robustness and controllability of accessible text generation systems. The submitted workflow was designed not only to generate simplified text, but also to monitor semantic preservation, factual consistency, readability, lexical difficulty and generation errors through internal signals and traceable routing decisions.

Future work will focus on document-aware error analysis, combining local segment-level inspection with assessment of consistency, terminology stability, coherence and E2R style across reconstructed documents. Expert review, user-centred validation and the integration of Human-in-the-Loop and Human-on-the-Loop mechanisms will also be explored in non-competitive deployments. This will make it possible to complement automatic metrics with accessibility-oriented evidence about comprehension, adequacy and usefulness for target users.

Declaration on Generative AI

Generative AI tool OpenAI ChatGPT was used to improve the English wording and to assist with editing some LaTeX text. The authors reviewed and edited all AI-assisted changes and take full responsibility for the final content of the paper.

Availability of Code and Supplementary Material

The implementation code associated with the HULAT2-UC3M MER-TRANS 2026 submissions is available in the task-associated GitHub repository: https://github.com/hulat-group/mertrans_2026. The accompanying reproducibility material, including the signal taxonomy, ECA rule specification, routing logic, validation signals and configuration documentation, is archived in Zenodo [37].

Acknowledgments

This work has been supported by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

References

- [1] International Organization for Standardization, ISO 21801-1:2020 Cognitive accessibility – Part 1: General guidelines, <https://www.iso.org/standard/71711.html>, 2020.
- [2] International Organization for Standardization, ISO 24495-1:2023 Plain language – Part 1: Governing principles and guidelines, <https://www.iso.org/standard/78907.html>, 2023.
- [3] Asociación Española de Normalización, UNE 153101:2018 EX. Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos, <https://www.une.org/>, 2018.

- [4] M. Nomura, G. S. Nielsen, B. Tronbacke, Guidelines for easy-to-read materials, <https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/professional-report/120.pdf>, 2010.
- [5] Inclusion Europe, Information for all: European standards for making information easy to read and understand, <https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/>, 2009.
- [6] H. Saggion, Automatic Text Simplification, Synthesis Lectures on Human Language Technologies, Springer Cham, 2017. URL: <https://link.springer.com/book/10.1007/978-3-031-02166-4>. doi:10.1007/978-3-031-02166-4.
- [7] S. Nisioi, S. Štajner, S. P. Ponzetto, L. P. Dinu, Exploring neural text simplification models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 85–91. URL: <https://aclanthology.org/P17-2014/>. doi:10.18653/v1/P17-2014.
- [8] P. Martínez, A. Ramos, L. Moreno, Exploring large language models to generate easy to read content, Frontiers in Computer Science 6 (2024) 1394705. URL: <https://doi.org/10.3389/fcomp.2024.1394705>. doi:10.3389/fcomp.2024.1394705.
- [9] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: <https://aclanthology.org/P02-1040/>. doi:10.3115/1073083.1073135.
- [10] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: <https://aclanthology.org/Q16-1029/>. doi:10.1162/tac1_a_00107.
- [11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [12] F. Alva-Manchego, C. Scarton, L. Specia, The (Un)Suitability of automatic evaluation metrics for text simplification, Computational Linguistics 47 (2021) 861–889. doi:10.1162/coli_a_00418.
- [13] M. Maddela, Y. Dou, D. Heineman, W. Xu, LENS: A learnable evaluation metric for text simplification, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 16383–16408. URL: <https://aclanthology.org/2023.acl-long.905/>. doi:10.18653/v1/2023.acl-long.905.
- [14] M. Maddela, F. Alva-Manchego, Adapting sentence-level automatic metrics for document-level simplification evaluation, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 6444–6459. URL: <https://aclanthology.org/2025.naacl-long.327/>. doi:10.18653/v1/2025.naacl-long.327.
- [15] L. Cripwell, J. Legrand, C. Gardent, Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation, in: Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 1–14. URL: <https://aclanthology.org/2024.readi-1.1/>.
- [16] S. Agrawal, M. Carpuat, Do text simplification systems preserve meaning? a human evaluation via reading comprehension, Transactions of the Association for Computational Linguistics 12 (2024) 432–448. URL: <https://aclanthology.org/2024.tacl-1.24/>. doi:10.1162/tac1_a_00653.
- [17] L. Carrer, A. Säuberli, M. Kappus, S. Ebling, Towards holistic human evaluation of automatic text simplification, in: Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 71–80. URL: <https://aclanthology.org/2024.humeval-1.7/>.
- [18] H. Saggion, S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, M. Zampieri, Findings of the TSAR-2022 shared task on multilingual lexical simplification, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 271–283. URL: <https://aclanthology.org/2022.tsar-1.31/>. doi:10.18653/v1/2022.tsar-1.31.
- [19] S. Štajner, H. Saggion, M. Shardlow, F. Alva-Manchego (Eds.), Proceedings of the Second Workshop

