

Overview of MER-TRANS at IberLEF 2026: First Shared Task on Multilingual Easy-to-Read Translation

Resumen de MER-TRANS en IberLEF 2026: Primera tarea compartida multilingüe de traducción a lectura fácil

Horacio Saggion,¹ Mehrzad Tareh,¹ Nouran Khallaf,²
Stefan Bott,¹ Daniel Adanza,¹ Almudena Rascón Alcaina,³
Nelson Pérez Rojas,⁴ Sandra Szasz¹

¹Universitat Pompeu Fabra

²University of Leeds

³Plena Inclusión Madrid

⁴Instituto Tecnológico de Costa Rica

{horacio.saggion, mehrzad.tareh, stefan.bott, daniel.adanza, sandra.szasz}@upf.edu
N.Khallaf@leeds.ac.uk almudenascon@plenamadrid.org nperez@itcr.ac.cr

Abstract: This paper presents MER-TRANS, a shared task on multilingual easy-to-read translation organised within the context of IberLEF, the evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages. The objective of MER-TRANS is to challenge the NLP community to develop solutions that produce easy-to-read versions of texts and sentences in Arabic, Catalan, Italian, and Spanish. The task features a multilingual corpus of original documents and their easy-to-read adaptations in Catalan, Italian, and Spanish, as well as a sentence-aligned dataset of original sentences and their simplifications in Arabic. MER-TRANS attracted seven teams which submitted a total of 42 systems (including the organisers' baselines), all of which were evaluated using automatic metrics.

Keywords: Automatic Text Simplification, Easy-to-Read, Linguistic Accessibility, Multilingual Text Processing.

Resumen: Este artículo presenta MER-TRANS, una tarea compartida sobre traducción multilingüe a lectura fácil, organizada en el marco de IberLEF, la campaña de evaluación de sistemas de procesamiento del lenguaje natural (PLN) en español y otras lenguas ibéricas. El objetivo de MER-TRANS es desafiar a la comunidad de PLN a desarrollar soluciones que generen versiones en lectura fácil de textos y oraciones en árabe, catalán, italiano y español. La tarea cuenta con un corpus multilingüe compuesto por documentos originales y sus adaptaciones a lectura fácil en catalán, italiano y español, así como con un conjunto de datos alineado a nivel de oración que incluye oraciones originales y sus simplificaciones en árabe. MER-TRANS contó con la participación de siete equipos que presentaron un total de 42 sistemas (incluidos los sistemas de referencia de los organizadores), todos los cuales fueron evaluados mediante métricas automáticas.

Palabras clave: Simplificación Automática de Textos, Fácil Lectura, Accesibilidad Lingüística, Procesamiento de Texto Multilingüe.

1 Introduction

Linguistic accessibility is increasingly recognised as a fundamental citizens' right. The United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) in-

cludes accessibility as one key enabler for a more inclusive society. The European Union has ratified this convention, and therefore, accessible information is now a binding legal obligation. Most importantly, the

EU adopted the European Accessibility Act (EAA) in 2019, which requires that many everyday products and services (including digital information services) meet strict European accessibility standards. The UNCRPD and the EAA stress that information has to be provided in accessible language, such as plain language or easy-to-read (E2R) language. Only in this way can the population segments with language comprehension difficulties exercise their now recognised right to participate in society and public life. In response, the European Commission and other bodies have begun to produce “easy-to-read” versions of key documents. However, providing information in easy language in a real-world setting is challenging. High-quality E2R content must typically be written or translated by human experts following specialised guidelines, and then often validated by target readers. This manual process requires considerable effort, time, and expertise, which limits how much content can be produced or translated into easy formats. As a result, only a fraction of official communications or legal texts might ever be available in simplified form. Society faces a clear gap between the aspiration of accessible information for all and the practical reality of inaccessible content. Over the years there has been considerable research in text simplification (Saggion, 2017; Al-Thanyyan and Azmi, 2021). While early automatic text simplification suffered from an abstract definition of its goal, contemporary approaches target specific, measurable standards like plain language and Easy-to-Read text (Ginel et al., 2025).

The MER-TRANS shared task organised in the context of the IberLEF 2026 evaluation forum (Bonet-Jover, González-Barba, and Chiruzzo, 2026) aims at advancing the state of the art in the field of easy-to-read translation with a specific emphasis on multilinguality, addressing two Romance languages from the Iberian Peninsula: Catalan and Spanish, as well as Italian, while Arabic was introduced as a surprise language to encourage flexible, true multilingual approaches.

