Translate or Simplify First: An Analysis of Cross-lingual Text Simplification in English and French

Anonymous ACL submission

Abstract

Cross-Lingual Text Simplification (CLTS) aims to make content more accessible across languages by simultaneously addressing both linguistic complexity and translation. This study investigates the effectiveness of different prompting strategies for CLTS between English and French using large language models (LLMs). We investigate five distinct prompting systems: a direct prompt instructing the LLM to perform both translation and simplification simultaneously, two chain-of-thought (CoT) approaches that either translate-then-simplify or simplify-then-translate within a single prompt, and two pipeline approaches that perform the same operations in separate, consecutive prompts. These systems are evaluated across a diverse set of five corpora of different genres (Wikipedia and medical texts) using three state-of-the-art LLMs. Output quality is assessed through a multi-faceted evaluation framework comprising automatic metrics complemented by comprehensive linguistic feature analysis. Our findings reveal that while direct prompting consistently achieves the highest BLEU scores, indicating meaning fidelity, Translate-then-Simplify approaches demonstrate the highest simplicity, as measured by the linguistic features.[1]

1 Introduction

Text simplification (TS) is the task of reducing linguistic complexity while preserving meaning, thereby improving accessibility for diverse audiences such as language learners, children, or individuals with reading difficulties. Traditional text simplification focuses on reducing linguistic complexity within a single language by employing strategies such as lexical substitution, syntactic restructuring, and content condensation (Siddharthan, 2014). However, with diverse global audiences requiring accessible content, there is a growing need for Cross-Lingual Text Simplification (CLTS) systems that can simultaneously translate and simplify text across different languages.

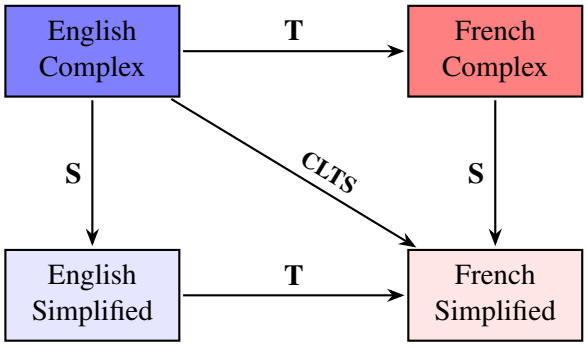

Figure 1: Conceptual diagram of text transformation tasks. Arrows represent the core operations: Translation (**T**), Simplification (**S**), and their combined operation, Cross-lingual text Simplification (**CLTS**).

CLTS presents unique challenges that go beyond the individual tasks of translation and simplification. While machine translation systems have achieved remarkable performance in recent years, they typically prioritize accuracy and fluency over accessibility (Leiter et al., 2024). Similarly, monolingual text simplification systems excel at reducing complexity within a single language but do not address the needs of multilingual audiences. The intersection of these two tasks requires careful consideration of how simplification strategies interact with translation processes, as linguistic complexity manifests differently across languages due to varying grammatical structures, vocabulary distributions, and cultural contexts (Miestamo, 2008).

CLTS can be conceptualized as a direct transformation from complex text in one language to simplified text in another. However, this single-step operation can alternatively be decomposed into sequential operations: either translating first then simplifying (T→S), or simplifying first then translating (S→T). Crucially, while Figure 1 depicts

[1] All the data and code will be made available upon publication.

these three paths as distinct routes to the same destination, they do not necessarily produce equivalent outputs. Each sequence of operations introduces different transformations, potential error propagation patterns, and linguistic trade-offs that may significantly impact both quality and characteristics of the final simplified text. This complexity is further compounded by the phenomenon of *translationese* – the distinct linguistic patterns that emerge in translated texts, which often include implicit simplification effects such as reduced lexical variety, syntactic normalization, and explicitation of implicit information (Volansky et al., 2014).

Traditional automatic text simplification approaches that rely on explicit training data and predefined rules face significant limitations in cross-lingual scenarios, including challenges in quality evaluation and cross-lingual adaptation (Alva-Manchego et al., 2021). The advent of large language models offers a fundamentally different approach to address these limitations. These models demonstrate remarkable capabilities in both translation and simplification tasks, often without explicit training on parallel simplified data. However, the optimal way to leverage LLMs for cross-lingual simplification remains an open question. Different prompting strategies may lead to varying trade-offs between translation quality and simplification effectiveness, and the order of operations (translate-then-simplify versus simplify-then-translate) may significantly impact the final output quality.

