Impact of automatic post-editing and prompting strategies on the linguistic features of English-to-French translations of editorials

Valentin Scourneau

revista.tradumàtica
tecnologies de la traducció

Valentin Scourneau
University of Mons (FTI-EII) & Polytechnic University of Hauts-de-France (LARSH);
valentin.scourneau@umons.ac.be;
valentin.scourneau@uphf.fr;
ORCID: 0009-0006-4295-7443

Abstract

This article consists in a comparison of the raw machine translation (MT), from English to French, and automatic post-editing (APE) of editorials, using three systems for MT (DeepL, Google Translate and GPT-4o) and three prompts with different levels of instruction specificity for APE. According to linguistic metrics, lexical and syntactic diversity increase across the board in APE as compared to MT, with the least constrained prompt generally leading to higher gains than the most constrained one. Meanwhile, quality estimation scores follow an opposite trend: less constrained prompts achieve lower COMETKiwi scores. A qualitative comparison of some machine-translated and automatically post-edited excerpts shows that APE can correct MT errors, reduce fluency errors or calques, and lead to more natural translations overall, but may also introduce new errors.

Keywords: machine translation, automatic post-editing, prompting strategies, linguistic metrics, lexical diversity, syntactic equivalence.

Resumen

El presente artículo consiste en una comparación entre traducción automática (TA) en bruto y posedición automática (PA). Se recupera la TA de tres sistemas (DeepL, Google Translate y GPT-4o) y se usan tres *prompts* que presentan instrucciones con distintos niveles de precisión para obtener las versiones poseditadas automáticamente. En cuanto a las métricas lingüísticas, el grado de diversidad léxica y sintáctica aumenta sistemáticamente en la PA en comparación con la TA en bruto: el *prompt* que presenta el menor grado de restricción desemboca generalmente en mejoras más marcadas que el de mayor grado de restricción. Sin embargo, las métricas de evaluación de la calidad producen resultados contrarios: los *prompts* con menor grado de restricción obtienen puntuaciones COMETKiwi inferiores. La comparación cualitativa de unos fragmentos de texto traducidos automáticamente con fragmentos poseditados automáticamente demuestra que la PA puede corregir los errores producidos por la TA, reducir la falte de fluidez y los calcos, y proponer traducciones más naturales, aunque también pueda llega a introducir errores nuevos.

Rebuda: 30 de juny de 2025| Acceptació: 9 de octubre de 2025 | Publicació: 15 de desembre de 2025

Revista Tradumàtica. Número 23 (Desembre 2025)
Cita recomanada: Scourneau, Valentin (2025). «Impact of automatic post-editing and prompting strategies on the linguistic features of English-to-French translations of editorials». Revista Tradumàtica. Tecnologies de la Traducció, 23, 065-100. <https://doi.org/10.5565/rev/tradumatica.533>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559
Universitat Autònoma de Barcelona

Palabras clave: traducción automática, posedición automática, técnicas de *prompting*, métricas lingüísticas, diversidad léxica, equivalencia sintáctica

.

Resum

Aquest article consisteix a comparar una traducció automàtica (TA) en brut i una postedició automàtica (PA). Es recupera la TA de tres sistemes (DeepL, Google Translate i GPT-4o) i s'utilitzen tres *prompts* que presenten instruccions amb diversos nivells de precisió per obtenir les versions posteditades automàticament. Quant a les mètriques lingüístiques, el grau de diversitat lèxica i sintàctica augmenta sistemàticament a la PA en comparació amb la TA en brut: el *prompt* que presenta el menor grau de restricció desemboca generalment en millores más marcades que el de major grau de restricció. Tanmateix, les mètriques d'avaluació de la qualitat produeixen resultats contraris: els *prompts* amb menor grau de restricció obtenen puntuacions COMETKiwi inferiors. La comparació qualitativa d'uns fragments de text traduïts automàticament amb fragments posteditats automàticament demostra que la PA pot corregir els errors produïts per la TA, reduir la falta de fluidesa i els calcs, i proposar traduccions més naturals, tot i que també pugui arribar a introduir errors nous

Paraules clau: traducció automàtica, postedición automàtica, tècniques de *prompting*, mètriques lingüístiques, diversitat lèxica, equivalència sintàctica.

