COMPLEXITYMT: Benchmarking the Interaction Between Text Complexity and Machine Translation

Joseph Marvin Imperial[1,3], **Junhong Liang**[4], **Belal Shoer**[4], **Abdullah Barayan**[2,10], **Rodrigo Wilkens**[5], **Omar Mussa**[11], **Dawn Knight**[2], **Eugénio Ribeiro**[8,9], **Ekaterina Kochmar**[4,6], **Sowmya Vajjala**[7], **Fernando Alva-Manchego**[2], **Harish Tayyar Madabushi**[1]

[1]University of Bath, [2]Cardiff University, [3]National University Philippines, [4]MBZUAI, [5]University of Exeter, [6]University of Victoria, [7]National Research Council, Canada, [8]INESC-ID Lisboa [9]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR [10]King Abdulaziz University, [11]Saudi Electronic University

jmri20@bath.ac.uk, htm43@bath.ac.uk, alvamanchegof@cardiff.ac.uk

Abstract

When a text is translated, does the translation retain the complexity of the original? We introduce **COMPLEXITYMT**, a new challenge for assessing how text complexity and machine translation interact with and influence each other, using the Common European Framework of Reference for Languages (CEFR) levels as the measure of text complexity. Across six languages, including Arabic, Dutch, English, French, Hindi, and Russian, we evaluate three open-weight models, one closed model, and a commercial machine translation system on two tasks: i) *correlation* of CEFR with translation difficulty, and ii) *shifts* in CEFR levels of the source texts. Our experiments show that higher CEFR levels make texts more difficult to translate and that machine translation lowers the target text's CEFR level for most languages at the document level. These findings provide new insights for researchers and practitioners working on multilingual pedagogical content generation and MT difficulty estimation.

🤗 UniversalCEFR/ComplexityMT
UniversalCEFR/ComplexityMT

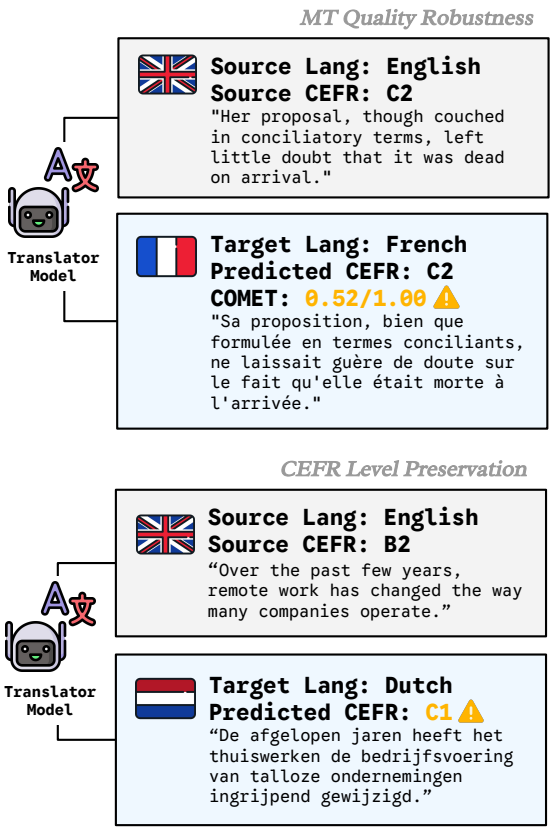

Figure 1: We observe two limitations in existing machine translation models: i) a *robustness* problem where they tend to produce lower quality translations (via COMET or GEMBA) for higher-complexity source texts, and ii) a *preservation* problem where they tend to change the pedagogical complexity levels (e.g., CEFR) of the source text.

1 Introduction

Creating reading materials tailored to a learner's proficiency level has been a topic of interest in educational research for a long time. Approaches such as traditional readability formulas (Kincaid et al., 1975; DuBay, 2004; Crossley et al., 2017) and proprietary measures such as Lexile scores (Lennon and Burdick, 2004) were used in the past to evaluate existing materials and adapt them to specific reading levels. More recently, Natural Language Processing (NLP) research has enabled automation of two foundational tasks in this direction, namely automatic readability assessment (Aluisio et al., 2010; Ciobanu et al., 2015; Xia et al., 2016; Deutsch et al., 2020; Vajjala, 2022) and text simplification (Maddela and Xu, 2018; Scarton and Specia, 2018; Nishihara et al., 2019; Alva-Manchego et al., 2020; Maddela et al., 2021; Sheang and Saggion, 2021; Alva-Manchego et al., 2025; Barayan et al., 2025), which could together support the creation of targeted reading level-specific texts. Re-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 9269–9283
October 24-29, 2026 ©2026 Association for Computational Linguistics

search in this direction has understandably focused on high-resource languages like English due to the abundance of data and models. Machine Translation (MT) offers a natural method to extend reading level-appropriate texts into more languages (Xu et al., 2016; Marchisio et al., 2019; Alva-Manchego and Shardlow, 2022; Zouhar et al., 2026). But understanding whether existing MT models are actually fit for this purpose requires empirical evidence of how text complexity affects translation quality, and if it preserves the complexity of the source text being translated. Understanding these interactions will guide how to leverage MT and text simplification to scale pedagogical content generation across languages.

The assumption that complex text could be difficult to translate is not new. Text simplification was considered a pre-processing step to reduce translation difficulty and improve translation quality, both in earlier MT models as well as more recently (e.g., Chandrasekar et al., 1996; Mehta et al., 2020), and recent work has shown that *complex texts are difficult to translate* (Shardlow and Alva-Manchego, 2022). At the same time, a separate line of work describes *simplification as a translation universal*, i.e., translated text is shown to be more readable and less complex, as it relies on high-frequency tokens in the target language (Corpas Pastor et al., 2008; Lu et al., 2021; Wastl et al., 2025). In all these cases, the notion of complexity has predominantly been binary. Drawing on what appears to be contrasting viewpoints on the relationship between text complexity and machine translation, we revisit these questions in our paper, specifically in the context of pedagogical difficulty, which we operationalize using the CEFR scale (Council of Europe, 2001). Thus, we pose the following research questions:

- RQ1: How does the CEFR level of a text *correlate* with translation difficulty?
- RQ2: How does the CEFR level of a text *change* under translation?