2 Task Description

In our proposed task, we challenge participants to automatically produce easy-to-read versions of texts and/or sentences from a recently developed corpus. The texts are sam-

pled from the iDEM corpus, a multilingual dataset in the domain of democratic participation created in the context of a European project (Saggion et al., 2024), which has been simplified by human experts following easy-to-read recommendations and high-quality standard procedures (Bott et al., 2026) a version of which is released for this shared task. The **surprise track** released Arabic text simplification resources to the community (Khallaf and Sharoff, 2021). The task, which features a dedicated Web page¹ was publicized via mailing list and social networks and requested participants to submit up to three runs per language for evaluation. As a result of our call, thirteen teams registered for the task, seven of which ultimately submitted their systems’ outputs.

3 Relevance and Novelty

Access to information in easy-to-read formats is relevant for many communities, including people with intellectual disabilities, people with low literacy, or more generally, people who struggle with language complexity. This is the first task that addresses multilingual easy-to-read adaptations. Previous related tasks were in the area of lexical simplification (Shardlow et al., 2024; Saggion et al., 2022) (e.g. TSAR 2022, BEA 2024) or targeted CEFR simplification in English (Alva-Manchego et al., 2025) (e.g. TSAR 2025) or easy-to-read or plain language adaptations in Spanish (Botella Gil et al., 2025) (e.g. CLEARS 2025). Moreover, this task addresses two Iberian languages: Catalan and Spanish, with Catalan being an under-resourced language where text simplification is of concern. In addition, the task also targets Italian with Arabic included as a surprise language.

4 Easy Language

Plain language is defined as “a communication that puts readers first. It considers: (1) what readers want and need to know; (2) readers’ level of interest, expertise, and literacy skills; and (3) the context in which readers will use the document” (ISO, 2023). This methodology has the aim of ensuring readers can find what they need, understand it, and use it.

¹<https://lastus-taln-upf.github.io/mertrans-iberlef-2026/>

Plain language acts as an umbrella term under which various levels of text simplification and adaptation fall. Easy-to-read (E2R) is the methodology specifically designed for people with reading comprehension difficulties. In Spain, we have a UNE standard that sets out specific guidelines for writing and designing in easy-to-read Spanish (AENOR, 2018). This standard is also used as a reference for Catalan, with the necessary adjustments to account for the specific characteristics of that language.

The strategies used in easy-to-read texts make it possible to reduce the complexity of the texts whilst retaining their essential information. To this end, strategies such as lexical substitution, syntactic restructuring, and occasional explanation or clarification, amongst others, are employed.

Furthermore, a key element of this UNE standard is the requirement to test and validate the comprehension of easy-to-read texts with users who have comprehension difficulties. Given the specific context of this project, the corpus available for this task has not yet been validated by final users.

5 Shared Task Data

The task features two datasets which are described below.

5.1 The iDEM Corpus

The main dataset for the shared task was the iDEM corpus (Bott et al., 2026), which covers Italian, Catalan, and Spanish. Table 1 shows the size of the language-specific parts in sentences and words.

Language	Sentences Segments		Words		Words per Sentence Segment	
	Orig	Simp	Orig	Simp	Orig	Simp
Spanish	354	1290	11665	10883	32.95	8.44
Catalan	380	405	11570	14279	41.32	35.26
Italian	325	718	10398	12230	31.99	17.03

Table 1: Distribution of number of sentence segments and words, and the word per sentence segment ratio for each language.

The corpus contains texts from the domain of public deliberation from different sources, like informative texts from institutions, political articles, and news articles, among others. In terms of content, it covers social policy, justice and services, health

policy, accessibility, governance and similar topics.

The simplified versions of the texts were produced by professional E2R translators, on the basis of variants of three language-specific versions of the E2R recommendations. While translating, translators annotated which E2R recommendations they were applying. The recommendations varied somewhat across languages, but targeted the same linguistic phenomena. Tables 2 to 4, show the frequencies of simplification operations. The most common ones across languages are the use of *frequent and familiar words* (verbs, nouns, adjectives, etc), *separation of ideas* into separate sentences, lines or clauses, and the use of *simple syntax*.

For the shared task, the corpus is distributed as CSV files (one per language) where each line contains an original sentence and its simplification. The document identifier and sentence number are also provided to facilitate and encourage context-aware simplification experiments. A JSON version of the corpus with embedded simplification criteria will be published at a later stage.

Simplification Criteria	Count	%
Use of paragraph breaks or line breaks to separate ideas into different sentences	241	22.8%
Use of frequent/common verbs	116	11.0%
Use of short, simple, and commonly used nouns	111	10.5%
Reordering of sentences into logical structure: subject + verb + predicate	84	7.9%
Elimination or substitution of abstract terms, foreign words, vague content, and/or technical jargon	79	7.5%
Use of indicative mood and simple tenses (present, simple future, simple past, and imperfect)	48	4.5%
Definition of technical or complex terms necessary for the document	45	4.3%
Addition of subject to avoid ellipsis	41	3.9%

Table 2: Most frequently used annotation tags: Spanish.