Recent advances in prompt engineering have shown that the way tasks are presented to LLMs can dramatically influence their performance (Wei et al., 2023). Chain-of-thought prompting, which encourages models to break down complex tasks into intermediate steps, has proven effective across various natural language processing tasks. Similarly, pipeline approaches that explicitly separate different sub-tasks may offer advantages over single-step approaches. However, the effectiveness of these different prompting strategies for CLTS has not been systematically investigated.

Furthermore, while automatic evaluation metrics provide valuable insights into system performance, they may not capture all aspects of simplification quality. Different metrics may favor different types of simplification strategies, and their effectiveness may vary depending on the target language and domain. The linguistic features that characterize effective simplification may also differ across languages and domains, highlighting the need for comprehensive analysis beyond traditional evaluation.

This study presents the first comprehensive evaluation of CLTS between English and French that systematically compares multiple prompting strategies across state-of-the-art language models. Specifically, we focus on bidirectional cross-lingual simplification: from English to French and from French to English, allowing us to investigate directional asymmetries in simplification strategies and their effectiveness. We investigate five distinct approaches: Direct prompting, Chain-of-Thought Translate-then-Simplify, Chain-of-Thought Simplify-then-Translate, Pipeline Translate-then-Simplify, and Pipeline Simplify-then-Translate. Our evaluation encompasses different language models, applied to five corpora. Table 1 illustrates example outputs from three of these approaches applied to an English source text from the ASSET corpus (Alva-Manchego et al., 2020).

This study advances the field through two pivotal contributions: First, we provide, focusing on English and French, the first systematic comparison of different prompting strategies for CLTS, revealing which approaches are most effective under different conditions with multiple automatic evaluation metrics and extensive linguistic analysis. Our findings indicate that Direct prompting achieves the best translation, while the highest simplicity is obtained by Translate-then-Simplify approaches. Second, we provide a comprehensive evaluation protocol for CLTS in English and French, including multiple datasets and evaluations adapted to the task, which can be extended to additional language pairs.

2 Related Work

2.1 Cross-Lingual Simplification

While monolingual text simplification has seen significant advancements, cross-lingual simplification remains an understudied area. Early work by Mallinson et al. (2020) on English-to-German simplification demonstrated the feasibility of zero-shot knowledge transfer from English to German. Later, the medical-domain `MultiCochrane` dataset introduced by Joseph et al. (2023) provides sentence-aligned parallel data enabling direct simplification from complex English texts to simplified versions in Spanish, French, and Farsi, facilitating research on how simplification strategies transfer across specific language pairs. This works has established that cross-lingual simplification benefits from un-

Original Text (English):

A few animals have chromatic response, changing color in changing environments, either seasonally (ermine, snowshoe hare) or far more rapidly with chromatophores in their integument (the cephalopod family).

French Simplification	English Translation
Direct Simplification: Quelques animaux changent de couleur pour s'adapter à leur environnement, soit au fil des saisons (hermine, lièvre d'Amérique), soit plus rapidement grâce à des cellules spéciales dans leur peau (les céphalopodes).	**Translation:** Some animals change color to adapt to their environment, either with the seasons (ermine, snowshoe hare), or more quickly thanks to special cells in their skin (cephalopods).
CoT T>S Simplification: Certains animaux changent de couleur pour s'adapter à leur environnement. Ce changement peut être saisonnier (comme chez l'hermine ou le lièvre d'Amérique) ou très rapide grâce à des cellules spéciales dans leur peau (comme chez les calmars ou les poulpes).	**Translation:** Some animals change color to adapt to their environment. This change can be seasonal (as in ermine or snowshoe hare) or very rapid thanks to special cells in their skin (as in squid or octopus).
Pipe S>T Simplification: Certains animaux changent de couleur pour se fondre dans leur environnement, soit saisonnièrement (comme l'hermine), soit rapidement grâce à des cellules cutanées spéciales (comme les pieuvres).	**Translation:** Some animals change color to blend into their surroundings, either seasonally (like stoats) or rapidly using special skin cells (like octopuses).

Table 1: Original English text from the ASSET corpus with multiple French simplifications generated by Gemini 2.5 Flash-Lite using 3 different prompting techniques. **Abbreviations:** CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline.

derstanding language-specific transfer patterns and operation ordering, though the optimal strategies for combining translation and simplification with current LLMs, across different domains, corpora, and language directions remain underexplored.