These questions, motivated by problems in multilingual, pedagogically grounded content generation, also contribute to MT difficulty estimation research. To our knowledge, the interaction between MT and pedagogical difficulty-based content generation has not been studied in past NLP research, and this paper introduces a framework to study this through two tasks (see Section 3).

2 Related Work

We consider two strands of existing NLP research as directly related to our current work: MT difficulty estimation, and complexity-controlled MT.

MT Difficulty Estimation MT is one of the core research problems in NLP, and the topics of MT difficulty estimation, i.e., how difficult a text is for an MT system to translate and how to mitigate this effect, have all received some attention in past research. Various text-level features (e.g., text length and lexical diversity) have been explored to model source text difficulty in machine translation (Hale and Campbell, 2002; Mishra et al., 2013; Li et al., 2014; Bugliarello et al., 2020; Araghi and Palangkaraya, 2024; Proietti et al., 2025; Zouhar et al., 2026). Proietti et al. (2025) have recently considered the CEFR scale (Council of Europe, 2001) as a way to characterize translation difficulty. Text simplification has been explored as a preprocessing step before MT, to make the text easier to (machine) translate and improve translation quality (Chandrasekar et al., 1996; Mehta et al., 2020). However, to our knowledge, all prior research on MT difficulty estimation has focused primarily on sentence-level text and English source data and has been motivated primarily by improving MT translation quality, rather than by a pedagogical justification like ours.

In this paper, we look at MT difficulty estimation by considering the CEFR scale as the text difficulty measure, and MT quality as the difficulty measure for machine translation. Unlike other works on this topic, we go beyond sentence-level analyses and English and also consider the document-level across multiple languages to explore this question.

Complexity Controlled MT Generating translation variants by controlling for aspects such as politeness (Sennrich et al., 2016), formality (Nadejde et al., 2022), and personalization (Mirkin and Meunier, 2015) is well-studied in NLP research. In this line of research, controlling for text complexity in MT (Agrawal and Carpuat, 2019; Marchisio et al., 2019; Tani et al., 2022; Zouhar et al., 2026) is somewhat related to our second research question, although the specific question of whether translation can preserve the original text complexity is not explored in previous work. Further, this strand of research assumes that the target translation's text complexity level is prespecified.

In this paper, we explore whether translation

preserves the source language's text complexity in the target language. Compared to previous works, we use a pedagogical complexity construct, CEFR, as our main reference for complexity for measuring shifts from MT models.

3 COMPLEXITYMT: Benchmarking the Interaction between Text Complexity and Machine Translation

As discussed in Section 2, previous work has expressed contrasting views on the relationship between text complexity and MT difficulty (as measured by MT quality). In that context, we introduce COMPLEXITYMT, a framework for assessing the impact of text complexity on MT across two core aspects: robustness and preservation. **Robustness** captures the expectation that good MT models should maintain translation quality across the text-complexity spectrum, thereby addressing RQ1. **Preservation** builds on the expectation that text complexity is maintained across translations, addressing RQ2. The following sections describe our experimental pipelines to assess both aspects.

3.1 COMPLEXITYMT-ROBUSTNESS

This task assesses whether MT quality correlates with the complexity of the source text. Given a set of CEFR-labeled source texts and a set of target languages, the robustness of an MT system is assessed as follows:

1. We translate each source text x with an assigned gold-standard CEFR level $\ell_x \in \{A1, A2, B1, B2, C1, C2\}$ into each target language L_{tgt} using the MT model under evaluation to produce translation y;

2. We then estimate a reference-free MT quality score $q(y) \in [0, 1]$;

3. Finally, we compute the Spearman correlation $\rho = \text{corr}(\ell_x, q(y))$ between the source CEFR level and the MT quality score.

We use the computed correlation ρ between textual complexity and MT quality as the main robustness metric, with values closer to zero indicating higher robustness. A significant negative correlation ($\rho < 0$) indicates that the translation quality decreases as the CEFR level increases, implying that higher-level texts pose greater challenges for the MT system. Conversely, a positive correlation ($\rho > 0$) suggests that higher-level texts receive better quality scores, a less intuitive but theoretically possible outcome.

3.2 COMPLEXITYMT-PRESERVATION

This task assesses whether MT systems preserve the CEFR levels of the source texts upon translation. Since obtaining manual text-complexity annotations for the translations generated by each evaluated MT system is not feasible, we rely on a pretrained multilingual CEFR level classifier (see §4.4 for details). To separate translation effects from the CEFR classifier's calibration errors, we evaluate CEFR level preservation via a backtranslation procedure[1] with a model-anchored shift, rather than a direct gold-vs-classifier comparison, as described below:

1. Given a source text x in language L_{src} with the gold-standard CEFR level ℓ_{gold}, we translate x into a pivot language L_{piv} with an MT model, producing the forward translation y_{fwd};

2. We then translate y_{fwd} back into L_{src} with the same MT model, producing the backtranslation y_{back};

3. Next, we apply the CEFR level classifier f independently to each translation output to obtain the forward and backtranslation CEFR level predictions: $\hat{\ell}_{\text{fwd}} = f(y_{\text{fwd}})$ and $\hat{\ell}_{\text{back}} = f(y_{\text{back}})$;

4. Finally, we compute the model-anchored CEFR level shift:

$$\Delta\ell_{\text{model}} = \hat{\ell}_{\text{back}} - \hat{\ell}_{\text{fwd}}. \quad (1)$$

We use the computed CEFR level shift $\Delta\ell_{\text{model}}$ as the main preservation metric, where a value of $\Delta\ell_{\text{model}} = 0$ signals that the backtranslation process preserved the original CEFR level, while a value of $\Delta\ell_{\text{model}} \neq 0$ indicates a net CEFR level shift induced by the MT model during translation. The purpose of anchoring on the classifier's forward prediction is two-fold. First, it minimizes the classifier's disagreement with the gold label of the

[1] A more direct test one can visualize is to compare the gold CEFR level of the source text assessed by human experts with a classifier-provided CEFR level of its forward translation. However, we avoid this setup since we cannot efficiently obtain human CEFR assessments for our translations. Comparing shifts between human expert and classifier-provided labels introduces a *label discrepancy* that would need to be mitigated.