For each language, the trial data consisted of one complete document containing original segments of varying lengths together with their Easy-to-Read reference simplifications. The trial documents contained 4 Spanish, 7 Catalan, and 8 Italian original segments. The remaining data formed the official test set, comprising 350 Spanish, 373 Catalan, and 317 Italian original segments.

Simplification Criteria	Count	%
Choose words that are easily understandable and familiar to the target group.	261	55.30%
Only use common and frequent adverbs. Avoid adverbs ending in -ment.	48	10.17%
When listing elements related to a single idea in one sentence, use commas to separate them.	34	7.20%
Write quantities and figures using numbers instead of words.	22	4.66%
Create a glossary or include glosses in the document if there are terms that are difficult to understand.	22	4.66%
The most common punctuation marks may be used: period (.), comma (,), question mark (?), exclamation mark (!), and the at symbol (@).	19	4.03%
Prioritize simple syntax. Use the neutral order of the elements of the sentence. Prefer short sentences	15	3.18%
Avoid using percentages when addressing people with significant comprehension difficulties	10	2.12%

Table 3: Most frequently used annotation tags: Catalan.

Simplification Criteria	Count	%
Use of commonly used verbs	292	13.34%
Use of periods or commas to separate sentences with multiple ideas	271	12.38%
Elimination or modification of abstract terms, foreignisms, vague content, and/or technical jargon	264	12.06%
Use of short, simple, and commonly used nouns	255	11.65%
Use of commonly used adjectives	163	7.45%
Use of commonly used adverbs	143	6.53%
Avoid use of semicolons (;), ellipses (...), parentheses, quotation marks, and other unusual punctuation	77	3.52%

Table 4: Most frequently used annotation tags: Italian.

5.2 Arabic Dataset

We use a subset of 398 aligned sentence pairs from the *Saaq al-Bambuu* Arabic simplification corpus introduced by (Khallaf and Sharoff, 2021). The original corpus was created by aligning sentences from Saud Al-Sanousi’s novel *Saaq al-Bambuu* with its authorised simplified edition for learners of Arabic as a second language, resulting in 2,980 complex-simple sentence pairs. The subset used in this study comprises 20 trial pairs with reference simplifications and 378 test pairs. The trial examples were manually selected to represent a range of sentence structures and simplification operations, providing

Measure	Mean	Q1	Med.	Q3	IQR
Orig. words	18.33	10.00	15.00	23.00	13.00
Simpl. words	13.69	7.50	12.00	17.00	9.50
Word red.	4.64	0.00	2.00	7.00	7.00
Red. (%)	20.73	0.00	16.67	35.71	35.71

Table 5: Length statistics for the 398 Arabic sentence pairs. Q1, Q3, and IQR denote the first quartile, third quartile, and interquartile range, respectively.

demonstrations of the expected simplification behaviour for the unseen test instances.

The source sentences frequently contain subordinate clauses, coordination, descriptive expressions, named entities, and literary narrative structures. To characterise the transformations applied in the reference simplifications, we use the existing annotations based on the simplification ontology introduced in (Khallaf, Eugeni, and Sharoff, 2025). The ontology captures operations including compression, omission, lexical substitution, syntactic change, modulation, transcription, transposition, and explanation. Since the annotation scheme is multi-label, an individual sentence pair may contain several simplification operations simultaneously. Compression and omission are the most frequent strategies in the selected subset, followed by lexical substitution, although the data also includes syntactic restructuring, semantic re-expression, minimal-change transcription, clarification, and explanation.

Table 5 presents the sentence-length statistics for the aligned subset. The mean length decreases from 18.33 words in the original sentences to 13.69 words in the simplified references, while the median decreases from 15 to 12 words. The interquartile range also declines from 13 to 9.5 words, indicating that the simplified sentences are generally shorter and less variable in length. Of the 398 pairs, 295 simplified sentences are shorter than their original counterparts, 76 retain the same number of words, and 27 become longer. The longer simplifications typically involve explanation, clarification, or explicit reference resolution.

6 Baselines

In order to provide comparison points, three baselines were established. A *DoNothing* baseline produces no simplification, copying the input sentence. The two other baselines employ current off-the-shelf LLM-based prompting approaches.

BASELINE1 results were computed using the Large Language Model (LLM) Llama version 3, with 8 billion parameters (Meta AI, 2024), deployed in a local environment with 8 GB of RAM and a model size of 4.34 GB. The baseline functionality has been implemented in an API server whose code is publicly available². To carry out the simplification task, the baseline undertakes the following steps:

- A language identifier process is carried out to identify the language of the input text.
- The received input is being chunked into one or more paragraphs, which are processed separately to avoid overflowing the LLM.
- A simple prompt is prepared based on the following sentence “Please simplify the following sentence:” adapted to Catalan, Spanish, Italian, or Arabic, depending on the identified language.
- Finally, the Llama model is being invoked with prompt and text chunk to produce a single simplification.