- on Text Simplification, Accessibility and Readability, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023. URL: <https://aclanthology.org/2023.tsar-1.0/>.
- [20] M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024. URL: <https://aclanthology.org/2024.tsar-1.0/>. doi:10.18653/v1/2024.tsar-1.0.
 - [21] 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. URL: <https://aclanthology.org/2025.tsar-1.0/>. doi:10.18653/v1/2025.tsar-1.0.
 - [22] F. Alva-Manchego, R. Stodden, J. M. Imperial, A. Barayan, K. North, H. Tayyar Madabushi, Findings of the TSAR 2025 shared task on readability-controlled text simplification, in: Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), Association for Computational Linguistics, Suzhou, China, 2025, pp. 116–130. URL: <https://aclanthology.org/2025.tsar-1.8/>. doi:10.18653/v1/2025.tsar-1.8.
 - [23] R. Wilkens, R. Cardon, A. Todirascu, N. Gala (Eds.), Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024. URL: <https://aclanthology.org/2024.readi-1.0/>.
 - [24] B. Botella-Gil, I. Espinosa-Zaragoza, A. Bonet-Jover, M. Madina, L. Molino Piñar, P. Moreda, I. Gonzalez-Dios, M. T. Martín-Valdivia, L. A. Ureña-López, Overview of clears at iberlef 2025: Challenge for plain language and easy-to-read adaptation for spanish texts, *Procesamiento del Lenguaje Natural* (2025) 393–400. doi:10.26342/2025-75-28.
 - [25] H. Saggion, M. Tareh, N. Khallaf, D. Adanza, S. Bott, N. Pérez-Rojas, A. Rascón-Alcaina, S. Szasz, Overview of MER-TRANS at IberLEF 2026: First shared task on multilingual easy-to-read translation, *Procesamiento del Lenguaje Natural* 77 (2026).
 - [26] 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.
 - [27] S. Bott, V. Riegler, H. Saggion, A. Rascón Alcaina, N. Khallaf, A multilingual human annotated corpus of original and easy-to-read texts to support access to democratic participatory processes, in: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 1117–1128. URL: <https://lrec.elra.info/lrec2026-main-087>. doi:10.63317/4b56cza6e7zk.
 - [28] N. Khallaf, S. Sharoff, Automatic difficulty classification of Arabic sentences, in: N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouni, F. Bougares, N. Tomeh, I. Abu Farha, S. Touileb (Eds.), Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 2021, pp. 105–114. URL: <https://aclanthology.org/2021.wanlp-1.11/>.
 - [29] MER-TRANS Organizers, MER-TRANS 2026: Multilingual easy-to-read translation shared task, <https://lastus-taln-upf.github.io/mertrans-iberlef-2026/>, 2026.
 - [30] LaSTUS-TALN-UPF, MER-TRANS Evaluator 2026: Evaluator for shared-task submissions, <https://github.com/LaSTUS-TALN-UPF/mertrans-evaluator-2026>, 2026.
 - [31] HULAT-UC3M, Bilingual Simplification Metrics: A unified evaluation framework for text simplification (ES/EN), https://github.com/hulat-group/bilingual_simplification_metrics, 2026.
 - [32] L. Campillos-Llanos, A. R. Terroba Reinales, S. Zakhir Puig, A. Valverde-Mateos, A. Capllonch-Carrión, Building a comparable corpus and a benchmark for spanish medical text simplification, *Procesamiento del Lenguaje Natural* (2022) 189–196. doi:10.26342/2022-69-16.
 - [33] L. Moreno, P. Martínez, A human-in/on-the-loop framework for accessible text generation, in: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 7236–7247. URL: <https://lrec.elra.info/lrec2026-main-574>. doi:10.63317/2gtngp2nmx63.

- [34] LangChain, LangGraph overview, 2026. URL: <https://docs.langchain.com/oss/python/langgraph/overview>.
- [35] Google AI for Developers, Gemini 2.5 flash, 2026. URL: <https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash>.
- [36] G. Santamaría Gómez, G. García Subies, P. Gutiérrez Ruiz, M. González Valero, N. Fuertes, H. Montoro Zamorano, C. Muñoz Sanz, L. Rosado Plaza, N. Aldama García, D. Betancur Sánchez, K. Sushkova, M. Guerrero Nieto, Á. Barbero Jiménez, RigoChat 2: An adapted language model to spanish using a bounded dataset and reduced hardware, 2025. URL: <https://arxiv.org/abs/2503.08188>. arXiv:2503.08188.
- [37] L. Moreno, P. Martínez, AIGov-Access: AI Governance for Accessibility-Oriented Text Adaptation — MER-TRANS 2026 Profile, 2026. URL: <https://doi.org/10.5281/zenodo.20855013>. doi:10.5281/zenodo.20855013.
- [38] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [39] D. Beauchemin, H. Saggion, R. Khoury, MeaningBERT: Assessing meaning preservation between sentences, *Frontiers in Artificial Intelligence* 6 (2023). URL: <https://www.frontiersin.org/articles/10.3389/frai.2023.1223924>. doi:10.3389/frai.2023.1223924.
- [40] J. Fernández Huerta, Medidas sencillas de lecturabilidad, *Consigna* 214 (1959) 29–32.

A. Public Resources Used for Calibration and Lexical Support

The following public resources were used during system calibration, prompt refinement or glossary construction:

- Plena Inclusión España, The Spanish Constitution in Easy-to-Read Format, Easy-to-Read publication, <https://www.plenainclusion.org/publicaciones/buscador/la-constitucion-espanola-en-lectura-facil/>.
- Easy-to-Read citizen participation guide from the Spanish transparency portal: https://transparencia.gob.es/content/dam/transparencia_home/multimedia/pdf/guias-de-lectura-facil/Guia_participacion_ciudadana_v3_sinplanes_n.pdf
- Citizen participation material available through SlideShare: <https://es.slideshare.net/slideshow/lectura-facil-participacin-ciudadanapdf/255259357>
- Accessible citizen participation manual from CARM: <https://participa.carm.es/documents/5690123/6505112/C%C3%B3mo+participar.+Manual+Accesible/b6624045-65ce-43c3-8ab8-ece6f8852d23>
- Plena Inclusión opposition-exam materials: <https://www.plenainclusion.org/coleccion/temario-de-oposicion/>
- Easy-to-Read guide for electoral participation: <https://mivotocuenta.es/wp-content/uploads/2019/02/Guia-para-participar-en-las-elecciones-en-Lectura-F%C3%A1cil.pdf>
- SimpMedLexSp [32] and related Spanish medical simplification resources.