Source	Lang	Size	CEFR Level Breakdown					
			A1	A2	B1	B2	C1	C2
Document-Level								
ELG-CEFR-NL (Breuker, 2023)	NL	200	4	21	69	77	22	7
ELG-CEFR-EN (Breuker, 2023)	EN	72	3	13	13	17	17	9
CAMBRIDGEEXAMS (Xia et al., 2016)	EN	43	0	6	9	7	10	11
KWIZIQ (Vásquez-Rodríguez et al., 2022)	FR	200	9	15	80	65	31	0
Total	–	**515**	**16**	**55**	**171**	**166**	**80**	**27**
Sentence-Level								
README++ (Naous et al., 2024)	FR	200	23	42	52	35	32	16
	RU	200	41	36	52	35	24	12
	AR	200	12	20	57	66	30	15
	HI	200	38	34	42	42	28	16
	EN	45	3	6	10	16	9	1
CEFR-SP (Arase et al., 2022)	EN	155	2	17	51	55	28	2
Total	–	**1,000**	**119**	**155**	**264**	**249**	**151**	**62**

Table 1: The distribution of multilingual reference texts extracted from the UNIVERSALCEFR test split ($n = 1,515$) we used for this study across the CEFR levels and formats (document- and sentence-level).

source text, thereby leaving the effect exclusively to the classifier's predictions. Second, as mentioned earlier, we lack human expert CEFR labels for the translations. We note that the classifier is applied to texts of two different languages on the two legs of backtranslation (L_{piv} for y_{fwd} and L_{src} for y_{back}), thus per-language prediction differences in the classifier are not canceled in $\Delta\ell_{\text{model}}$. We address this by investigating robustness across three structurally different classifiers (Section 5). Given this setup, $\Delta\ell_{\text{model}}$ represents the *net effect* of forward and backward translations combined together, which we treat as the core measure of CEFR preservation.

4 Experimental Setup

In this section, we detail an implementation and application of COMPLEXITYMT. The framework can easily be extended to additional languages, used to evaluate other MT systems, and improved through advances in automatic CEFR level prediction.

4.1 Data

We use a subset of CEFR-labeled texts from the test split of **UNIVERSALCEFR** (Imperial et al., 2025) at both the sentence and document levels.[2] We filtered UNIVERSALCEFR to extract the reference-level texts that are associated with gold-standard CEFR levels as reported in Table 1. The sentence-level data cover multiple source languages, including English, French, Arabic, Hindi, and Russian, while the document-level data include English, French, and Dutch. Each text is associated with an original CEFR label from A1 to C2. For analysis, CEFR levels are assigned to an ordinal scale, where A1 = 1, A2 = 2, B1 = 3, B2 = 4, C1 = 5 and C2 = 6. Table 1 shows the distribution of CEFR levels in the document-level and sentence-level subsets. The document-level subset contains 515 instances across three languages, while the sentence-level subset contains 1,000 instances across five languages, with 200 instances per language.

4.2 Translation Models

We evaluate five diverse MT systems spanning different training objectives and scales: a general-purpose LLM GPT-5.4 (OpenAI, 2025); three translation-specialized LLMs, TowerInstruct-7B (Alves et al., 2024) and TranslateGemma (Google Translate Research Team et al., 2026) in its 4B and 12B variants; and a commercial translation system, Google Cloud Translation API.[3] All models were accessed between February and May 2026.

4.3 Translation Quality Metrics

We select two reference-free MT quality metrics for COMPLEXITYMT-ROBUSTNESS, specifically COMET[4] (Rei et al., 2020) which measures quality via an encoder-based multilingual BERT model, and GEMBA-DA[5] (Kocmi and Federmann, 2023), which measures quality via prompting GPT-5.4 for a direct assessment. Both metrics have been shown to correlate strongly with human evaluation. We

[2] https://huggingface.co/UniversalCEFR

[3] https://cloud.google.com/translate

[4] https://github.com/Unbabel/COMET

[5] https://github.com/MicrosoftTranslator/GEMBA

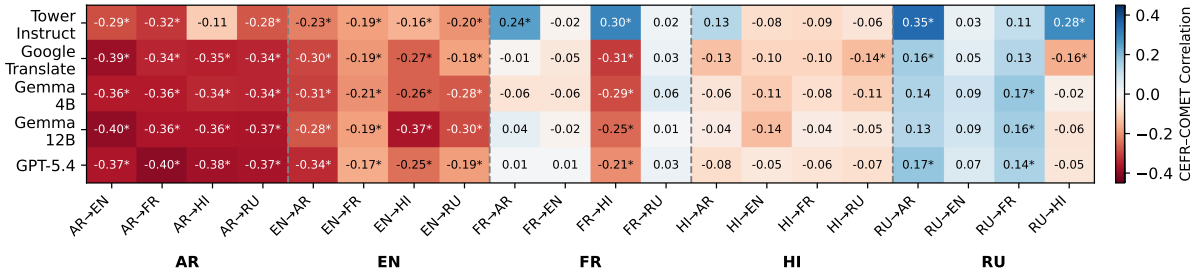

Figure 2: Spearman correlation between source CEFR levels and COMET scores across MT models for sentence-level texts. * indicates statistical significance ($p < 0.05$). Negative values indicate that translation quality, as measured by COMET, decreases as the CEFR level of the source text increases.

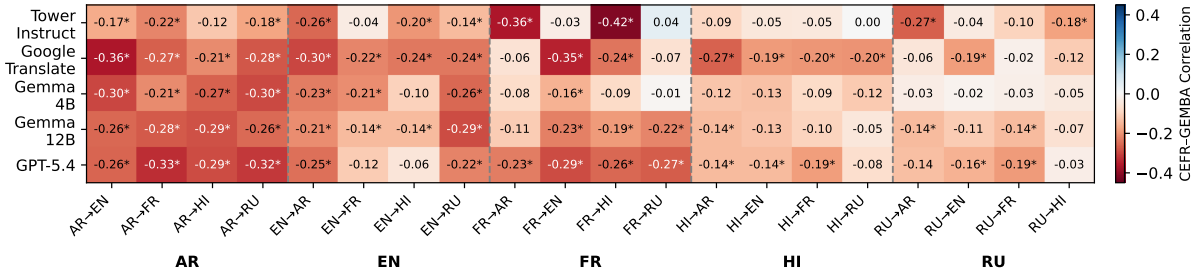

Figure 3: Spearman correlation between source CEFR levels and GEMBA-DA scores across MT models for sentence-level texts. * indicates statistical significance ($p < 0.05$). The significance pattern matches the results of COMET except for Russian where it diverges for GEMBA, although the absolute correlations are smaller.

use a uniform score scale of $[0, 1]$ where higher values indicate better translations.