BASELINE2 was produced by prompting the instruction-tuned version of the SALAMANDRA (Gonzalez-Agirre et al., 2025) model with 2 billion parameters, accessed via the HuggingFace transformer library (Barcelona Supercomputing Center, 2024), as a very small model, with prompt 2 taken from the BLESS benchmark (Kew et al., 2023), which was only slightly modified to make the system produce only one output solution. The prompt was translated into each of the languages and used for the runs produced in those languages. The English version of the prompt is as follows:

Please rewrite the following complex sentence to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning. Return one rephrasing. Do not generate any text except the reformulation.

²<https://github.com/danielAdanzaDopazo/BaselineIberlef>

7 Participants

In the following subsections, we provide an overview of the approaches proposed by the seven teams.

7.1 HumanAI-UCM

The HumanAI-UCM team (Broto-Ortega, Francisco, and Hervás, 2026) prepared three submissions for the Spanish language. They based their systems on LLM prompting, relying on Llama, Phi-4 Mini, and Mistral models, assessing various prompting strategies over available data to identify the most suitable configurations. More specifically, they experimented with zero-shot, one-shot, and few-shot configurations utilizing prompts derived from the Easy-to-Read UNE norm (AENOR, 2018) and tested with the corpus-based simplification criteria reported in the MER-TRANS Corpus (Bott et al., 2026). The impact of incorporating contextual information via a context window mechanism was also assessed.

7.2 VICOMTECH

The VICOMTECH team (Calleja, Ponce, and Etchegoyhen, 2026) presented a comprehensive approach to multilingual text adaptation into Easy-Read (ER) format for Arabic, Catalan, Italian, and Spanish. To achieve this, they contrasted three distinct modelling paradigms: a foundational few-shot LLM approach, an ER fine-tuned LLM approach, and an iterative post-editing framework guided by a learned evaluation metric adjusted specifically to the task. Across all configurations, the team heavily utilised the medium-sized Llama-3 (8B) model family.

7.3 ClearText

The ClearText approach (Estevanell-Valladares et al., 2026) used an agentive workflow with three specialised agents that were dedicated to different groups of rules from the UNE 153101 EX Easy-to-Read standard (AENOR, 2018). They applied orthographic, lexical and syntactic rules in sequence and in this order. This follows the intuition that the three agents represent an expert in each field and the three linguistic levels are separated from each other. Each processing stage was completed before its output was passed to the next agent. The team submitted one run using Qwen3.6-27B.

7.4 FACILE

The FACILE team (Diab-Lozano et al., 2026) participated in the Spanish track of MER-TRANS; the authors present an automatic text-simplification approach utilizing a Qwen3.5-9B model trained via a two-stage fine-tuning pipeline. In the first stage, the model is fine-tuned on a binary style-classification task to accurately recognize Easy-to-Read (E2R) writing styles. To support this classification phase, the team compiled a substantial corpus of Spanish texts labelled as either "easy" or "difficult", from which a stratified subset was used. In the second stage, the model is fine-tuned to perform the actual automatic adaptation, utilizing a parallel corpus of over 3,000 original-E2R sentence pairs assembled from two complementary resources.

7.5 HULAT1

The HULAT1 approach (Martín Ybañez and Segura-Bedmar, 2026) relied on the Llama 3.3 70B Versatile model to perform prompt-based simplification for Italian and Spanish since it effectively preserved the original content while strictly producing the requested simplification. They tested four distinct prompt configurations by combining two types of instructions, simple and advanced, with both zero-shot and few-shot strategies. The simple prompt instructed the model to assume the role of a text simplification expert. Meanwhile, the advanced prompt extended this basic instruction by incorporating a set of explicit simplification rules selected and adapted from the UNE 153101:2018 EX standard (AENOR, 2018). For the few-shot variants, the configurations incorporated three language-specific examples from the MER-TRANS trial data. To better select the final submissions, the authors assessed their configurations on an augmented dataset produced via back-translation. Few-shot prompting consistently outperformed zero-shot prompting, whereas the additional UNE-based rules produced only marginal improvements.

7.6 NIL-UCM

Building on a previous method used in CLEARs 2025 (Botella Gil et al., 2025), the NIL-UCM team (Díaz and Fernández, 2026) used an LLM (Mistral-7B-Instruct-v0.3) with explicit linguistic instructions based on the

easy-to-read UNE 153101 EX basis standard (AENOR, 2018) to produce a system for Spanish. The prompt focused on aspects related to orthotypography, lexis, morphosyntax, style, and information organization. They also tested a less restrictive prompt which they argue may favour better language adaptation in the MER-TRANS context.