2.2 LLMs for (Cross-lingual) Text Generation

Recent advances in large language models have demonstrated remarkable capabilities in zero-shot text generation tasks across multiple domains. Brown et al. (2020) established the foundational principle with GPT-3, showing that large-scale pre-trained models can perform various NLP tasks without task-specific fine-tuning through appropriate prompting strategies. Building on this foundation, Wei et al. (2023) introduced chain-of-thought prompting, demonstrating that encouraging models to generate intermediate reasoning steps dramatically improves performance on complex reasoning tasks, and establishing the critical importance of prompt engineering strategies for LLM performance. In the context of cross-lingual tasks, Wang et al. (2023) explored zero-shot cross-lingual summarization, proposing effective prompting strategies that we adapt in our work by replacing summarization instructions with simplification objectives. Our work builds on these insights by systematically investigating how different prompting approaches, particularly the ordering of translation and simplification operations, affect CLTS performance across multiple models and domains.

2.3 Evaluation of Text Simplification

Evaluating text simplification quality presents unique challenges that differ from traditional NLP tasks, as simplification involves complex trade-offs between meaning preservation, readability improvement, and linguistic adequacy (Alva-Manchego et al., 2021). This has led to the development of specialized automatic metrics and feature-based approaches tailored specifically for assessing simplification effectiveness.

2.3.1 Automatic Evaluation of Simplification

The evaluation of text simplification has traditionally relied on machine translation metrics such as BLEU (Papineni et al., 2002). SARI (Xu et al., 2016) was introduced to explicitly reward additions, deletions, and retentions, and is now widely used for simplification evaluation. Later studies by

Sulem et al. (2018) confirmed that BLEU penalizes legitimate simplification operations like sentence splitting and may negatively correlate with simplicity, validating the need for specialized metrics like SARI, though BLEU remains valuable for assessing fluency. More recently, semantic similarity metrics such as BERTScore (Zhang et al., 2020) have been adopted to better capture meaning preservation in text generation tasks. Surveys highlight that no single metric fully captures simplification quality and recommend combining multiple complementary measures (Grabar and Saggion, 2022; Alva-Manchego et al., 2021).

2.3.2 Feature-Based Evaluation of Text Simplicity

Beyond automatic metrics, several studies have emphasized the role of linguistic features in assessing text simplicity. Vajjala and Meurers (2016) showed that lexical and syntactic features are effective predictors of sentence readability and can distinguish simplified from complex sentences. Recent work has expanded this direction by combining transformer-based models with diverse linguistic features, including lexical richness, readability measures, and psycholinguistic indicators, to improve the explainability of simplification systems (Qiao et al., 2022). Kreutz et al. (2024) introduced BATS (BenchmArking Text Simplicity), a comprehensive framework that implements 37 literature-derived features across multiple dimensions. We adopt the majority of BATS features in our evaluation framework, applying them to assess cross-lingual simplification—a novel extension beyond BATS original monolingual focus. These findings support our approach of integrating linguistic features to complement automatic metrics in evaluating CLTS, providing a strong foundation for understanding how different prompting strategies affect specific textual properties.

3 Prompting Strategies in CLTS

CLTS can be approached through various operational strategies that differ in 3 fundamental dimensions: (1) whether to perform CLTS directly or sequentially, (2) if sequential, which operation to perform first, and (3) whether to execute operations within a single prompt or across separate prompts. This section examines these strategic choices and their implications for CLTS system design.

CLTS involves two distinct transformations: converting text from one language to another (translation) and reducing linguistic complexity while preserving meaning (simplification). These operations can be combined in different ways:

Direct Approach: Performing both translation and simplification together in a single prompt, instructing the model to produce simplified text directly in the target language without explicit intermediate steps.

Sequential Approaches: Decomposing the task into two explicit operations performed in sequence, which can follow two different orderings:

Translate-then-Simplify (T→S): First converting text to the target language, then simplifying it.

Simplify-then-Translate (S→T): First simplifying in the source language, then translating the simplified version.

Implementation Methods: Sequential approaches can be implemented in two distinct ways:

Chain-of-Thought (CoT): Both operations specified within a single prompt, with explicit instructions to perform them sequentially.

Pipeline: Operations performed through separate consecutive prompts, with the output of the first serving as input to the second

This yields five distinct prompting strategies all explored in this work:

- Direct prompting
- CoT T→S (sequential, single prompt)
- CoT S→T (sequential, single prompt)
- Pipeline T→S (sequential, separate prompts)
- Pipeline S→T (sequential, separate prompts)

3.1 Theoretical Considerations

Each strategic choice involves distinct trade-offs:

Parallel vs. Sequential Processing: Direct prompting allows the model to optimize both objectives simultaneously but provides no explicit guidance on how to balance them. Sequential approaches provide clear task decomposition but may introduce error propagation between stages.