4.4 CEFR Level Classifiers

To assess whether the CEFR level classifier impacts the results of the benchmark, we use a recent state-of-the-art XLM-R model finetuned with the massively multilingual train split of the UNIVERSALCEFR data. We verified that there were no overlaps between the train split used for training the XLM-R CEFR classifier and the curated test split used for the evaluation for COMPLEXITYMT-ROBUSTNESS and COMPLEXITYMT-PRESERVATION. Thus, we guarantee that there is no data leakage within the experiments. For comparison, we also perform cross-analysis with two additional CEFR classifiers that were trained with the same data but that differ architecturally – one is based on the ModernBERT architecture (Warner et al., 2025) and another is trained with 146 hand-engineered linguistic features using Random Forest (see Appendix A.3).

5 Results

In this section, we discuss the results of our experiments exploring how text complexity and MT affect each other across languages.

5.1 Translation Quality Declines with CEFR Level

We address RQ1 through the COMPLEXITYMT-ROBUSTNESS task described in Section 3.1. Figures 2 and 3 present Spearman correlation heatmaps capturing the relationship between text complexity and MT quality across the five multilingual MT models at the sentence level with COMET and GEMBA-DA scores respectively, while Figure 4 does the same for the document level.

Across most source languages at the sentence level, we observe a distinct pattern where *translation quality correlates negatively with the source CEFR levels* for Arabic and English, meaning that texts of higher CEFR levels are associated with lower translation quality according to COMET scores. These significant correlations are observed within the range of -0.16 to -0.40, and this pattern is consistent with the general expectation that higher-proficiency texts, which tend to contain more complex syntactic structures, richer vocabulary, and denser information, pose greater challenges for MT models. For GEMBA, we observe the same predominantly negative correlation pat-

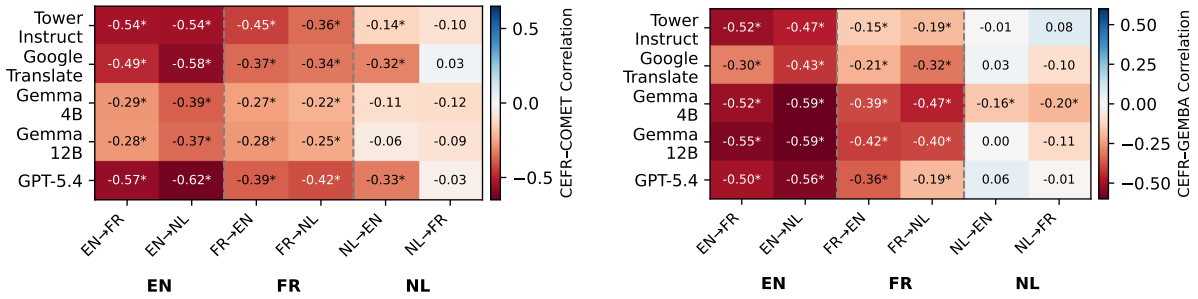

Figure 4: Spearman correlation between source CEFR levels and COMET scores (left) and GEMBA scores (right) across MT models for document-level texts. * indicates statistical significance ($p < 0.05$).

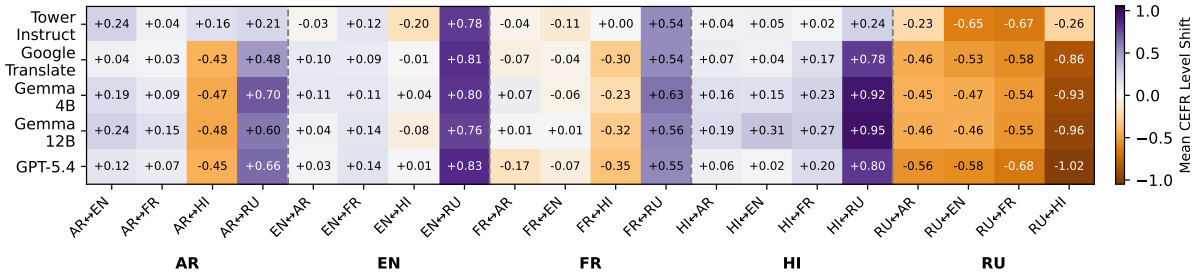

Figure 5: Mean model-anchored CEFR level shifts from back-translations across MT models for sentence-level texts. We anchor the UNIVERSALCEFR classifier's forward prediction to isolate the back-translation's effects from the classifier bias.

tern at smaller magnitudes. Interestingly, Russian stands out as an outlier with positive COMET correlations with CEFR level not seen under GEMBA. We posit that most MT models have been trained on higher-level Russian texts, which are more likely to conform to conventional written registers, thereby yielding higher translation scores.

At the document level, we observe that *negative correlation between translation quality and CEFR levels* is stronger and more uniform. For English and French, all MT models show significant negative correlations with both translation quality metrics where $\rho \approx -0.47$ and -0.34 for COMET, and $\rho \approx -0.50$ and -0.31 for GEMBA. For NL, this correlation is less pronounced with $\rho \approx -0.13$ and -0.04. We posit that the effect is more pronounced for documents than for sentences because longer inputs may carry more linguistic complexity, allowing translation difficulty to converge and yield less noise than for a single sentence.

5.2 Translation Shifts CEFR at Document Level

We observe how the CEFR level of a text changes under MT by visualizing the mean model-anchored level shifts from the backtranslation process for COMPLEXITYMT-PRESERVATION (Section 3.2) experiments across the five multilingual MT models. Figures 5 and 6 show sentence-level and document-level results, respectively, and cross-classifier robustness checks are reported in Table 2.

Table 2: Mean model-anchored CEFR level shift under three independently designed CEFR classifiers across MT models and granularities (25% sample), based on $n = 1500$ sentence-level and $n = 270$ document-level instances.