7.7 HULAT2

For the Spanish track, the HULAT2 team (Moreno et al., 2026) submitted three runs leveraging distinct architectural strategies. Two (RUN1/RUN2) were generated using the same LangGraph-based multi-agent workflow that combines Gemini 2.5 Flash and RigoChat-7B-v2—a Spanish-oriented model based on Qwen2.5-7B-Instruct tailored for Spanish queries. Specific features within the workflow are used to dynamically determine whether a text requires conservative rewriting, syntactic simplification, lexical support, or stronger semantic control. One of the workflows (RUN3) integrated specialized lexical resources, drawing from a detailed list of public resources providing a glossary of learning words and short explanatory forms. In contrast, the third run served as a linear generate-evaluate-regenerate baseline built entirely around the RigoChat model, utilizing advanced prompt engineering and fine-tuned using LoRA on the original Spanish Constitution and its E2R version.

8 Metric-based Evaluation

The official evaluation was conducted with automatic metrics commonly used in text simplification. In particular, our evaluation code relies on EASSE (Alva-Manchego et al., 2019), a Python package designed to facilitate and standardise the automatic evaluation of sentence simplification systems. EASSE provides implementations of standard simplification metrics, including SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002), and has been widely used in previous text simplification research.

We report four metrics: BLEU-Orig, BLEU-Gold, SARI, and BERTScore. BLEU is computed at the corpus level for all languages, while SARI is computed at the corpus level for Catalan, Italian, and Spanish and as the average of sentence-level scores for Arabic. BERTScore is reported as the

mean of the sentence-level F1 scores. All metrics are computed against a single easy-to-read reference per input sentence, as the iDEM corpus provides one expert adaptation for each original sentence.

BLEU-Orig measures the similarity between the system output and the original complex sentence. This score is useful for identifying conservative systems that remain very close to the input. However, a high BLEU-Orig score does not necessarily indicate successful simplification, since copying the original sentence can also result in high similarity. BLEU-Gold measures the similarity between the system output and the gold, easy-to-read reference. This metric provides an indication of how close the generated sentence is to the reference simplification at the surface level. Nevertheless, BLEU alone is limited for text simplification evaluation, because valid simplifications may differ from the reference while still being correct and accessible.

SARI is the main simplification-oriented metric used in our evaluation and the primary criterion for ranking systems within each language, with BLEU-Gold, BLEU-Orig, and BERTScore used as successive secondary ranking criteria. It compares the system output with both the original sentence and the gold reference and evaluates whether the system appropriately keeps, deletes, and adds textual units. For this reason, SARI is more suitable than BLEU for assessing whether a system has actually simplified the input rather than preserving or paraphrasing it. As only one reference is available per sentence, SARI is computed in a single-reference setting.

BERTScore (Zhang et al., 2020) was included as a complementary semantic similarity metric. Unlike BLEU, which is based on lexical overlap, BERTScore uses contextual embeddings to estimate semantic similarity between the system output and the gold reference. This allows the evaluation to capture cases where the system preserves the meaning of the reference using different wording. We compute BERTScore with the multilingual `xlm-roberta-large` model (Conneau et al., 2020) and report its F1 variant without baseline rescaling.

Because the task is multilingual, the evaluation pipeline applies language-appropriate text processing. For Catalan, Italian, and

Spanish, BLEU and SARI use EASSE’s international tokenizer. For Arabic, we apply a dedicated preprocessing step consisting of Unicode NFKC normalisation, removal of tatweel and diacritics, and word tokenisation with CAMEL Tools (Obeid et al., 2020), as well as an Arabic-specific SARI implementation operating over n -grams of order 1 to 4.

These metrics provide a multi-dimensional view of system performance. Since easy-to-read translation requires both simplification and meaning preservation, the results should be interpreted by considering all metrics jointly rather than relying on a single score.

9 Results

Tables 6, 7, 8, and 9 present the official MER-TRANS results for Arabic, Catalan, Spanish, and Italian, respectively. The tables include all participating teams, as well as the official baselines and the DoNothing setting. When a team submitted multiple runs for the same language, we report the run with the highest SARI score, since SARI is the main simplification-oriented evaluation metric. Systems were evaluated using BLEU-Orig, BLEU-Gold, SARI, and BERTScore. BLEU-Orig measures similarity with the original input, BLEU-Gold measures surface-level overlap with the reference simplifications, SARI evaluates simplification operations, and BERTScore captures semantic similarity. Higher BLEU-Gold, SARI, and BERTScore values generally indicate greater similarity to the reference or better simplification performance. BLEU-Orig is instead a diagnostic measure, a high value indicates that the output remains close to the source and may therefore signal insufficient simplification.