Operation Ordering: Translate-first approaches preserve source-language context for accurate translation but may introduce target-language complexities difficult to subsequently simplify. Simplify-first approaches reduce source complexity but may lose contextual information necessary for accurate translation.

Single vs. Multiple Prompts: Chain-of-thought prompting maintains both objectives within a single generation, potentially enabling better integration but increasing cognitive load. Pipeline approaches allow focused attention on each task separately but risk suboptimal coordination between operations.

3.2 Research Questions

Our systematic evaluation addresses key empirical questions about these strategic choices:

- Which prompting strategy yields the best performance across different evaluation metrics (BLEU, SARI, semantic similarity)?
- Does explicit task decomposition (sequential approaches) outperform simultaneous processing (direct approach)?
- Which approach achieves better results – translate-first or simplify-first, and does this vary by context/genre?
- How do different strategies differ in the specific linguistic transformations they perform?

3.3 Implications

Understanding the effectiveness of different prompting strategies has significant implications for CLTS system design, prompt engineering practices, evaluation methodology, and practical deployment. Through systematic comparison of these approaches across multiple models, domains, and language directions, this study provides empirically-grounded recommendations that advance both research and practice in CLTS.

4 Experimental Setup

This section describes the datasets, models, and preprocessing procedures in our CLTS experiments.

4.1 Datasets

We evaluate our approaches on 5 corpora that represent different domains and language pairs (Details about corpus sizes are in Appendix A):

ASSET: An English Wikipedia dataset with multiple human simplifications crowdsourced for each sentence (Alva-Manchego et al., 2020).

WIKI-AUTO: A large-scale English dataset automatically aligned from English Wikipedia and Simple English Wikipedia (Jiang et al., 2020).

MultiCochrane: The first multilingual, sentence-aligned medical text simplification dataset, which includes complex and simplified texts in English, Spanish, French, and Farsi (Joseph et al., 2023).

CLEAR: A French medical corpus of comparable texts from sources like encyclopedias and drug leaflets, with a subset of manually aligned parallel sentences (Grabar and Cardon, 2018).

WikiLarge-Fr: A French version of WikiLarge (Zhang and Lapata, 2017), translated into French using Google Translate (Ormaechea and Tsourakis, 2024).

4.2 Data Preprocessing

Current text simplification datasets are primarily designed for monolingual settings and lack the parallel structures necessary for evaluating cross-lingual simplification across both source and target languages. To address this limitation, we used machine translation to create the required cross-lingual evaluation data for monolingual corpora. Specifically, for each original-simplified text pair in a source language, we translated both texts to the target language using machine translation, creating the original text and reference simplified texts in the target language. The `MultiCochrane` dataset, which already provides cross-lingual parallel data, did not require this translation step. To ensure data quality and semantic consistency across all corpora, we applied a filtering process. We used Sentence-BERT (Reimers and Gurevych, 2019) to compute semantic similarity scores between original texts and their corresponding simplified versions. We filtered out all sentence pairs with semantic similarity scores below 0.6, ensuring that simplified versions maintain sufficient semantic overlap with their original counterparts while allowing for meaningful simplification transformations.

4.3 Models

We evaluate our prompting strategies using four state-of-the-art large language models:

GPT-3.5-Turbo: OpenAI's efficient conversational model with strong multilingual capabilities.[2]

GPT-4o-Mini: A GPT-4 small efficient variant.[3]

Gemini 2.5 Flash-Lite: Google's lightweight multimodal model with competitive multilingual generation capabilities.[4]

[2] https://platform.openai.com/docs/models/gpt-3.5-turbo
[3] https://platform.openai.com/docs/models/gpt-4o-mini
[4] https://deepmind.google/models/gemini/flash-lite/

Feature Category	Features
Lexical Features	lexical richness; infrequent words ratio; long words ratio; content words ratio; average word length
Syntactic and Structural Features	words before main verb; noun phrases ratio; relative clauses ratio; appositions ratio; conditional clauses ratio; conjunctions ratio; passive voice ratio; syntactic tree depth; sentences number; words per sentence; short sentences ratio
Readability Features	Flesch Reading Ease; Flesch Kincaid Grade; syllables ratio
Named Entity Features	max same entity distances; unique entities; unique entities average; avg same entity distance; entity to token ratio; unique entities to total num of entities; consecutive entity distance
Grammatical Features	modifiers ratio; negations ratio; past perfect verbs; past tense verbs; punctuation ratio; third person pronouns ratio

Table 2: Comprehensive linguistic features used for CLTS analysis.

mT5: We also use mT5 (Xue et al., 2021) as a baseline model to the prompting-based approaches. We fine-tune the model on the training set of each corpus and evaluate on the corresponding test set.

