Vicomtech at MER-TRANS 2026: Incremental Easy Read Adaptation with Fine-tuned Models, Post-Editing Cycles and Learned Metrics

Jesús Calleja[1,2,†], David Ponce[1,2,†] and Thierry Etchegoyhen[1,*,†]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia – San Sebastián, Spain
[2]University of the Basque Country - UPV/EHU, Donostia – San Sebastián, Spain

Abstract

We describe Vicomtech's participation in the MER-TRANS challenge on multilingual text adaptation to Easy Read in Arabic, Catalan, Italian and Spanish. We explored three modelling variants for each of the four languages in the shared task. Our first approach leverages a pretrained LLM with a structured prompt, incremental context processing, and few shot demonstrations sampled from the trial training data. The second variant extends the first one with a model fine-tuned on Easy Read corpora. Finally, our third variant builds upon the second variant, adding post-editing cycles guided by learned metrics. We achieved robust results overall and present a detailed analysis of our findings regarding the comparative strengths and limitations of the selected approaches.

Keywords

MER-TRANS, Easy Read, Multilingual Text Adaptation, Large Language Models

1. Introduction

In this work, we describe Vicomtech's participation in the MER-TRANS shared task [1] on multilingual text adaptation to Easy Read (ER), organised within IberLEF 2026 [2]. ER adaptation involves multiple factors, such as simple vocabulary and grammar, short sentence length, information selection, natural sentence segmentation, and concept explanations as necessary, among other criteria. The challenge centred on one of the core aspects of ER adaptation, namely text simplification, where participants were tasked to adapt text fragments to ER content in Catalan, Spanish, Italian and Arabic. System outputs were to be evaluated on a range of metrics that included BLEU [3], SARI [4] and BERTScore [5].

Our participation aimed to address the multilingual ER adaptation task across different dimensions. We contrasted different types of modelling, exploring few-shot adaptation with Large Language Models (LLM), model fine-tuning, and post-editing cycles [6] with a learned metric adjusted to the task. Our aim in this case was to assess the relative adequacy of methods with varying degrees of complexity across languages. Additionally, since ER resources are scarce overall, with only a limited set of publicly available corpora for Spanish [7, 8, 9] among the languages of the task, as of this writing, we explored the viability of models fine-tuned on machine-translated Spanish ER data. If effective, this strategy could partially alleviate the need for ER datasets that are costly to produce.

We submitted three different variants for all four languages of the task, achieving strong results over the strong baselines in Spanish, Catalan and Arabic, with relatively lower performance in Italian. We describe in more detail our approaches and results in the remainder of this paper.

2. Related Work

Text adaptation to Easy Read, also referred to as Easy-to-read, is one of the main recognised means to provide accessible content for persons with intellectual and developmental disabilities. ER text

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
jcalleja@vicomtech.org (J. Calleja); adponce@vicomtech.org (D. Ponce); tetchegoyhen@vicomtech.org (T. Etchegoyhen)

adaptation is subjected to multiple constraints aiming at enhancing readability for the target audience, including simpler vocabulary and grammar, short sentences and segmented lines as needed, formatting constraints with larger fonts and line spacing, explanation of concepts as necessary, and presenting the information in clearly separated blocks focusing on the core ideas of the original text. ER is regulated by specific guidelines[1], and a standardised norm (UNE 153101:2018 EX [10]) for Spanish

The number of currently available ER corpora is fairly limited at present. One of the first such corpora, the LäsBarT corpus, was built from varied sources that included news, fiction and community information in Swedish [11]. For that same language, Rennes and Jönsson [12] prepared an ER corpus from public administration websites. In Dutch, the Wablieft corpus was constructed from an ER newspaper source [13].

For the languages of the shared task, to our knowledge only Spanish ER resources were available at the onset of the task, although a multilingual corpus will be released in the near future for all four languages [14]. Other resources have been built for Spanish in recent years. The IREKIER corpus [9], which also includes Basque, was prepared from the Irekia open data portal of the Basque Government; this source was also used to develop the IrekiaLF, a smaller dataset in Spanish for ER evaluation [15]. The ClearSim corpus [7] is a large corpus of 15K texts automatically adapted to ER and Plain Language using ChatGPT, with human validation, post-editing, or both. Finally, an ER corpus was provided within the CLEARS shared task [8], collected from Spanish municipalities and covering different topics such as sports, culture, leisure, and festivities.

Most current approaches to ER adaptation follow an end-to-end methodology exploiting LLMs with different types of prompt engineering, based on ER guidelines, model fine-tuning [16, 17], or both. This was the case for instance for most participants in the aforementioned Clears shared task: San-Martín et al. [18] used zero-shot prompting with a Gemma 3 model [19]; Ayesh et al. [20] also explored adaptation prompt variants and few-shot demonstrations, comparing Llama 3.2 and Gemma 3 models; Fernández and Díaz [21] leveraged Mistral-7B-Instruct-v0.3 model with a prompt that included guidelines and examples; Plaza et al. [22] used a Doplhin Mistral model under a zero-shot approach and prompts based on the specifications of the UNE standard for Spanish.

Although also relying on LLMs and structured few-shot prompting for ER adaptation, Calleja et al. [6] proposed a novel approach based on automatic refinement of ER hypotheses via post-editing cycles, a method they referred to as APEC. We exploit this specific approach in our experiments.