1. Introduction

Since the introduction of large language models (LLMs), different ways of harnessing their language capabilities for translation-related tasks have been explored in the field of machine translation (MT). They include chain-of-thought prompting (Wei et al., 2022) and additional pretraining (Guo et al., 2024) or fine-tuning steps (Xu et al., 2024), LLMs for MT (e.g. Hendy et al., 2023), translation evaluation (e.g. Fernandes et al., 2023), and automatic post-editing (APE) (e.g. Raunak et al., 2023), which will be the focus of this work.

Various studies have examined the use of LLMs as MT systems (e.g. Hendy et al., 2023; Jiao et al., 2023; Kocmi et al., 2024; Moslem et al., 2023; Vilar et al., 2023; Wang et al., 2023; Zhang et al., 2023; Zhu et al., 2024). As models have scaled up in size, they have quickly become more and more effective. While LLMs were first introduced as MT systems at the 2023 Conference on Machine Translation (WMT) (Kocmi et al., 2023), they "exhibited some of the best quality translations" at the 2024 edition (Kocmi et al., 2024: 19). According to a literature review by Chan & Tang (2024), most studies on GPT for translation indicated it can generate translations that outperform those of neural machine translation (NMT) systems and are comparable to human translation (HT) based on automatic metrics. Despite these promising advances, LLMs can produce hallucinations — in which unrelated content that does not appear in the source text is inserted (Hendy et al., 2023) — and prompting strategies can significantly affect the quality of LLM-generated translations (He et al., 2024; Hendy et al., 2023; Jiao et al., 2023; Peng et al., 2023; Vilar et al., 2023; Wang et al., 2023; Yamada, 2024).

Most of the studies cited above focus on assessing translation adequacy through automatic metrics or human evaluation. As a matter of fact, in early 2025, Li et al. (2025) noted that limited attention had been paid to fluency and style in LLM-generated translation, but recent works by Alvarez-Vidal et al. (2025), Briva-Iglesias (2025), and Du et al. (2025) have shed light on those aspects, showing growing interest in this research avenue. Key findings include: i) LLM-generated translations can contain fewer fluency errors than NMT and may employ similar translation strategies to human translators, albeit to a much lesser extent (Alvarez-Vidal et al., 2025); ii) LLM-based multi-agent translation workflows may achieve higher fluency than NMT alone (Briva-Iglesias, 2025); and iii) with appropriate prompting, LLMs may produce more creative translations than NMT (Du et al., 2025).

A closely related research area is that of “translationese” (Gellerstam, 1986), or “third code” (Frawley, 1984), often discussed in relation to the translation universals (Baker, 1993). The concept of translationese, also called “translated language” or “language of translation”, especially in the broader field of translation studies (Jiménez-Crespo, 2023), refers to the specific features of translated texts, manifesting as fingerprints of the source text that may cause them to sound unnatural to native speakers (Gellerstam, 1986). It was originally studied in the context of human translations, in corpus studies by Laviosa-Braithwaite (1996) and Laviosa (1998). Based on lexical density and variety, evidence of the simplification translation universal (Baker, 1993; 1995) was found: translated newspaper articles (Laviosa-Braithwaite, 1996) and translated narrative prose (Laviosa, 1998) had lower lexical density and higher frequencies for high-frequency words than their non-translated counterparts. Later studies also established the presence of translationese in MT in the form of “machine translationese”, where MT results in lower lexical diversity (Hansen & Esperança-Rodier, 2022; Loock, 2018; Toral, 2019; Vanmassenhove et al., 2019; 2021) and higher syntactic equivalence (Hansen & Esperança-Rodier, 2022). MT systems also seem to produce more translations using cognates, i.e. words sharing a similar form and meaning in two languages (Čulo & Nitzke, 2016), although this might no longer be the case in newer systems. More recently, Niu & Jiang (2024) conducted a corpus study using metrics of lexical diversity as proxies for the translation universal of simplification in MT, HT, and original texts. They showed that overall, simplification seemed to occur in both MT and HT when compared to original texts, although to a greater extent in MT. Regarding LLMs for MT, while Raunak et al. (2023) concluded that GPT systems produce translations from English that are less literal than those of NMT systems, Chen et al. (2024) and Li et al. (2025) came to a more nuanced conclusion, with zero-shot translations exhibiting features of translationese. Similarly, there is mixed evidence regarding the existence of “post-editese” in human post-editing (HPE), with opposing results being obtained in different studies and sometimes even within the same study (Castilho et al., 2019; Castilho & Resende, 2022; Daems et al., 2017; Farrell, 2023a; Toral, 2019; Volkart & Bouillon, 2022; 2023; 2024), and virtually no evidence on “automatic post-editese” so far.