5 Methodology

This section details our specific approaches for CLTS, including prompting strategies, evaluation metrics, and analysis methods.

5.1 Prompting Strategies

We investigate five distinct prompting strategies that represent different approaches to combining translation and simplification tasks (prompts shown for French as target language; the same structure applies when English is the target):

1. Direct Prompting: A single-step approach where the model is instructed to perform both translation and simplification simultaneously: *"Please simplify the following text in French: "*

2. Chain-of-Thought Translate-then-Simplify: A single prompt that explicitly guides the model through a two-step reasoning process, first translating and then simplifying: *"Please first translate the following text to French and then simplify the translated text in French: "*

3. Chain-of-Thought Simplify-then-Translate: A single prompt that reverses the order, first simplifying in the source language and then translating: *"Please first simplify the following text and then translate the simplification to French: "*

4. Pipeline Translate-then-Simplify: A two-step pipeline where translation and simplification are performed in separate consecutive prompts:
Step 1: *"Please translate the following text to French: "*
Step 2: *"Please simplify the following text in French: "*

5. Pipeline Simplify-then-Translate: A two-step pipeline with reversed order:
Step 1: *"Please simplify the following text: "*
Step 2: *"Please translate the following text to French: "*

These prompting strategies allow us to investigate the impact of task decomposition (single-step vs. multi-step), reasoning guidance (chain-of-thought vs. direct), and operation order (translate-first vs. simplify-first) on CLTS performance.

5.2 Evaluation Metrics

We employ a comprehensive evaluation framework using multiple automatic metrics to assess different aspects of cross-lingual simplification quality:

BLEU (Papineni et al., 2002): Measures n-gram overlap between generated and reference texts, providing an indication of translation quality.

SARI (Xu et al., 2016): Specifically designed for text simplification evaluation, SARI measures the quality of lexical transformations by evaluating addition, deletion, and keeping operations against reference simplifications and the source. In our case, we translate the source to the language of the output, using Google Translate.

Semantic Similarity Metrics: To capture meaning preservation across languages, we use language-specific semantic similarity scores: **BERTScore** (Zhang et al., 2020), for texts simplified from French to English, and **CamemBERTScore** for texts simplified from English to French, using CamemBERT (Martin et al., 2020) instead of BERT (Devlin et al., 2019) embeddings.

To ensure statistical validity, we apply t-tests to compare the performance of different prompting strategies. We conduct separate statistical analyses for each combination of corpus, model, and evaluation metric. This approach allows us to identify sta-

Corpus	Method	GPT-3.5-turbo			GPT-4o-mini			Gemini 2.5 Flash-Lite		
		BLEU	SARI	Camem.	BLEU	SARI	Camem.	BLEU	SARI	Camem.
ASSET	Direct	**66.091**	39.184	**0.839**	**70.140**	39.173	**0.837**	**68.127**	39.712	**0.826**
	CoT T>S	57.105	42.558	0.805	58.775	43.911	0.799	47.134	41.445	0.763
	CoT S>T	59.474	41.748	0.795	60.496	44.282	0.805	48.923	40.505	0.744
	Pipe T>S	51.888	42.855	0.815	60.017	43.464	0.814	52.500	42.021	0.767
	Pipe S>T	48.027	43.675	0.807	60.384	44.191	0.814	54.242	42.097	0.765
WikiAuto	Direct	**30.966**	32.237	**0.793**	**32.406**	33.281	**0.795**	**29.950**	33.240	**0.785**
	CoT T>S	21.887	33.361	0.756	21.366	34.439	0.756	16.635	32.866	0.728
	CoT S>T	23.786	33.477	0.760	24.317	35.000	0.768	17.257	32.740	0.712
	Pipe T>S	22.024	33.703	0.768	25.634	35.264	0.777	18.879	33.654	0.735
	Pipe S>T	20.271	**34.252**	0.761	25.465	35.412	0.774	19.751	33.835	0.735
MultiCochrane	Direct	10.778	32.259	0.590	**13.462**	33.166	0.664	**12.931**	34.780	**0.663**
	CoT T>S	10.047	36.425	0.637	9.544	38.680	0.635	7.191	38.866	0.608
	CoT S>T	12.732	34.558	0.652	10.096	38.985	0.648	8.262	39.624	0.614
	Pipe T>S	9.883	37.929	0.650	10.451	37.794	0.653	7.918	39.049	0.623
	Pipe S>T	9.564	38.928	0.652	10.627	39.205	0.657	8.891	40.276	0.632

Table 3: Automatic Evaluation Metrics for Text Simplification Systems (English to French). **Bold** indicates that the best score is significantly higher than all others for the same corpus, model, and metric. **Abbreviations:** CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline, Camem. = CamembertScore.

tistically significant differences between prompting strategies, accounting for the variability introduced by different domains and models.