Finally, as previously noted, ER adaptation involves a large number of aspects beyond text simplification. Specific aspects such as line segmentation into grammatical constituents have notably been studied by Calleja et al. [23]; since this specific component of ER was not included in the shared task, we left it aside in our experiments.

3. Approach

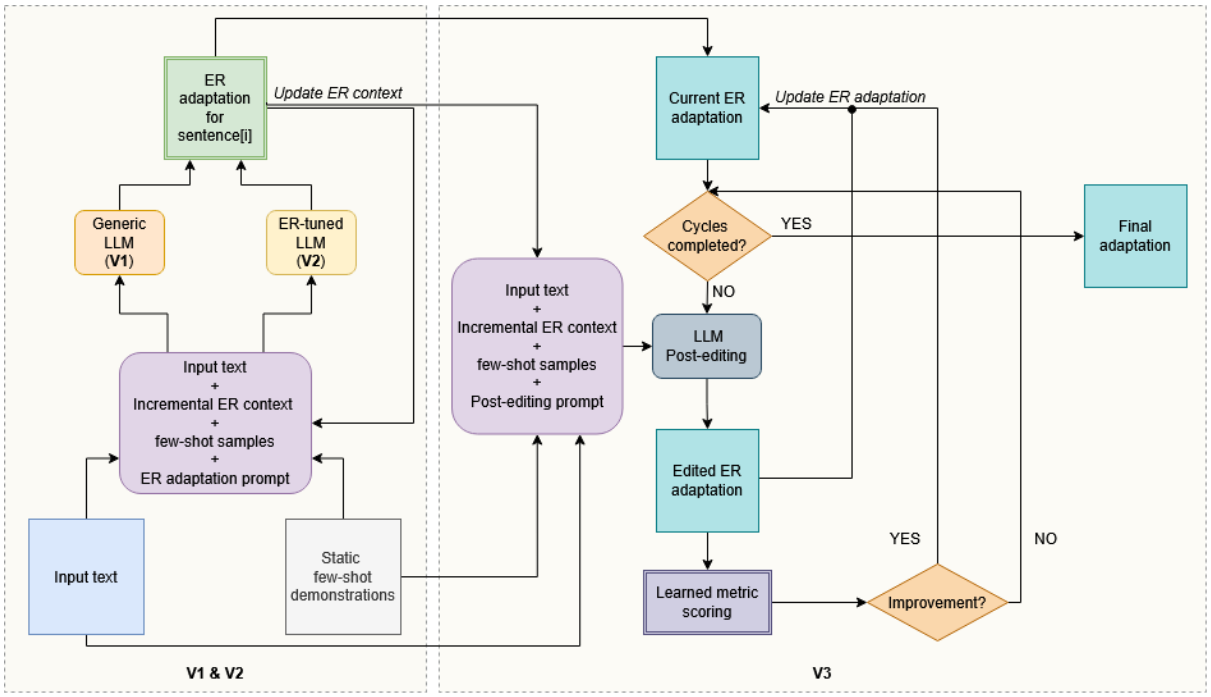
As previously indicated, we explored three different types of systems for ER adaptation, illustrated in Figure 1 and described below in more detail.

Variant 1: Few-shot Incremental Adaptation. Our first variant (hereafter, V1) was a few-shot LLM-based approach that incrementally updated its context as the input was processed. For each input sentence to be adapted, we provide as context the complete input document, all sentences ER-adapted so far, and all pairs of input and corresponding ER adaptation in the provided trial data. Including the adapted text up to the current point was necessary to circumvent sentence-level limitations and provide some form of global continuity to the adaptation process.

For V1, we used a pretrained medium-sized LLM of the Llama 3 family with 8B parameters (see section 4), as it provided robust results in our own preliminary experiments in multilingual ER adaptation; we did not consider alternative models in this work. The prompt for this variant is provided in its

[1] https://www.inclusion-europe.eu/easy-to-read/

Figure 1: Overview of our three adaptation variants.



generic English form in Appendix A; a localised version of the prompt was used for each specific language.

Variant 2: Leveraging Fine-tuned Models. Our second variant (V2) builds upon V1, maintaining its incremental context update and few-shot demonstrations, but replacing the pretrained generic LLM with a model fine-tuned on ER data. In this case, the model was tuned with LoRA [24] on pairs of complex and ER-adapted documents. For Spanish, we compiled a collection of professionally created ER data; for the other languages, we machine-translated all Spanish data with a Gemma 3 model. More details are provided in Section 4.