Another area of research on LLMs addresses their potential as editors of machine-translated texts (Chen et al., 2024; Farrell, 2023b; Ki & Carpuat, 2024; Kunilovskaya et al., 2024; Li et al., 2025; Macken, 2024; Raunak et al., 2023). For the sake of simplicity, we will refer to this method as "automatic post-editing" (APE), even though the prompts used in the studies above do not always specifically include the notion of post-editing. Although the examined language pairs, as well as prompt length and specificity (i.e. the mention or otherwise of "post-editing"), varied considerably across studies, interesting results have been reported across the board. With regard to translation quality, the above studies generally demonstrate that APE improves texts when compared to raw machine-translated output. Other studies on APE describe changes to linguistic and stylistic features, which will be the focus of this research. In particular, Macken (2024) observed that GPT, when used as a post-editor (GPT-APE), "made the most lexico-semantic changes in all texts compared to the human post-editors" and "improved lexical richness over the machine translation for all texts" (2024: 79). Li et al. (2025) reduced perplexity, which measures a language model's "decision (un)certainty" (Čulo & Nitzke, 2016: 108) and was employed as a proxy for translation unnaturalness, with a "polishing prompt" similar to APE prompts used in other studies. Similarly, Kunilovskaya et al. (2024) leveraged prompts including specific linguistic instructions to mitigate translationese as measured by morphosyntactic features in translations at the sentence level.

Nevertheless, to the best of our knowledge, no research has assessed the impact of different document-level APE prompts on lexical density, lexical variety, and syntactic equivalence metrics. Li et al. (2025) come closest to our efforts in that respect, and only a few studies have compared metrics of lexical diversity and syntactic equivalence in MT and APE. In this case study, we adapted and explored three APE prompts from the literature, each with a different level of instruction specificity. GPT-4o was used for English-to-French APE of editorials in a direct setting (i.e. without any prior evaluation of the raw MTs). Besides analysing the impact of the different prompts, we looked to investigate whether the APE process seems to lead to lexical and syntactic changes similar to those caused by the HPE process. We complemented the analysis of linguistic features with an automatic evaluation using COMETKiwi (Rei et al., 2022b) and a qualitative analysis of some of the changes found between the MT and APE steps. In doing so, we aimed to answer the following research questions:

RQ1. Are there significant differences between the raw MT outputs and the APE outputs in terms of lexical and syntactic metrics?

RQ2. Does APE contribute to reducing known machine translationese markers, such as decreased lexical variety and increased syntactic equivalence?

RQ3. What are the differences between the different APE prompts in terms of lexical density, lexical diversity, and syntactic equivalence?

RQ4. How do differences in lexical and syntactic metrics affect COMETKiwi scores?

RQ5. What qualitative differences can be found between MT and APE and between texts generated using different APE prompts?

2. Previous work

While the principle of APE is far from new, as evidenced by a literature review by do Carmo et al. (2021), LLM-powered APE is a relatively new topic and research specifically looking into aspects of variety, style, and literalness is still scarce. In this section, we will focus on studies dealing specifically with those aspects.

Farrell (2023b) authored one of the first studies dealing with lexical diversity in APE. He compared GPT3.5-generated MT, DeepL-MT, HT, monolingual GPT3.5-APE, and bilingual GPT3.5-APE using the existential construction "there is/are" as a previously identified English-to-Italian MT marker (Farrell, 2018). Out of four occurrences in a short Wikipedia article, he found that "there are" was translated with its literal equivalent "*ci sono*" four times, three times, twice, twice, and once, respective to each of the above conditions. Accordingly, he argued that the bilingual GPT3.5-APE reached higher lexical variety than the HT for the studied marker and the studied text. Nevertheless, that preliminary study did not include any metric of lexical or syntactic diversity. Another limitation was that the students who translated the article did not receive any specific instruction regarding lexical variety, while ChatGPT did.