MT Model	XLM-R	ModernBERT	RF
Sentence Level			
GPT-5.4	+0.020	+0.037	+0.087
Google Translate	+0.045	+0.060	+0.069
TranslateGemma-4B	+0.071	+0.098	+0.106
TranslateGemma-12B	+0.062	+0.071	+0.115
Tower-7B	+0.011	+0.049	+0.157
Document Level			
GPT-5.4	−0.248	−0.374	−0.378
Google Translate	−0.196	−0.396	−0.222
TranslateGemma-4B	−0.163	−0.374	−0.215
TranslateGemma-12B	−0.204	−0.396	−0.193
Tower-7B	−0.237	−0.333	−0.400

At the sentence level, the mean CEFR shift is minimal at $\Delta\ell \leq +0.07$ across all MT models with 95% confidence interval excluding zero, indicating a small but statistically distinguishable upward shift. However, this effect is not uniform across languages and is most evident with Russian, where it receives a higher CEFR classifier predic-

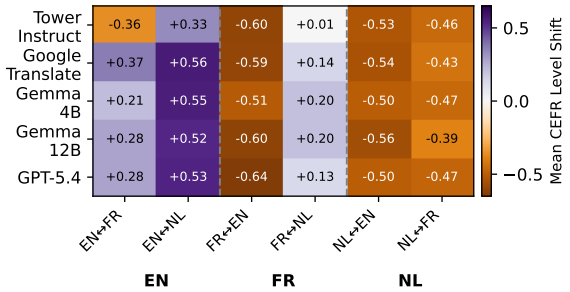

Figure 6: Mean model-anchored CEFR level shifts from back-translations across MT models for document-level texts. Level shifts are larger at the document level but remain consistent in direction across MT models for each back-translation.

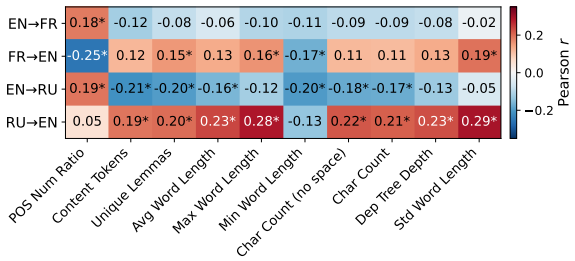

Figure 7: Pearson correlation coefficients between source-text linguistic features and COMET scores across four translation directions.

tion by $\approx +0.6$ levels on average and up to $+0.95$ max, whereas using it as the source text lowers it by ≈ -0.6 levels, down to roughly -1.0.

At the document level, we observe that *the CEFR level shift is larger and consistently negative for documents* with mean $\Delta\ell$ ranging from -0.16 to -0.31 with confidence intervals excluding zero for all MT models. Considering that these selected MT models differ in their training data, language coverage, and architectures, their similar level shifts in document-level texts indicate that this is a systematic phenomenon across all models rather than an outlier related to a single model. The three-model comparison of CEFR level shifts in Table 2 further confirms this observation, where all three CEFR classifiers produce downward shifts for all MT models, with a mean $\Delta\ell$ of -0.21, -0.38, and -0.28 for XLM-R, ModernBERT, and the Random Forest classifiers, respectively. As discussed in the setup in Section 3.2, these observed shifts reflect the total net effect of using backtranslation.

5.3 Linguistic Features Expose Finer-Grained Complexity Shifts

To systematically investigate the correlation between source text complexity and translation quality in a finer-grained setup, we design a small-scale control experiment with sentence-level texts evaluating four translation directions: EN→FR, FR→EN, EN→RU, and RU→EN. We use 200 sentence pairs per model-direction combination (3,200 instances in total), with translation quality measured via COMET. We use `spaCy`[6] to extract 37 source text linguistic complexity features, including length statistics, POS distribution, syntactic complexity, lexical richness, and argument structures, and compute Pearson correlation coefficients against COMET.

Our linguistic feature analysis, as reported in Figure 7, shows that translation directions involving Russian exhibit the strongest feature sensitivity. EN→RU and RU→EN display markedly opposite correlation patterns across nearly all complexity features, indicating substantial directional asymmetry. For instance, vocabulary-related features such as content tokens and unique lemmas are negatively correlated with quality in the EN→RU direction ($r \approx -0.20^*$), yet positively correlated in the RU→EN direction ($r \approx 0.19$–0.20^*). Word length features show the strongest associations in the RU→EN direction (up to $r = 0.29^*$), while POS ratio yields opposing significant correlations across FR→EN ($r = -0.25^*$) and EN→RU ($r = 0.19^*$). These findings suggest that *linguistic sensitivity is highly direction-dependent and that source-text complexity affects translation quality asymmetrically* across language pairs.

5.4 Translation Quality Does not Predict CEFR Shift

We visualize per-pair computed MT quality against the model-anchored CEFR level shifts in Figure 8 for COMET and GEMBA, and across sentence and document level texts. We randomly sampled 100 and 30 data points for sentence- (5 systems $\times$ 20 language pairs) and document-level (5×6 language pairs), respectively.

In all four experiment runs, we obtain non-significant near-zero results for Spearman correlation ($\rho = +0.05$ and -0.01 at the sentence level for COMET and GEMBA, and -0.12 and -0.12 at the document level), indicating that *translation quality and CEFR level shift are statistically independent*. From this, we con-