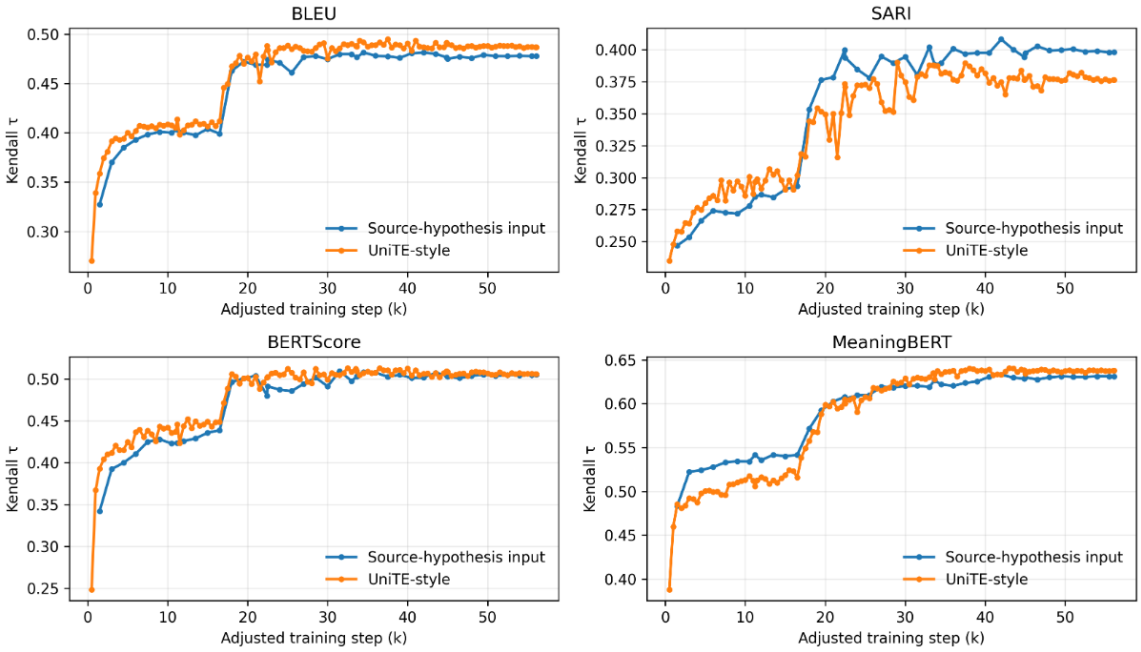
Variant 3: Learned Metric & Post-editing Cycles. For our third variant (V3), we opted for a more sophisticated approach. We employed the same core components as in V2, namely incremental context updates and fine-tuned models, complementing them with the APEC cyclic post-editing approach of Calleja et al. [6] and a learned evaluation metric.

APEC is based on iteratively refining ER adaptations according to internal metrics results. A fixed number of cycles is predefined, 5 in the original work, where the latest version of ER adaptation is either retained for the next cycle if it improved on the selected reference-free metrics, or discarded if it does not. The best adaptation according to the metrics is thus selected as the final hypothesis once all cycles have completed. One of the main reasons for undergoing a fixed number of cycles is to provide alternative generation paths via positive temperature.

In the original APEC work, the internal metrics were BERTScore, aiming to preserve the information between the source text and the adaptation as much as possible, and the Fernandez Huerta Readability Index, as a reference-free measure of readability. For the current task, we used a learned metric derived from the metrics indicated as candidates for system evaluation in the task: BLEU, SARI, BERTScore and MeaningBERT.[2]

[2]MeaningBERT, although initially mentioned among the candidate metrics, was not part of the official evaluation results. We maintained it nonetheless as we had included it in our metrics training set.

Figure 2: Evolution of Kendall τ correlation for each selected metric, contrasting simple concatenation of adapted and original sentences against Unite-style mixed training.



Our learned metric aimed to produce a single real-valued score from all selected metrics, given an input consisting of an original sentence and its corresponding ER adaptation. We used multilingual modern BERT [25] to encode the input, which consists in the concatenation of a system-generated adaptation and the original sentence, separated by a [SEP] token. The model jointly predicts the scores for all four aforementioned metrics, where the target scores for each metric are precomputed over the reference adaptations. The final score is taken as the mean of the four scores, after performing z-score normalisation for each individual score on the basis of the mean and standard deviation of each metric on the training set. We used Mean Squared Error as our loss function in all cases, and the best checkpoint was determined with Kendall τ.

Sentence representations are obtained through masked mean pooling over the encoder hidden states and passed to independent MLP regression heads for each target metric. We also explored a UniTE-style training formulation [26], where the model is trained with multiple input combinations involving the source, the system hypothesis, and optionally the reference adaptation. However, as shown in Figure 2, this strategy did not consistently improve performance over the simpler source-hypothesis formulation, while also requiring substantially more training time and computational cost. Consequently, we selected the simpler hypothesis+source input concatenation for our final system, due to its comparatively optimal efficiency-performance trade-off.

Although our learned metric is not *per se* an ER metric that takes into account all aspects of ER adaptation, it allowed us to test the APEC approach with a reference-free metric derived from the individual metrics selected by the organisers for the shared task.

4. Experimental Setup

Corpora. The core statistic for the Spanish corpora we used in the task are shown in Table 1. We used four separate corpora, available for research purposes. Two of them have been mentioned in Section 2: Irekier, consisting of professionally adapted texts from the Irekia open portal of the Basque Government, and Clears, the ChatGPT-generated and revised ER adaptations of the Clears challenge. We also aligned pairs of complex and ER-adapted documents from two additional sources: Web news from Plena Inclusión, one of the leading associations for the inclusion of persons with intellectual

Table 1
Corpus statistics for the Irekier, Plena Inclusión, Administración Fácil, and Clears datasets.