According to SARI, VICOMTECH achieved the best result for Arabic, while ClearText obtained the highest scores for Catalan, Spanish, and Italian. The DoNothing baseline achieved very high BLEU-Orig scores, as expected, because it preserves the original input; however, its substantially lower SARI scores show that copying the input does not perform effective simplification.

Overall, ClearText achieved the best SARI scores for Catalan, Spanish, and Italian, with scores of 48.78, 47.02, and 49.84, respectively. For Arabic, the best result

Team	Run	BLEU-O	BLEU-G	SARI	BERT
VICOMTECH	1	44.02	30.79	39.31	0.935
BASELINE1	1	13.47	10.39	35.77	0.925
BASELINE2	1	9.96	5.16	30.32	0.875
DoNothing	1	100.00	43.88	20.50	0.949

Table 6: Official results for Arabic ($n = 378$).

Team	Run	BLEU-O	BLEU-G	SARI	BERT
ClearText	1	18.70	12.95	48.78	0.921
VICOMTECH	1	15.33	9.97	44.71	0.910
BASELINE1	1	14.43	10.21	44.01	0.905
BASELINE2	1	28.04	8.70	40.57	0.912
DoNothing	1	100.00	15.10	11.96	0.924

Table 7: Official results for Catalan ($n = 373$).

was obtained by VICOMTECH, with a SARI score of 39.31. This difference may reflect the greater difficulty of the Arabic setting for the participating systems in terms of simplification-oriented evaluation.

The results also show the importance of evaluating simplification separately from meaning preservation or surface similarity. A system may obtain a high BLEU-Orig score by staying close to the original sentence. SARI is therefore especially important in this task because it rewards appropriate rewriting operations rather than simple copying.

Across languages, the best SARI scores for Catalan, Spanish, and Italian were relatively close, ranging from 47.02 to 49.84. In contrast, the best Arabic SARI score was lower, at 39.31. This difference may reflect the greater difficulty of the Arabic setting, especially because Arabic was included as a surprise language.

In summary, the official results show that MER-TRANS is a challenging, multilingual, easy-to-read translation task. The strongest systems were those that balanced simplification with semantic preservation. The contrast between BLEU-based scores and SARI confirms that easy-to-read translation should be evaluated as a simplification task, not only as a sentence generation or translation task.

9.1 Statistical System Comparison

We ran systematic statistical comparisons (paired one-sided t-tests) between the outputs of the different systems and the baselines to see if their performances were significantly different from each other. As data points, we used the SARI results of individual sentences. We were only interested in the significance of differences, not

Team	Run	BLEU-O	BLEU-G	SARI	BERT
ClearText	1	18.00	15.45	47.02	0.922
VICOMTECH	3	20.79	14.11	45.58	0.924
HULAT1	2	33.18	17.71	45.02	0.927
HULAT2	1	35.57	17.57	44.05	0.926
BASELINE1	1	18.95	10.89	43.29	0.913
HumanAI-UCM	2	26.98	11.84	41.62	0.911
FACILE	1	3.00	2.86	41.61	0.878
NIL-UCM	2	11.38	4.80	40.52	0.903
BASELINE2	1	29.09	9.39	37.18	0.902
DoNothing	1	100.00	23.22	13.87	0.930

Table 8: Official results for Spanish ($n = 350$).

Team	Run	BLEU-O	BLEU-G	SARI	BERT
ClearText	1	18.63	13.88	49.84	0.92
HULAT1	3	23.44	16.00	48.68	0.92
HULAT1	2	24.71	15.11	48.67	0.93
BASELINE1	1	16.58	10.15	46.51	0.91
VICOMTECH	1	9.04	8.01	46.48	0.91
HULAT1	1	34.65	15.59	43.98	0.92
VICOMTECH	3	30.66	8.70	41.63	0.91
BASELINE2	1	24.74	6.25	39.52	0.90
VICOMTECH	2	45.02	9.49	37.01	0.92
DoNothing	1	100.00	12.40	10.37	0.91

Table 9: Official results for Italian ($n = 317$).

in effect sizes. For Spanish, there was the highest number of participating teams and submissions. The results were correspondingly very close to each other. Table 10 presents a cross-comparison by significance level. Only the highest-scoring runs (SARI) of each participant are listed. It can be seen that the two top-scoring systems, ClearText and VICOMTECH, are so close together that the difference between them is only significant at the lowest level ($p < 0.05$), and so are the difference between rank 2 (VICOMTECH) and 3 (HULAT1). Other pairs with this low level of significance in differences are FACILE vs. HumanAI-UCM and HumanAI-UCM vs. NIL-UCM, as well as HumanAI-UCM vs. NIL-UCM. HULAT1 and HULAT2 differed at the $p < 0.01$ level and HULAT2 vs. BASELINE1 at the $p < 0.05$ level.