[6] https://spacy.io

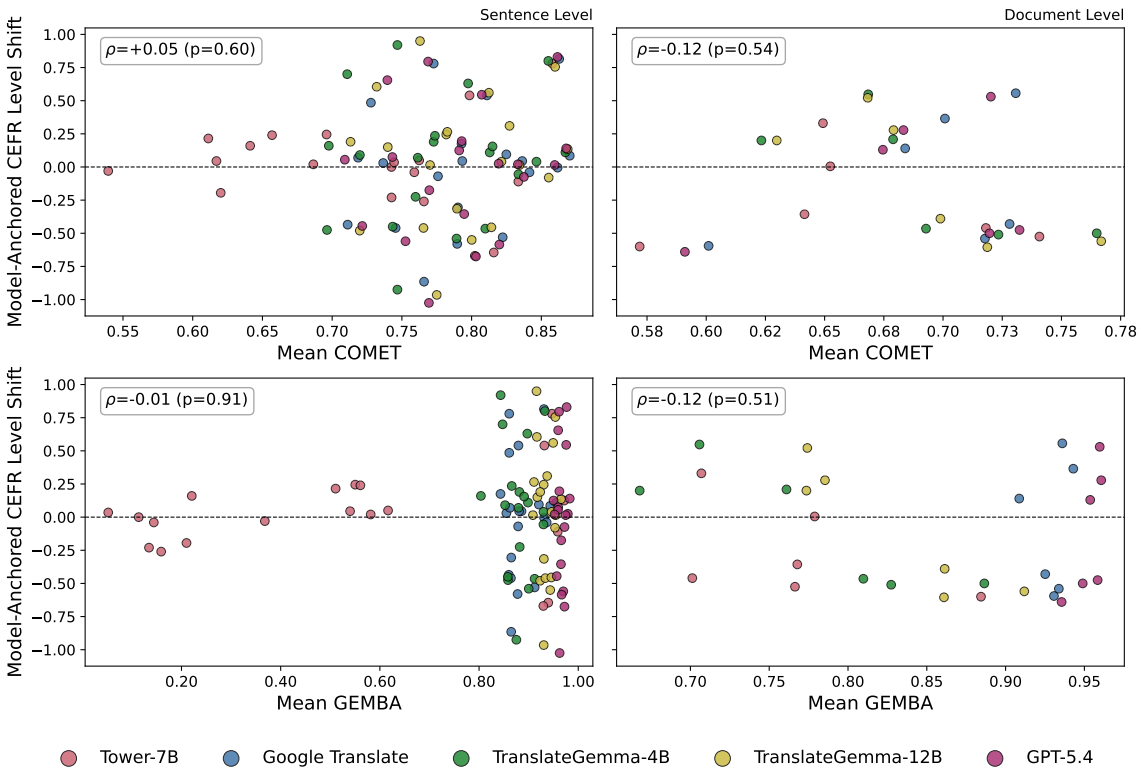

Figure 8: Scatterplot of MT quality via COMET and GEMBA versus model-anchored CEFR level shifts for sentence and document-level texts across MT models. The Spearman correlation is near zero in every panel ($|\rho| \leq 0.12$), indicating that translation quality does *not* predict the magnitude or direction of its CEFR level shift. This means that a high-quality translation is just as likely to shift the CEFR level of a text as a lower-quality translation.

clude that the two subtasks under COMPLEXITYMT, COMPLEXITYMT-ROBUSTNESS and COMPLEXITYMT-PRESERVATION, therefore measure distinct yet complementary properties of MT behavior. Since a high COMET or GEMBA does not lead to effective preservation of source CEFR levels, both a translation metric and a preservation metric may be needed for translation tasks, especially in MT applications sensitive to text complexity, such as educational content generation across multiple languages. This result demonstrates the importance of the COMPLEXITYMT framework as a strong evaluation challenge for current and future MT models.

6 Discussion

Relationship between Machine Translation and Complexity Prior works have formed two distinct viewpoints on the relation of translation and complexity: complex texts are complex to translate (Shardlow and Alva-Manchego, 2022), and the translation process inherently simplifies texts (Corpas Pastor et al., 2008; Wastl et al., 2025). Our empirical findings with COMPLEXITYMT *offer support for both viewpoints*. COMPLEXITYMT-ROBUSTNESS results show that translation quality indeed declines when the source text is of a higher CEFR level, while COMPLEXITYMT-PRESERVATION results provide evidence that the simplification universal phenomenon persists at the document level. Beyond confirming both, we also showed that translation quality and complexity level shifts do not correlate with each other. Therefore, our contribution to this discourse is not to resolve any disagreement—if such a disagreement exists—but to support both aspects by showing that they are *distinct dimensions of MT behavior*.

Pedagogically Motivated Content Generation Readability-controlled simplification is studied as a method to adapt content to different levels of text complexity, while MT enables scaling the adapted content into multiple languages, which is one of the motivations for this research. Our findings with COMPLEXITYMT show that translation can shift CEFR levels differently across languages, and these results *can inform content generation approaches by guiding when to simplify (at source or at target)*. For language pairs where translation tends to

increase the complexity of the source text, our findings show *the need for technical pipeline improvements, such as adding a readability-controlled simplification module* at the target end, and vice versa.

7 Conclusion

We introduced **COMPLEXITYMT**, a new challenge for assessing how text complexity and machine translation interact, using the CEFR as the measure of text complexity. Across six languages and five MT systems, our experiments showed that higher CEFR levels make translation more difficult and that MT via backtranslation lowers the target text's CEFR level relative to the source in most languages at the document level. These two effects are also statistically independent, where translation quality does not predict the magnitude of CEFR level shifts. Together, these findings highlight the need to consider both translation quality and CEFR preservation when evaluating MT applications sensitive to text complexity, such multilingual educational content generation.

Limitations

We identify a few limitations to our work, in terms of how we operationalize text complexity and the data used, which we discuss below.

Focus on CEFR We use CEFR (Council of Europe, 2001) as the reference pedagogical construct for text complexity in our study, as it is the most widely recognized language proficiency framework across the broader education and learning community. We recognize that our findings may not directly translate to other region- or country-specific pedagogical constructs, such as the Common Core Standards (CCS) in the United States, or China's Standards for English (CSE), to name a few. However, the methodology we followed can be replicated with other constructs, if relevant data resources are available.

Use of Automatic Classifiers for CEFR We use three diverse, state-of-the-art CEFR classifiers from (Imperial et al., 2025), which were trained on the massively multilingual gold-standard UNIVERSALCEFR dataset. While automated CEFR classifiers may have inherent prediction errors, their utility in our study remains appropriate and needed, considering that our main goal includes investigating how MT models can preserve CEFR levels of texts for educational content generation, for which these automatic CEFR classifiers will be a necessary resource in the process. We also emphasize that the classifier agreement we measured and reported among our three chosen CEFR classifiers in Appendix A.3 measures consistency in producing CEFR labels and should not be treated as an equal substitute for CEFR levels from human experts. Our overall findings for this paper are anchored in the relative CEFR level shifts that can be reproduced using these three classifiers, rather than relying on absolute human judgments for every language tested. This setup keeps our analyses self-contained and reproducible.