	Samples					Avg words/sample	Avg words/sample
Corpus	Train	Dev	Test	Total	Documents	(complex)	(adapted)
Irekier	3,479	170	201	3,850	444	33.80	36.41
Plena Inclusión	994	65	64	1,123	279	27.69	21.21
Administración Fácil	472	50	45	567	16	13.89	16.92
Clears	6,465	67	76	6,608	1,248	27.01	20.49
Total	11,410	352	386	12,148	1,987	28.61	25.44

disabilities in Spain[3], and the adapted texts from the Spanish Administration available on their website[4]. For these two additional sources, document alignment was performed by exploiting hyperlinks, with further paragraph-level alignment computed with CATS [27].

To be able to fine-tune the selected LLMs on ER corpora beyond Spanish, we machine-translated all Spanish corpora into the remaining three languages: Arabic, Catalan and Italian. We used the Gemma 3 12B Instruct model for the task, and five few-shot demonstrations generated with ChatGPT for each specific language.[5]

The training data for the metric were generated by first splitting both source and adapted training data at the sentence level, using the Moses sentence splitter [28]. Sentences were aligned with CATS [27], using ngram similarity and filtering all aligned sentences under a threshold of 0.3. The aligned references were used for our experiments with the Unite-like approach. To generate variants of ER adaptation, we used the adaptation prompt described in Appendix A in its localised version, and generated 5 adaptation samples from each complex source sentence by providing 3 few-shot demonstrations, performing inference with Llama 3.1 8B with a temperature of 1.0, top_p of 0.95, and deduplicating the generated adaptations. For languages other than English, the prompts were machine-translated under the same conditions described above.

Finally, for all languages, we used the trial data provided by the organisers [14, 29] as few-shot demonstrations. There were 21, 8, 7 and 4 pairs for Arabic, Catalan, Italian, and Spanish, respectively.

Models. We used Llama 3.1 Instruct 8B for all adaptation tasks [30]. All our experiments were performed on a single NVIDIA L40 GPU with 48GB vRAM.

For fine-tuning we used LoRA with 8 bit quantisation, r=16, alpha=32, dropout=0.05, max length of 5200, batch size of 1, gradient accumulation step of 16, and a learning rate of 2e-4. The model was evaluated every 50 steps, and the best checkpoint according to perplexity was stored after 3 epochs.

For inference, we used vLLM v0.19.1, with a temperature of 0.3, top_p of 1.0 and a repetition penalty of 1.0. All other parameters were set to default values.

We used a two layer MLP for our learned metric, of dimension 3072 for the first layer and 1024 for the second layer. The encoder parameters were frozen for stabilization during the first 30% of training, and later unfrozen for full fine-tuning. The model was trained with a multi-task mean squared error objective over the predicted metric scores. We trained for 5 epochs with a batch size of 32, evaluated with a batch size of 16, and used a maximum sequence length of 1024.

Due to a lack of time to complete the training of our learned metric for all languages, we resorted to using the model trained on Spanish data in all cases. Although not an ideal scenario, this approach allowed us to measure the portability of the metric when trained on a single language.

Finally, we used a fixed number of 5 cycles for the APEC approach.

[3]https://www.plenainclusion.org/
[4]https://administracion.gob.es/pag_Home/en/Lectura-Facil.html
[5]See Appendix B for details.

5. Results

5.1. Internal Results

Table 2
Results on our test set for our three variants. We indicate in bold the best results on each metric amongst the three variants, and the second-best in italics. *Avg* indicates the average of all three metrics.

Approach	Lang	BLEU	SARI	BERTScore	Avg
V1	ar	*10.87*	35.55	89.36	45.26
	ca	13.03	45.17	89.68	49.29
	es	11.70	45.26	90.11	49.02
	it	7.65	42.25	89.16	46.35
V2	ar	**11.72**	*42.31*	**90.47**	**48.17**
	ca	**17.66**	*48.80*	**91.20**	**52.55**
	es	**19.48**	*51.61*	**91.99**	**54.36**
	it	**16.91**	*46.35*	**90.88**	**51.38**
V3	ar	9.33	**44.11**	*90.28*	*47.91*
	ca	*16.58*	**48.98**	*90.76*	*52.11*
	es	*17.56*	**52.28**	*91.61*	*53.82*
	it	*14.79*	**46.59**	*90.56*	*50.65*

We first evaluated our approaches on the development and test partitions of our own data, indicated in Table 1, with the results shown in Table 2. Note that BLEU scores are those computed against the reference adaptation, not against the source text.[6] BERTScore values were ported to a [0-100] range and we computed the average of all scores to gain a general view of system results across metrics.

The first notable result is that the simplest approach, V1, was outperformed across the board by the other two variants, achieving significantly lower scores on all metrics except in the single case of BLEU in Arabic, where it ranked second, although all models performed rather poorly overall in this case. Given that V2 only differs from V1 by the use of a model fine-tuned on ER data, with that same model used in V3 as well, this shows that, on data sampled from similar sources, fine-tuning can provide large gains across metrics.

A second observation relates to the differences between V2 and V3. The use of APEC, while consistently ranked a close second compared to V2, achieved the best SARI scores in all cases but Arabic. From these results, it is rather difficult to extract robust conclusions regarding the best approach, as each metric favours a specific aspect in the evaluation, and features specific limitations: BLEU will favour adaptations that are lexically close to the reference, although multiple adaptations might be equally correct; SARI also tends to exhibit a similar strong dependency to the specific choices made in the reference adaptation, which can be detrimental to alternative valid adaptation hypotheses; finally, BERTScore results, without further normalisation, tend to gravitate around similar results which are hard to interpret beyond the fact that all models seem to preserve semantic information when measured against the references.