5.3 Feature-based Analysis

To gain deeper insights into the characteristics of successful CLTS, we extract and analyze a comprehensive set of linguistic features from the outputs of each prompting strategy. These features were designed to capture various linguistic and structural properties, which are detailed in Table 2.

We apply t-tests to compare these linguistic features between the outputs of different prompting strategies for each model, and corpus combination. This analysis helps identify which linguistic transformations are most characteristic of effective cross-lingual simplification and how these patterns vary across different experimental conditions.

6 Results

6.1 Automatic Evaluation Metrics Results

Performance of the different prompting approaches, using the different models across datasets are provided in Table 3 (English-to-French), Table 4 (French-to-English), and in Appendix B (mT5).

BLEU Score Performance: Direct prompting consistently achieves the highest BLEU scores across both translation directions and all three models (GPT-3.5-turbo, GPT-4o-mini, and Gemini 2.5 Flash-Lite), significantly outperforming other approaches on general domain corpora. Notably, translate-first approaches (CoT and Pipeline Translate-then-Simplify) consistently show the poorest BLEU performance in the French-to-English direction, suggesting significant operational ordering challenges for this language pair.

SARI Score Performance: Multi-step approaches dominate SARI scores across both directions and all models, with Pipeline Simplify-then-Translate emerging as the consistent leader. However, in almost all the cases, the differences are not statistically significant.

CamemBERT Score Performance (English-to-French): Semantic similarity for English-to-French CLTS reveals consistent patterns across models. Direct prompting achieves significantly superior CamemBERT scores on general domain datasets indicating strong semantic preservation for encyclopedic content across all three models.

BERT Score Performance (French-to-English): The French-to-English direction shows markedly different semantic preservation patterns. Direct prompting demonstrates superior BERT scores across both datasets and all models with statistical significance particularly pronounced for GPT-3.5-turbo and Gemini. Translate-first approaches (CoT and Pipeline Translate-then-Simplify) consistently show the lowest BERT scores across all models, indicating substantial semantic drift when French text is translated before simplification. This

Corpus	Method	GPT-3.5-turbo			GPT-4o-mini			Gemini 2.5 Flash-Lite		
		BLEU	SARI	BERT	BLEU	SARI	BERT	BLEU	SARI	BERT
WikiLargeFR	Direct	**34.516**	36.250	**0.648**	33.856	36.479	0.612	**33.486**	34.658	**0.642**
	CoT T>S	23.030	36.074	0.584	19.801	35.511	0.534	19.157	35.087	0.556
	CoT S>T	31.978	35.459	0.624	32.749	36.854	0.605	25.722	35.032	0.589
	Pipe T>S	19.771	35.676	0.566	23.832	37.248	0.569	17.943	34.381	0.538
	Pipe S>T	24.280	36.295	0.590	26.784	37.777	0.578	29.635	34.725	0.612
Clear	Direct	26.776	35.989	**0.634**	27.993	35.737	0.651	**27.250**	34.363	**0.639**
	CoT T>S	13.873	33.002	0.536	12.580	32.527	0.536	11.364	31.179	0.509
	CoT S>T	25.688	34.374	0.609	26.481	35.799	0.632	19.969	33.738	0.571
	Pipe T>S	11.338	31.829	0.506	15.541	34.330	0.569	10.507	30.402	0.479
	Pipe S>T	19.110	35.395	0.574	21.845	36.632	0.612	23.257	34.119	0.608

Table 4: Automatic Evaluation Metrics for Text Simplification Systems (French to English). **Bold** indicates that the best score is significantly higher than all others for the same corpus, model, and metric. **Abbreviations:** CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline, BERT = BERTScore.

directional asymmetry in semantic preservation is consistent across all three models, reflecting inherent linguistic properties of the language pair.