For the other three languages - Catalan, Italian and Arabic - the performance differences were generally larger, but for Italian, again, the two best-scoring systems, ClearText and HULAT1, were significantly different only at the $p < 0.05$ level. Results are given in tables 11, 12, and 13.

Team	ClearText	VICOMTECH	HULAT1	HULAT2	BASLINE1	HumanAI-UCM	FACILE	NIL-UCM	BASLINE2
ClearText									
VICOMTECH	*								
HULAT1	*	*							
HULAT2	***	***	***						
BASLINE1	***	***	***	*					
HumanAI-UCM	***	***	***	***	*				
FACILE	***	***	***	*	***	*			
NIL-UCM	***	***	***	***	***	*	*		
BASLINE2	***	***	***	***	***	***	***	***	
DoNothing	***	***	***	***	***	***	***	***	***

Table 10: Statistical comparison in terms of significant differences for Spanish ('-': no significance, '*': p_i0.05, '**': p_i0.01, '***': p_i0.001).

TEAM	ClearText	VICOMTECH	BASLINE1	BASLINE2
ClearText				
VICOMTECH	***			
BASLINE1	***	*		
BASLINE2	***	***	***	
DoNothing	***	***	***	***

Table 11: Statistical comparison in terms of significant differences for Catalan ('-': no significance, '*': p_i0.05, '**': p_i0.01, '***': p_i0.001).

10 Conclusion

This paper presented MER-TRANS, the first shared task on multilingual Easy-to-Read translation covering Catalan, Spanish, Italian, and Arabic. Seven teams participated and explored a diverse range of methods, including prompting, fine-tuning, multi-agent systems, lexical resources, and iterative post-editing. This diversity is an important outcome of the task, as it allows different approaches to be compared using the same data and evaluation framework, helping identify promising directions for future research.

ClearText achieved the highest SARI scores for Catalan, Spanish, and Italian using a sequential multi-agent workflow in which specialised agents applied orthographic, lexical, and syntactic Easy-to-Read rules. VICOMTECH obtained the best result for Arabic using a pretrained Llama 3.1 8B model with few-shot and incremental context processing, in which previously simplified sentences and the complete source document were provided as context. The statisti-

TEAM	ClearText	HULAT1	BASLINE1	VICOMTECH	BASLINE2
ClearText					
HULAT1	*				
BASLINE1	***	***			
VICOMTECH	***	***	*		
BASLINE2	***	***	***	***	
DoNothing	***	***	***	***	***

Table 12: Statistical comparison in terms of significant differences for Italian ('-': no significance, '*': p_i0.05, '**': p_i0.01, '***': p_i0.001).

TEAM	VICOMTECH	BASLINE1	BASLINE2	DoNothing
VICOMTECH				
BASLINE1	***			
BASLINE2	***	***		
DoNothing	***	***	***	

Table 13: Statistical comparison in terms of significant differences for Arabic ('-': no significance, '*': p_i0.05, '**': p_i0.01, '***': p_i0.001).

cal comparisons further showed that differences between neighbouring systems were not always significant, particularly among the highest-ranked Spanish and Italian submissions.

Future editions of MER-TRANS will extend the evaluation through human assessment and validation with target readers, while also considering multiple reference adaptations and document context. Expanding the task to further languages and domains will support the development of more robust and inclusive Easy-to-Read translation systems.

Acknowledgments

This document is part of two projects that have received funding from the European Union’s Horizon Europe research and innovation program under grants: Grant Agreement No. 101132431 (iDEM Project) and Grant Agreement No. 101178191 (IDEAL Project). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The University of Leeds (UOL) was funded by UK Research and In-

novation (UKRI) under the UK government’s Horizon Europe funding guarantee (Grant Agreement No. 10103529).