Finally, when comparing results across languages, consistently higher scores were achieved across metrics in Spanish, perhaps unsurprisingly considering that this was the only language for which training, development and test data were original and not translations. Despite having the data machine-translated in all other languages, fine-tuning the models led to consistent improvements over V1, showing that this might be a viable approach to partially overcome the lack of ER corpora across languages. On a similar note, despite using the learned metric trained only on Spanish data for the other languages in V3, the tendencies were similar for all four languages, with gains in SARI across the

[6] Although BLEU against the source text might provide some indications of a model's tendency to remain close to the input text, we view it as difficult to interpret, since ER adaptation is typically meant to differ from the source text to a significant degree, while preserving some core informational elements. We thus discarded it from our analyses.

board, for instance. We will further explore tuning a specific metric for each language in future work.

5.2. Task Results

Table 3
Gold data results for our approaches. We indicate the highest score for each metric per language in bold, and the second best in italics. We show baseline1 results and the best average result for a given language among other participants in the shared task. There were no other participants for Arabic.

Approach	Language	BLEU	SARI	BERTScore	Avg
Baseline1	ar	10.39	35.77	92.52	46.23
	ca	10.21	44.01	90.45	48.22
	es	10.88	43.29	91.30	48.49
	it	*10.15*	46.51	91.11	49.25
-	ar	-	-	-	-
ClearText (run1)	ca	**12.95**	**48.78**	92.12	**51.28**
Hulat1 (run2)	es	**17.71**	*45.02*	92.72	**51.82**
Hulat1 (run3)	it	**16.00**	**48.68**	**92.45**	**52.38**
V1	ar	*30.79*	**39.30**	*93.50*	*54.53*
	ca	9.97	*44.71*	91.01	*48.56*
	es	9.33	44.19	91.14	48.22
	it	8.01	*46.48*	91.37	*48.62*
V2	ar	20.48	37.19	91.94	49.87
	ca	*11.44*	36.18	**92.33**	46.65
	es	*14.75*	42.17	**92.46**	49.79
	it	9.49	37.01	*91.55*	46.02
V3	ar	**31.10**	*39.00*	**93.58**	**54.56**
	ca	9.61	42.55	*92.14*	48.10
	es	14.11	**45.58**	92.41	*50.70*
	it	8.70	41.63	91.48	47.27

In Table 3, we present the results for our approaches on the shared task test data. We also indicate the Baseline1 results, which was the strongest of the two baselines in all cases, and the scores of the participant who achieved the highest average score for a given language. Note that other participating systems may have scored higher on a given individual metric in a given language; selecting the best average is nonetheless indicative of balance across metrics, none of which being actually fully representative of adaptation quality in isolation, as previously noted.

Our first observation from these results is the underperformance of V2 in terms of SARI, where V1 performed markedly better in all languages. Considering that both V2 and V3, which rely on a model fine-tuned on ER data, performed significantly better on our own ER test data, a first conjecture could be that the adaptation style and domain of the gold data differed significantly from our own datasets, with the fine-tuned model over-fitting on a specific style and data characteristics. Since our data were extracted from four different sources of professionally adapted material, however, this result was rather surprising to us. Although languages other than English may have suffered from machine-translated data quality, leading to worse performance, the Spanish results with the original data followed similar trends, thus highlighting a general issue with the fine-tuned models over the gold data. Only APEC-style adaptation in V3 helped partially recover the negative impact of using a model fine-tuned on ER data, when adapting the gold test data.

Looking at the other metrics, the tendencies were more in line with those observed on our own data, with V2 and V3 achieving higher scores than V1 on BLEU and BERTScore overall, although BLEU scores were significantly lower overall. Taking all metrics into account, these trends are at least partially indicative of a fine-tuned model that led to significantly different simplification choices, while also

being closer semantically to the reference. Since we were limited to three runs, and based our selecting the fine-tuned model for V2 on V3 on the marked improvement achieved on our own test sets, we did not include a variant of V3 with a generic model, an option which could have shed further light on the observed discrepancies. We nonetheless further analyse the differences between our internal test data and the gold references in the next section.

Overall, whereas on average our best run was competitive in Spanish (1.12 points below the best system), the performance degraded in Catalan (-2.72 avg compared to the best system) and significantly more so in Italian (-5.11), to the extent that this was the only case where our best run fell below Baseline1. For these two languages, performance was degraded across all metrics for V2 and V3, compared to Spanish, whereas V1 achieved more balanced scores on SARI and BERTScore. These differences might be due to either the quality of the machine-translated ER data used to fine-tune the adaptation model, although translation quality between Romance languages is typically high; or, as conjectured above, they might be due to a significant difference in adaptation style between our data and the gold references in these languages.

As a final note on the gold data results, our systems achieved significantly higher BLEU scores, compared to the other languages, a trend that is actually the reverse of what we observed on our own test data. We also achieved our highest absolute BERTScore results on this language, again a result opposite what was achieved on our internal data.