The **mT5** baseline model achieved substantially lower BLEU and semantic scores across all datasets compared to the prompted LLMs, though it showed competitive SARI performance on some corpora.

6.2 Features Based Analysis Results

Our linguistic analysis, grounded in the feature set detailed in Table 2, revealed distinct patterns in how each prompting strategy achieves simplification, differentiating their impact across lexical, syntactic, and structural dimensions. Results of selected features appear on Appendix C.

The **Direct Prompting** approach, which instructs the LLM to perform simultaneous translation and simplification, consistently demonstrated the least significant changes in Syntactic and Structural Features (e.g., minimal reduction in syntactic tree depth and words per sentence). This approach also consistently produced outputs with higher complexity metrics, including elevated Flesch-Kincaid grade (Kincaid et al., 1975) levels, lower Flesch Reading Ease (Flesch, 1948), indicating higher complexity, and increased syntactic tree depth. This suggests that direct prompting tends to generate more complex and formal language structures. This suggests it prioritizes fluency and translation fidelity (as seen in high BLEU scores) over deep structural transformations.

In contrast, the **CoT** methods excelled at targeted complexity reduction by leveraging explicit intermediate steps. Specifically, **CoT translate→simplify** generally resulted in simplified outputs, with higher Flesch Reading Ease scores (indicating better readability), reduced syntactic complexity, more sentences, higher short sentence ratios, and shorter sentence lengths. The **pipeline** approaches showed intermediate characteristics, with **pipeline simplify→translate** producing more complex outputs than **pipeline translate→simplify** (higher syntactic tree depth, words per sentence and Flesch-Kincaid grade). This indicates that the order of operations in pipeline methods significantly influences the final output characteristics.

7 Conclusion

Our comprehensive evaluation of CLTS in English and French reveals that no single prompting strategy universally outperforms others across all evaluation dimensions—rather, the optimal approach depends critically on the specific objectives and constraints of the simplification task. For applications prioritizing translation fidelity and lexical accuracy, Direct Prompting emerges as the most effective strategy, consistently achieving the highest BLEU scores across both language directions and multiple corpora. However, when accessibility and readability are paramount, translate-then-simplify approaches, particularly CoT translate-then-simplify demonstrate clear advantages, achieving the highest Flesch Reading Ease scores and most effective syntactic simplification while maintaining moderate lexical richness. Importantly, these trade-offs manifest consistently across different models and domains, though with varying magnitudes, suggesting that practitioners must strategically select prompting approaches based on whether their applications prioritize translation quality, readability optimization, or structural accessibility.

Limitations

While this study provides comprehensive insights into cross-lingual text simplification strategies, several limitations highlight directions for future work. First, our evaluation is restricted to the English-French language pair. Future work should explore the generalizability of our findings to other language combinations, particularly those with greater typological differences or lower-resource contexts. Second, some nuanced aspects of simplification quality such as cultural appropriateness, or actual reader comprehension are not captured by the automatic evaluation metrics (BLEU, SARI, BERTScore, and CamemBERTScore) and feature-based analysis used in the paper. Finally, while we evaluate multiple state-of-the-art LLMs, the rapid evolution of these models means our findings represent a snapshot of current capabilities. Emerging models or fine-tuning approaches may yield different patterns and will thus require novel evaluation and analysis, for which we provide tools in this paper.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Matthew Honnibal and Ines Montani. 2020. spacy: Industrial-strength natural language processing in python. *To appear*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. 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, Defense Technical Information Center (DTIC) Document.

Christin Kreutz, Fabian Haak, Björn Engelmann, and Philipp Schaer. 2024. BATS: BenchmArking text simplicity. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11968–11989, Bangkok, Thailand. Association for Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *J. Mach. Learn. Res.*, 25:75:1–75:49.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *CoRR*, cs.CL/0205028.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matti Miestamo. 2008. Grammatical complexity in cross-linguistic perspective. *Language Complexity: Typology, Contact, Change*, pages 23–42.

Lucía Ormaechea and Nikos Tsourakis. 2024. Automatic text simplification for french: model fine-tuning for simplicity assessment and simpler text generation. *International Journal of Speech Technology*, 27:957–976.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yu Qiao, Xiaofei Li, Daniel Wiechmann, and Elma Kerz. 2022. (Psycho-)linguistic features meet transformer models for improved explainable and controllable text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 125–146, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Thilina C. Rajapakse, Andrew Yates, and Maarten de Rijke. 2024. Simple transformers: Open-source for all. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 209–215, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165:259–298.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *Preprint*, arXiv:1603.06009.