References

- AENOR. 2018. Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos (Norma UNE 153101:2018 EX).
- Al-Thanyyan, S. S. and A. M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2), March.
- Alva-Manchego, F., L. Martin, C. Scarton, and L. Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China, November. Association for Computational Linguistics.
- Alva-Manchego, F., R. Stodden, J. M. Imperial, A. Barayan, K. North, and H. Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In M. Shardlow, F. Alva-Manchego, K. North, R. Stodden, H. Saggion, N. Khallaf, and A. Hayakawa, editors, *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 116–130, Suzhou, China, November. Association for Computational Linguistics.
- Barcelona Supercomputing Center. 2024. Salamandra model card, bsc/salamanddra-2b. <https://huggingface.co/BSC-LT/salamandra-2b>.
- Bonet-Jover, A., J. Á. González-Barba, and L. Chiruzzo. 2026. 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.
- Botella Gil, B., I. Espinosa Zaragoza, A. Bonet Jover, M. Madina, L. Molino Piñar, P. Moreda Pozo, I. González Dios, M. T. Martín Valdivia, and L. A. Ureña López. 2025. Overview of CLEARs at IberLEF 2025: Challenge for Plain Language and Easy-to-Read Adaptation for Spanish texts. *Procesamiento del lenguaje natural*, 75:393–400.
- Bott, S., V. Riegler, H. Saggion, A. Rascón-Alcaina, and N. Khallaf. 2026. A multilingual human annotated corpus of original and easy-to-read texts to support access to democratic participatory processes. In *Proceedings of the Language Resources and Evaluation Conference 2026 (LREC 2026)*, Palma, Mallorca, Spain.
- Broto-Ortega, J., V. Francisco, and R. Hervás. 2026. HumanAI-UCM at MER-TRANS 2026: Translating Easy-to-Read Standards into Prompting Strategies for Spanish Text Simplification. 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.
- Calleja, J., D. Ponce, and T. Etchegoyhen. 2026. Vicomtech at MER-TRANS 2026: Incremental Easy Read Adaptation with Fine-tuned Models, Post-Editing Cycles and Learned Metrics. 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.
- Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.
- Diab-Lozano, I., J. Manobanda-Tutasig, C. Torres-Celorio, J. Rubira-Herráez, and M. C. Suárez-Figueroa. 2026. FACILE Team at MER-TRANS 2026: Easy-to-Read Adaptation in Spanish through Two-Stage Fine-Tuning of Large Language Models. 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.
- Díaz, A. and D. Fernández. 2026. NIL-UCM at MER-TRANS2026: Integrating Linguistic Knowledge into Prompting Strategies for Spanish Text Simplification. 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.
- Estevanell-Valladares, E., B. Botella Gil, A. Bonet Jover, I. Espinosa Zaragoza, J. Abreu Salas, P. Moreda Pozo, and A. Montoyo Guijarro. 2026. ClearText at MER-TRANS 2026: Multi-Agent Easy-to-Read Text Generation. 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.
- Ginel, M. I. R., P. Cadwell, P. Canavese, S. Hansen-Schirra, M. Kappus, A. Matamala, and W. Noonan, editors. 2025. *Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL)*, Geneva, Switzerland, June. European Association for Machine Translation.
- Gonzalez-Agirre, A., M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, and M. Villegas. 2025. Salamandra technical report.
- ISO. 2023. Plain language. Part 1: Governing principles and guidelines (ISO 24495-1:2023).
- Kew, T., A. Chi, L. Vázquez-Rodríguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, and M. Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. *arXiv preprint arXiv:2310.15773*.
- Khallaf, N., C. Eugeni, and S. Sharoff. 2025. Reading between the lines: A dataset and a study on why some texts are tougher than others. In M. Zock, K. Inui, and Z. Yuan, editors, *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, pages 24–34, Abu Dhabi, UAE, January. International Committee on Computational Linguistics.
- Khallaf, N. and S. Sharoff. 2021. Automatic difficulty classification of arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual), apr. Association for Computational Linguistics.
- Martín Ybañez, L. and I. Segura-Bedmar. 2026. HULAT1 at MER-TRANS 2026: UNE-Guided Prompting for Multilingual Easy-to-Read Text Simplification. 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.
- Meta AI. 2024. Meta-Llama-3-8B. <https://huggingface.co/meta-llama/Meta-Llama-3-8B>. Accessed 4 June 2026.
- Moreno, L., P. Martínez, M. A. Sanchez-Escudero, and M. Domínguez-Gómez. 2026. HULAT2 at MER-TRANS 2026: Governed Multi-Agent Simplification for Spanish Easy-to-Read Generation. 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.
- Obeid, O., N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the twelfth language resources and evaluation conference*, pages 7022–7032.
- Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

- Pennsylvania, USA, July. Association for Computational Linguistics.
- Saggion, H. 2017. *Automatic Text Simplification*, volume 10 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- Saggion, H., J. O’Flaherty, T. Blanchet, S. Sharoff, S. Sanfilippo, L. Muñoz, M. Gollegger, A. Rascón, J. L. Martí, S. Szasz, S. Bott, and V. Sayman. 2024. Making democratic deliberation and participation more accessible: The idem project. In *SEPLN (Projects and Demonstrations)*, volume 3729 of *CEUR Workshop Proceedings*, pages 71–76. CEUR-WS.org.
- Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, and W. Xu, editors, *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual), December. Association for Computational Linguistics.
- Shardlow, M., F. Alva-Manchego, R. Batista-Navarro, S. Bott, S. Calderon Ramirez, R. Cardon, T. François, A. Hayakawa, A. Horbach, A. Hülsing, Y. Ide, J. M. Imperial, A. Nohejl, K. North, L. Occhipinti, N. P. Rojas, N. Raihan, T. Ranasinghe, M. S. Salazar, S. Štajner, M. Zampieri, and H. Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico, June. Association for Computational Linguistics.
- Xu, W., C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.