5.3. Data Analysis

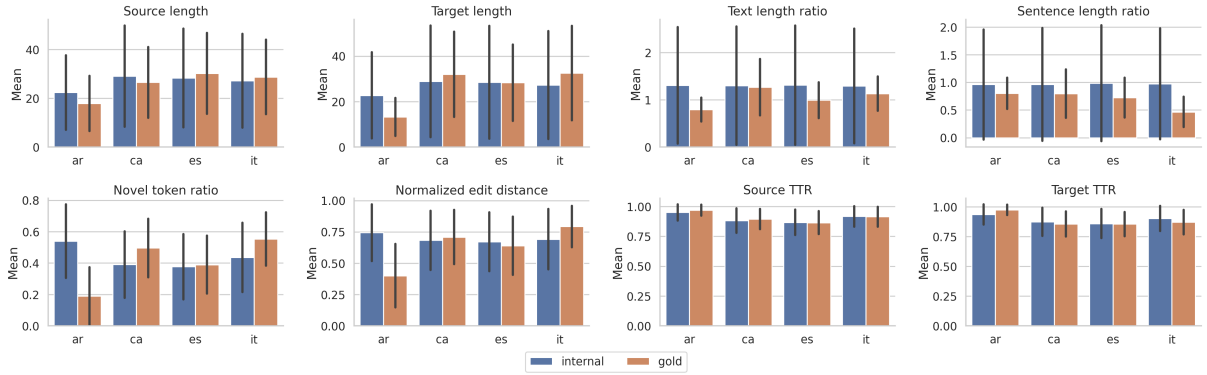
Table 4
Scoring differences between gold and internal test data. Positive values indicate higher gold data score.

Approach	Language	ΔBLEU	ΔSARI	ΔBERTScore	ΔAvg
V1	ar	+19.92	+3.75	+4.14	+9.27
	ca	-3.06	-0.46	+1.33	-0.73
	es	-2.37	-1.07	+1.03	-0.80
	it	+0.36	+4.23	+2.21	+2.27
V2	ar	+8.76	-5.12	+1.47	+1.70
	ca	-6.22	-12.62	+1.13	-5.90
	es	-4.73	-9.44	+0.47	-4.57
	it	-7.42	-9.34	+0.67	-5.36
V3	ar	+21.77	-5.11	+3.30	+6.65
	ca	-6.97	-6.43	+1.38	-4.01
	es	-3.45	-6.70	+0.80	-3.12
	it	-6.09	-4.96	+ 0.92	-3.38

In Table 4, we show the scoring differences for all three variants between gold scores and those achieved on our internal test set, which was composed from four different sources and domains. As noted in the previous section, the impact of fine-tuning over our own composite ER data, directly visible in the V2 variant, led to the most significant drop in scores for most languages. Arabic was the sole exception, with large gains in BLEU, but also significantly dropping on SARI. These effects were partially mitigated with V3 via post-editing recovery, although the effect of using a fine-tuned model was still notable in all languages but Arabic. The V1 variant, which only differs from V2 by using a generic model, underwent either minor losses in BLEU and SARI for Catalan and Spanish, or gains in all other cases.

To better understand these marked differences in behaviour, we analysed the characteristics of the data in both test sets, as shown in Figure 3. We indicate the following characteristics for both types of data, comparing the complex source and the ER target reference: source and target length in terms of tokens, the ratio of average number of target tokens over source tokens per sample (text length ratio), ratio of average target sentence length over source sentence length per sample (sentence length

Figure 3: Comparative data analysis across languages. For each subplot, bar height represents the mean value of the feature, and the gray vertical bar shows the standard deviation across sentence pairs.



ratio), the percentage of new tokens introduced in the adaptation (novel token ratio), the normalised edit distance at the token level, and the type to token ratio (TTR) in both source and target texts. The statistics on these features indicate important differences between both types of datasets, internal and gold.

First, in terms of global token ratios for source and target, our internal data featured comparable amounts of tokens between the source and adapted texts, in all languages, with following percentage changes in the amount of tokens from source to adaptation: +1.69% in Arabic; -0.63% in Catalan; +0.56% in Spanish; and +0.44% in Italian. These marginal differences can be derived from the different adaptation styles in our composed datasets, with adapted texts longer than the source in Irekia and Administración Fácil being balanced by the shorter adaptations in Plena Inclusión and Clears (see Table 1).

In the gold data however, the number of tokens differed significantly from the source to the adapted text: -25.69% in Arabic; +20.76% in Catalan; -6.22% in Spanish; and +13.28% in Italian. This is a first clear indication of different adaptation styles overall, not only between our internal data and the gold references, but also between languages in the gold data.

Another important difference is uniformity of the data distributions, as indicated by the standard deviations across the majority of features, with the gold data exhibiting significantly lower variation than our own test sets, in Spanish and all machine-translated derived test sets. This is an important aspect, due to differing adaptation styles in our collected data. Whereas the adaptation in the Plena Inclusión or Administración Fácil corpora is typically short and condensed, it is typically more dense in Irekia, despite all adaptations being done by experts in the field, as indicated in Table 1. The Clears data also have their own characteristics, being pre-adapted via ChatGPT rather than created from scratch by human experts. These differences are likely to account for a higher dispersion in core characteristics, which in turn led to a data distribution further removed from that of the gold data.

Looking at length ratios, both in terms of tokens and sentences, marked differences appeared as well, with denser and longer adaptations in our data overall. For instance, our test data featured a length ratio that was close to 1 between the source and target sentences, while in the gold the adapted sentences are shorter, as expected for an ER adaptation task. Models fine-tuned on our collected data were thus likely to promote adaptations further removed from the adaptation style in the gold data.