Data and Linguistic Coverage Our main results are anchored in the specific languages for which we obtained representative CEFR-labeled reference texts from UNIVERSALCEFR (Imperial et al., 2025). This includes English, Dutch, and French at the document level, and English, French, Russian, Arabic, and Hindi at the sentence level. We do not claim that our results will generalize to other languages, text formats (e.g., phrase-level), or text types (e.g., learner texts) not tested in this work. Since our sentence-level and document-level analyses cover different language test sizes and distributions, as seen in Table 1, we limit our cross-lingual findings to the languages we tested and their specific granularities. We do not claim that our findings broadly generalize to other languages we have not tested.

Focus on Quantitative Analysis Our results are primarily quantitative, given the straightforward goal of empirically investigating the effect of translation on text complexity and vice-versa. We acknowledge that a well-constructed qualitative analysis would enrich and complement our work, but recognize that this can be conducted as a separate study and leave it for future work.

Acknowledgments

This work was partly supported by the UKRI Strength in Places Fund MyWorld Project (SIPF00006/1).

JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

ER is supported by Portuguese national funds through Fundação para a Ciência e a Tecnolo-

gia, I.P. (FCT) under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025), UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025), and UID/04466/2025 (DOI: https://doi.org/10.54499/UID/04466/2025).

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. 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)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego and Matthew Shardlow. 2022. Towards readability-controlled machine translation of COVID-19 texts. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 287–288, Ghent, Belgium. European Association for Machine Translation.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 116–130, Suzhou, China. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. In *First Conference on Language Modeling*.

Sahar Araghi and Alfons Palangkaraya. 2024. The link between translation difficulty and the quality of machine translation: a literature review and empirical investigation. *Language Resources and Evaluation*, 58(4):1093–1114.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Mark Breuker. 2023. *CEFR Labelling and Assessment Services*, pages 277–282. Springer International Publishing, Cham.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Alina Maria Ciobanu, Liviu P. Dinu, and Flaviu Pepelea. 2015. Readability assessment of translated texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 97–103, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? a corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, pages 75–81, Waikiki, USA. Association for Machine Translation in the Americas.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, 54(5-6):340–359.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information.

Google Translate Research Team, Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report. *Computing Research Repository*, arXiv:2601.09012.

Sandra Hale and Stuart Campbell. 2002. The interaction between text difficulty and translation accuracy. *Babel*, 48(1):14–33.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Colleen Lennon and Hal Burdick. 2004. The lexile framework as an approach for reading measurement and success. *Electronic Publication on www.lexile.com*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288, Baltimore, Maryland. Association for Computational Linguistics.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-then-translate: Automatic preprocessing for black-box translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8488–8495.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351, Sofia, Bulgaria. Association for Computational Linguistics.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 260–266, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2025. GPT-5 System Card. *Computing Research Repository*, arXiv:2601.03267.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple TICO-19: A dataset for joint translation and simplification of COVID-19 texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3093–3102, Marseille, France. European Language Resources Association.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2022. A benchmark dataset for multi-level complexity-controllable machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6744–6752, Marseille, France. European Language Resources Association.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Michelle Wastl, Jannis Vamvas, and Rico Sennrich. 2025. Machine translation models are zero-shot detectors of translation direction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1054–1074, Vienna, Austria. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. 2026. Generating difficult-to-translate texts. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 204–219, Rabat, Morocco. Association for Computational Linguistics.

A Appendix

A.1 Libraries, Hyperparameters, and Configurations

We provide the full table of the Python libraries used and their corresponding versions in our experiments in Table 3. Likewise, we also provide the configurations and hyperparameter values used for the three CEFR classifiers (XLM-R, MODERNBERT, and RANDOM FOREST) in Table 4 across all CEFR level predicitons in this study.

Library	Version
`accelerate`	–
`bitsandbytes`	–
`huggingface-hub`	1.12.0
`joblib`	1.5.3
`matplotlib`	3.10.9
`numpy`	2.4.4
`openai`	–
`openpyxl`	3.1.5
`pandas`	3.0.2
`pyphen`	0.17.2
`requests`	2.33.1
`scikit-learn`	1.8.0
`scipy`	1.17.1
`sentencepiece`	–
`spacy`	3.8.13
`spacy-stanza`	1.0.4
`stanza`	1.11.1
`tokenizers`	0.22.2
`torch`	2.11.0
`tqdm`	4.67.3
`transformers`	5.6.2
`tseval`	1.0

Table 3: The Python libraries used in our experiments with with a main Python version of 3.14.4.

A.2 Utility Prompts

We provide the prompts we used for the LLM-based MT models we used in this study, including GPT-5.4, TowerInstruct-7B, TranslateGemma and for the LLM-based translation quality metric GEMBA.

GPT-5.4 Translation Prompt

```
Translate the following {src_name} sentence
to {tgt_name}. Output only the translation,
nothing else.

{text}
```

TowerInstruct-7B Translation Prompt

```
Translate the following text from {src_name}
into {tgt_name}.
{src_name}: {text}
{tgt_name}:
```

Parameter	Value
XLM-R (primary classifier)	
Identifier	UniversalCEFR/xlm-roberta-base-cefr-all-classifier
Max length	512
Truncation	enabled
Mode	`model.eval()`, `torch.no_grad()`
MODERNBERT (cross-classifier check)	
Identifier	UniversalCEFR/ModernBERT-base-cefr-all-classifier
Max length	512
Truncation	enabled
Mode	`torch.inference_mode()`
RANDOM FOREST (feature-based)	
Class	`sklearn.ensemble.RandomForestClassifier`
`n_estimators`	100 (default)
`max_depth`	None (default)
`random_state`	42
`n_jobs`	-1
Features	146, see (Imperial et al., 2025)
Training langs	`en, de, fr, ar, hi, nl, ru`
NaN handling	impute to 0

Table 4: Hyperparameter and model configurations for the three CEFR classifiers (XLM-R, MODERNBERT, and RANDOM FOREST) explored in the study.

Parameter	Value
COMET	
Checkpoint	`Unbabel/wmt22-cometkiwi-da`
Scale	$[0, 1]$
GEMBA	
Evaluator LLM	GPT-5.4
Temperature	0
Native scale	$[0, 100]$, rescaled to $[0, 1]$

Table 5: Configurations for the reference-free translation quality metrics COMET and GEMBA used in the study.