Pauli Virtanen, Ralf Gommers, Travis Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan Walt, Matthew Brett, Joshua Wilson, K. Millman, Nikolay Mayorov, Andrew Nelson, Eric Jones, Robert Kern, Eric Larson, and Yoshiki Vázquez-Baeza. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:1–12.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2014. On the features of translationese. *Digital Scholarship in the Humanities*, 30.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A Corpus Statistics

	Asset	WikiAuto	MultiCochrane	Clear	WikiLarge FR
Train set size	20,000	576,126	28,998	4,110	264,258
Test set size	359	4,690	264	389	1,067

Table 5: Corpus statistics showing dataset sizes for training and evaluation

B mT5 Results

Metric	Asset	WikiAuto	MultiCochrane	Clear	WikiLarge FR
Bleu	9.678	11.213	2.529	3.214	8.700
Sari	28.817	37.525	35.436	24.708	29.658
Semantic Score	0.331	0.496	0.269	-0.004	0.132

Table 6: mT5 model evaluation results across different corpora and metrics. Semantic Score uses CamembertScore for English→French simplification and BertScore for French→English simplification.

C Linguistic Features Results

Corpus	Prompting System	Flesch Reading Ease	Flesch Kincaid Grade	Lexical Richness	Sentences Number	Short Sentences Ratio	Words per Sentence	Syntactic Tree Depth
ASSET	direct	62.645	9.550	0.909	1.028	0.052	23.355	4.544
	CoT T>S	70.805	7.426	0.926	1.085	0.131	18.575	4.016
	CoT S>T	68.140	7.965	0.926	1.035	0.135	19.199	4.136
	Pipe T>S	68.939	7.944	0.921	1.068	0.120	19.868	4.195
	Pipe S>T	66.640	8.380	0.916	1.084	0.111	20.174	4.338
WikiAuto	direct	64.700	9.411	0.905	1.068	0.069	24.308	4.479
	CoT T>S	72.631	7.158	0.925	1.157	0.154	18.810	3.943
	CoT S>T	70.247	7.759	0.919	1.083	0.138	19.804	4.066
	Pipe T>S	70.648	7.710	0.919	1.133	0.125	20.183	4.090
	Pipe S>T	68.632	8.112	0.912	1.159	0.117	20.522	4.256
MultiCochrane	direct	47.391	12.869	0.883	1.033	0.051	28.609	5.398
	CoT T>S	60.574	9.895	0.900	1.074	0.093	22.797	4.859
	CoT S>T	55.448	11.002	0.893	1.062	0.075	24.543	5.112
	Pipe T>S	57.557	10.415	0.902	1.058	0.091	23.312	4.928
	Pipe S>T	54.124	11.110	0.896	1.052	0.075	24.194	5.131
CLEAR	direct	41.612	11.792	0.939	1.022	0.114	20.144	6.070
	CoT T>S	55.058	9.429	0.955	1.043	0.154	17.535	5.493
	CoT S>T	43.261	11.302	0.943	1.021	0.152	18.897	5.771
	Pipe T>S	55.300	9.347	0.953	1.045	0.169	17.486	5.462
	Pipe S>T	43.972	11.210	0.942	1.026	0.138	18.837	5.882
WikiLarge-FR	direct	59.044	9.570	0.900	1.064	0.081	21.794	5.754
	CoT T>S	67.756	7.657	0.921	1.122	0.149	18.123	5.211
	CoT S>T	61.273	8.935	0.909	1.067	0.127	20.142	5.441
	Pipe T>S	67.573	7.669	0.922	1.111	0.152	18.201	5.207
	Pipe S>T	61.631	8.812	0.906	1.076	0.108	19.791	5.461

Table 7: Selected Linguistic Features by Prompting System Across Corpora (Average of GPT 3.5 turbo, GPT 4o mini and Gemini 2.5 Flash Lite). **Higher values indicate simpler text for:** Flesch Reading Ease, Lexical Richness, Sentences Number, and Short Sentences Ratio. **Abbreviations:** CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline.

D Used Python Packages

In our implementation, we utilize the following packages:

- easse (Alva-Manchego et al., 2019)
- NLTK (Loper and Bird, 2002)
- Pandas (Wes McKinney, 2010)
- Pyphen (https://github.com/Kozea/Pyphen)
- SciPy (Virtanen et al., 2020)
- sentence-transformers (Reimers and Gurevych, 2020)
- simpletransformers (Rajapakse et al., 2024)
- SpaCy (Honnibal and Montani, 2020)
- textstat (https://github.com/textstat)