The introduction of novel tokens in the adaptation, and to a lesser degree the token-level edit distance, were also higher in the gold test set than in ours in all cases but Arabic, indicative of more pronounced lexical adaptation. This could be attributed to either a more systematic adaptation towards simpler terms in general in the gold data, or to a shared task domain that is more conducive to a need for adaptation compared to the domains in our test set. A more in-depth study would be necessary to more firmly establish the root causes of the differences in lexical variation between the datasets.

Finally, type-to-token ratios did not display marked differences, although our data featured slightly higher ratios, indicative of higher lexical diversity compared to the gold data. Note that in this case as well, Arabic displayed the opposite tendency.

These differences between the two datasets can provide a partial explanation for the significant differences we observed with our fine-tuned model, which may have overfitted on characteristics of our multi-source data that differed from those of the gold dataset. In future work, we will notably test our APEC method on the gold data with a generic model, expecting it to provide more direct adaptation refinements over initial adaptations that are less influenced by different adaptation styles.

6. Conclusions

We described Vicomtech's participation in the MER-TRANS shared task on multilingual adaptation to Easy Read. We submitted three runs in all four languages, contrasting three variants of ER adaptation on both our internal test sets and the gold data of the task.

Our first variant relied on a generic LLM, with an ER adaptation prompt that included incremental updates of the previously adapted sentences as context. Our second variant followed a similar setup, only differing in the use of a model fine-tuned on a varied range of ER corpora, which was machine-translated from Spanish to all other languages of the task. The third variant built upon the latter, adding post-editing cycles to refine the adaptation via the APEC approach, leveraging a learned metric trained to predict the scores of four metrics relevant to the task.

On our internal test sets, extracted or translated from the collected professionally-adapted Spanish ER data, the use of a fine-tuned model led to significant improvements, as did the use of post-editing cycles. On the task gold data however, using the fine-tuned model proved detrimental overall, outperformed by the generic model, with partial recoveries via the post-editing cycles. Our analysis of the internal and gold data showed marked differences across several dimensions, indicative of differences in adaptation style and significant differences across the four adaptation datasets we used in our tuning experiments.

Despite the somewhat negative impact of the fine-tuned model, we achieved robust results overall, outperforming the strongest baseline in all but one language. In future work, we will further explore data selection for more portable model fine-tuning, exploiting the new datasets provided by the MER-TRANS shared task.

Acknowledgments

We would like to thank the two anonymous reviewers for their very detailed and helpful comments. This work was partially supported by the Department of Economic Competitiveness of the Basque Government (Spri), via project ERAI (ZL-2026/00434).

Declaration on Generative AI

The authors have not employed any Generative AI tools for the preparation of this work.

References

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

[2] 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.

[3] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), 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.

[4] 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.

[5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020.

[6] J. Calleja, D. Ponce, T. Etchegoyhen, Text adaptation to plain language and easy read via automatic post-editing cycles, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.

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

[8] B. Botella-Gil, I. Espinosa-Zaragoza, A. Bonet-Jover, M. Madina, L. M. Pi, P. Moreda, I. Gonzalez-Dios, M. T. Martín-Valdivia, L. A. Ure, et al., Overview of CLEARS at IberLEF 2025: Challenge for Plain Language and Easy-to-Read Adaptation for Spanish texts, Procesamiento del Lenguaje Natural 75 (2025) 393–400.

[9] J. Calleja, T. Etchegoyhen, Irekier: An easy read corpus for basque and spanish, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 1633–1649. doi:10.63317/2e96m595cmfg.

[10] Asociación Española de Normalización, UNE 153101:2018 EX. Lectura fácil. Pautas y recomendaciones para la elaboración de documentos, Norma Técnica Experimental, 2018.

[11] K. Mühlenbock, Readable, legible or plain words–presentation of an easy-to-read swedish corpus, in: Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics, volume 8, Acta Universitatis Upsaliensis Uppsala, Sweden, 2008, pp. 327–329.

[12] E. Rennes, A. Jönsson, Towards a corpus of easy to read authority web texts, in: Proceedings of the Sixth Swedish Language Technology Conference (SLTC2016), Umeå, Sweden, 2016.

[13] V. Vandeghinste, B. Bulté, L. Augustinus, Wablieft: An easy-to-read newspaper corpus for dutch, in: Proceedings of CLARIN Annual Conference, 2019, pp. 188–191.

[14] S. Bott, V. Riegler, H. Saggion, A. Alcaina Rascón, N. Khallaf, A multilingual human annotated corpus of original and easy-to-read texts to support access to democratic participatory processes, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 1117–1128. doi:10.63317/4b56cza6e7zk.

[15] I. Gonzalez-Dios, I. Gutiérrez-Fandiño, O. m. Cumbicus-Pineda, A. Soroa, IrekiaLFes: a new open benchmark and baseline systems for Spanish automatic text simplification, in: S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, W. Xu (Eds.), Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual), 2022, pp. 86–97. URL: https://aclanthology.org/2022.tsar-1.8/. doi:10.18653/v1/2022.tsar-1.8.

[16] P. Martínez, A. Ramos, L. Moreno, Exploring large language models to generate easy to read content, Frontiers in Computer Science Volume 6 - 2024 (2024). URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1394705. doi:10.3389/fcomp.2024.1394705.