TranslateGemma Translation Prompt

```
{"role": "user", "content": [{
"type": "text",
"source_lang_code": "{src}",
"target_lang_code": "{tgt}",
"text": "{text}"
}]}
```

GEMBA Quality Scoring Prompt

```
Score the following translation from
{src_name} to {tgt_name} on a continuous scale
from 0 to 100, where a score of zero means "no
meaning preserved" and a score of one hundred
means "perfect meaning and grammar".
{src_name} source: "{src_text}"
{tgt_name} translation: "{tgt_text}"

Score (0-100):
```

A.3 CEFR Classifier Reliability

To investigate the robustness of the automatic CEFR classifiers we used in this work, we conduct a pairwise reliability test by computing Cohen's κ with quadratic weights and exact-match rate on a 25% random sample across the sentence and document-level texts across languages. We report the results in Table 6. Results show that all three structurally diverse CEFR classifier models exhibit moderate to high agreement (κ_{quad}), with XLM-R and ModernBERT showing the greatest agreement. All three CEFR classifiers also achieve $\approx 90.0+$ in adjacent agreement (± 1) and $\approx 50+$ in exact match rate.

Table 6: Pairwise CEFR-label agreement between XLM-R, ModernBERT, and the feature-based Random Forest model on round-trip text classification (forward and back-translations) across five MT models (25% sample). Cohen's κ with quadratic weights and exact-match rate.

Classifier pair	Scope	Exact	± 1	κ_{quad}
XLM-R vs. ModernBERT	All	0.662	0.973	0.835
	Sent	0.672	0.977	0.844
	Doc	0.610	0.947	0.695
XLM-R vs. RF	All	0.508	0.896	0.584
	Sent	0.515	0.896	0.576
	Doc	0.469	0.892	0.480
ModernBERT vs. RF	All	0.507	0.895	0.587
	Sent	0.518	0.893	0.572
	Doc	0.450	0.904	0.523

A.4 Source-Anchored Level Shift

In Section 3.2 with COMLEXITYMT-PRESERVATION, we discuss that we use a model-anchored shift that uses the forward and back-translation $\Delta\ell_{\text{model}} = \hat{\ell}_{\text{back}} - \hat{\ell}_{\text{fwd}}$. This represents the core translation process to quantify the CEFR level shift from the source and target languages. To investigate a within-language comparison from the back-translation process, we report a formula which is a source-anchored shift

$$\Delta\ell_{\text{source}} = \hat{\ell}_{\text{back}} - \hat{\ell}_{\text{orig}}, \tag{2}$$

where $\hat{\ell}_{\text{orig}} = f(x)$ is the CEFR classifier's prediction on the source text. This allows measuring the CEFR level shift from the same language of the source and back-translation texts. We produce the same heatmap visualization for the $\Delta\ell_{\text{source}}$ and report the results in Figure 9 for sentence level and Figure 10 for document level.

At the sentence level, we observe that this within-language shift is relatively small. For English source texts, in particular, receive lower CEFR levels across all MT models with around -0.21 to -0.41 shifts, while French, Russian, and Hindi source texts obtain slightly higher CEFR levels compared to their original for most MT models. We do note that this within-language comparison is observable with Russian where the high sentence-level CEFR level shifts as seen in Figure 5 using forward translation as anchor is not observed in Figure 9 using the source text with CEFR predicted level as anchor.

At the document level, we observe a sharper pattern where Tower-7B is the only outlier model that shows consistent within-language decrease in CEFR level shifts across all backtranslations, with the highest being English to French at -0.63. Meanwhile, the other MT models more or less retain the source CEFR level with very minor deviations. Again, we see the distinctiveness here in using the within-language comparison, where it can cancel the classifier's per-language offset, and the backtranslation is able to return to the source text's predicted CEFR level.

A.5 Use of LLMs

In producing this work, we used Grammarly for minor grammar and spelling corrections, ChatGPT and Claude Code for assistance with LaTeX table and figure formatting, troubleshooting code, and issues with Matplotlib visualizations. All suggestions from these tools were scrutinized by the authors before integration into the paper.

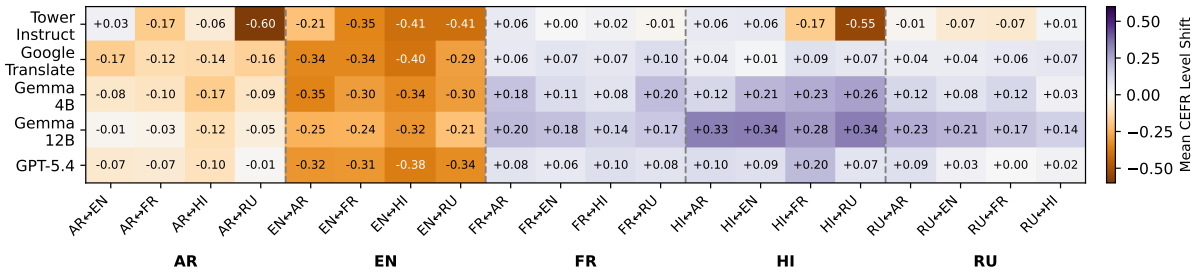

Figure 9: Mean model-anchored CEFR level shifts from back-translations across MT models for sentence-level texts. We anchor the UNIVERSALCEFR classifier's source text as anchor (vs. the forward translation shown in Figure 5) to investigate the effects of within-language differences.

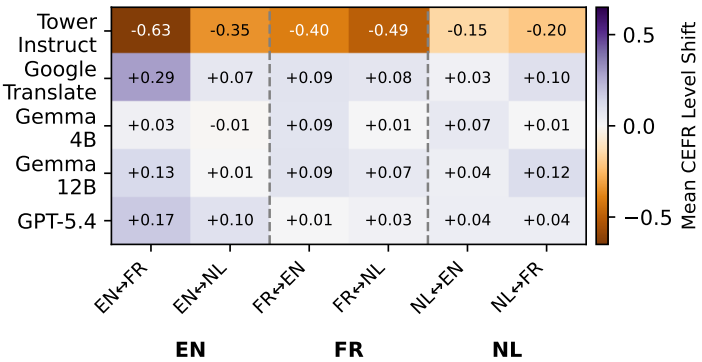

Figure 10: Mean model-anchored CEFR level shifts from back-translations across MT models for document-level texts using the source text as the anchor. Level shifts at the document level remain consistent with the sentence level of the same anchor.