[17] N. Freyer, H. Kempt, L. Klöser, Easy-read and large language models: on the ethical dimensions of llm-based text simplification, Ethics and Information Technology 26 (2024) 50.

[18] M. San-Martín, S. Gómez, J. Heras, I. Martínez, G. Mata, Zero-shot prompting for adapting texts to easy-to-read, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025),

CEUR-WS.org, 2025.
[19] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
[20] M. Ayesh, N. Gutiérrez-Rolón, F. Alva-Manchego, Cardiffnlp at clears-2025: Prompting large language models for plain language and easy-to-read text rewriting, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.
[21] D. Fernández, A. Díaz, Integrating linguistic knowledge into prompting strategies for spanish text simplification: Insights from the nil-ucm participation, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.
[22] L. Plaza, L. Araujo, F. López-Ostenero, J. Martinez-Romo, Prompting large language models for spanish easy-to-read text generation: Uned-ineda at clears, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.
[23] J. Calleja, T. Etchegoyhen, D. Ponce, Automating easy read text segmentation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11876–11894. URL: https://aclanthology.org/2024.findings-emnlp.694/. doi:10.18653/v1/2024.findings-emnlp.694.
[24] E. J. Hu, yelong 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.
[25] M. Marone, O. Weller, W. Fleshman, E. Yang, D. Lawrie, B. V. Durme, mmbert: A modern multilingual encoder with annealed language learning, 2025. URL: https://arxiv.org/abs/2509.06888. arXiv:2509.06888.
[26] Y. Wan, D. Liu, B. Yang, H. Zhang, B. Chen, D. Wong, L. Chao, Unite: Unified translation evaluation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8117–8127.
[27] S. Štajner, M. Franco-Salvador, P. Rosso, S. P. Ponzetto, CATS: A tool for customized alignment of text simplification corpora, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1615/.
[28] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al., Moses: Open source toolkit for statistical machine translation, in: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, 2007, pp. 177–180.
[29] N. Khallaf, S. Sharoff, Automatic difficulty classification of arabic sentences, in: Proceedings of the sixth Arabic natural language processing workshop, 2021, pp. 105–114.
[30] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

A. Adaptation Prompts

We provide generic English versions of the prompt below. All prompts were machine-translated into the corresponding language with ChatGPT.

Core ER adaptation prompt

```
You are an expert assistant in adapting texts to Easy Read. Your task
is to adapt the original sentences to make them easy to understand,
while maintaining the original structure.

Writing Rules:
1. Use short sentences of fewer than 15 words.
2. Use simple, everyday vocabulary.
3. If a complex term is essential, add a brief explanation.
4. Keep the titles and order of the original text. Do not invent
new sections or titles.
5. Avoid passive voice, relative clauses, and long sentences.
6. Replace percentages and difficult figures with words like
"almost all," "many," or "half."
7. Do not add personal comments or explain the changes made. Submit only
the adapted text.

Here is the original complex text:
<original_texts>
Here is the adapted text so far:
<prior_adapted_text>
Continue the adaptation with the following complex sentence:
<current_sentence>
```

APEC refinement prompt

```
You are an expert Easy Read editor.

Your task is to improve an adaptation that has already been created,
not to write a new version from scratch.

You must review the previous adaptation and correct it so that
it is easier to understand and more faithful to the original complex text.

Editing rules:

Preserve the meaning of the original complex text.
Use short sentences of fewer than 15 words.
Use simple, everyday vocabulary.
If a complex term is essential, add a brief explanation.
Keep the titles and order if they appear in the original text.
Avoid passive voice, long subordinate clauses,
and sentences with many ideas.
Do not add information that is not in the original text.
Do not explain the changes you made.
Return only the final edited version.
```

B. Corpus Translation

Below is an example, for Spanish-Catalan, of the few-shot demonstrations used to machine translate the Spanish corpus:

Few-shot demonstrations for corpus translation

```
{
    "role": "user",
    "content": (
        "Translate from Spanish into Catalan:\n\n"
        "El portavoz del Gobierno Vasco y consejero
        de Cultura y Política Lingüística, "
        "Bingen Zupiria, ha comparecido en la rueda de prensa habitual
        tras el Consejo de Gobierno."
    ),
},
{
    "role": "assistant",
    "content": (
        "El portaveu del Govern Basc i conseller
        de Cultura i Política Lingüística, "
        "Bingen Zupiria, ha comparegut a la roda de premsa habitual
        després del Consell de Govern."
    ),
},
{
    "role": "user",
    "content": (
        "Translate from Spanish into Catalan:\n\n"
        "El Gobierno Vasco ha aprobado hoy un decreto para facilitar
        el acceso a la vivienda."
    ),
},
{
    "role": "assistant",
    "content": (
        "El Govern Basc ha aprovat avui un decret per facilitar
        l'accés a l'habitatge."
    ),
},
{
    "role": "user",
    "content": (
        "Translate from Spanish into Catalan:\n\n"
        "Iñaki Arriola ha presentado un decreto importante.\n"
       "Este decreto ayuda a las personas a tener un lugar donde vivir."
    ),
},
{
    "role": "assistant",
    "content": (
```

```
        "Iñaki Arriola ha presentat un decret important.\n"
        "Aquest decret ajuda les persones a tenir un lloc on viure."
    ),
}
```