Macken (2024) compared HPE to GPT-APE for three literary short stories in English translated into Dutch. In the study, the 23 human post-editors were professional translators with varying experience in translation and post-editing. Macken found an increase in lexical diversity in all GPT post-edited texts as compared to the raw MTs. When compared to HPE, it was among the highest values for one of the texts. In the same line, the GPT post-edited version generally contained more lexico-semantic and syntactic changes than the average human post-edited version, with a specific tendency to replace words with synonyms. Concerning the type of edits, GPT-4 made more preferential and undesirable changes and left more MT errors and inconsistencies unaddressed than the professional post-editors. It is especially interesting to note that GPT-4 made so many changes when the prompt specified that any unnecessary edits had to be avoided, i.e. when the MT was both faithful to the meaning and the style of the author.

Chen et al. (2024) built on the idea that human translators can re-read and edit their translations to propose an LLM-based iterative refinement prompting strategy. They argued that their method could prevent post-editese as the LLM could regenerate an unconstrained translation, unlike neural APE methods trained on human post-edited texts. According to their human evaluation, the refined sentence-level outputs achieved higher fluency and naturalness than the raw LLM translation outputs in all language pairs (English-German and English-Chinese). When compared to WMT human references, the refined outputs were also considered more fluent and natural in the case of working into English and were nearly tied in the English-to-German direction. Meanwhile, they obtained higher COMET (Rei et al., 2020) and COMETKiwi scores, but markedly lower scores according to string-based metrics, which suggests lexical and syntactic variations with no negative impact on translation quality. Thus, the method seems effective in increasing MT fluency and naturalness.

In the English-German language pair, Kunilovskaya et al. (2024) utilised four rewriting prompts including or omitting a definition of translationese and/or linguistic instructions with GPT-4 in order to reduce translationese as measured by a set of 58 linguistic features, 15 of which were identified as translationese predictors. Even the retranslation prompt effectively mitigated translationese features, but rewriting prompts yielded better results. Prompts containing specific linguistic instructions reduced translationese features more than those that did not. However, they scored lower for both fluency and accuracy in human evaluation, as well as according to COMET-DA (Rei et al., 2022a) and COMET-QE (Rei et al., 2020), although differences were small.

Recently, Li et al. (2025) successfully reduced the proportion of overly literal translations using a "polishing" prompting strategy and supervised fine-tuning (SFT) data with higher translation naturalness, obtained through their polishing prompt. After polishing, the proportion of GPT-4-generated translations exhibiting translationese, as evaluated by holders of linguistics or translation degrees, fell from 43% to 25% and 50% to 42% in the English-to-Chinese and German-to-English language pairs respectively. Likewise, implementing their SFT during the training of the Llama and Qwen LLMs effectively reduced perplexity, which was found to be a good predictor of translationese, while the resulting translations obtained higher COMET-QE scores and were ranked higher by human evaluators.

3. Methodology

This section explains the compilation of the source corpus, as well as the process followed to select the source texts analysed and the MT systems and prompts used. It also describes the linguistic metrics chosen and the statistical tests used.

3.1 Source corpus and selected source texts

Instead of using texts from an existing English-French parallel corpora, we gathered English editorials that had been published in British national newspapers (*The Guardian*, *The Independent*, and *The Telegraph*) between November and December 2024. This was to avoid data contamination, as no English-French parallel corpus was recent enough to be sure it had not been used as part of the training of GPT (OpenAI et al., 2024).

In each newspaper, we selected 16 editorials dealing with social issues or world and domestic events, giving a total of 48 editorials with word counts ranging from 448 to 789 words. In contrast to news items, editorials are not purely informative (Poibeau, 2022), as they seek to persuade readers, and "may have more freedom in their writing styles" (Marques & Mont'Alverne, 2021: 1816). These factors make this text type particularly interesting to investigate as part of a study looking into stylistic aspects such as lexical and syntactic variety. Indeed, as Farrell (2018) puts it:

> "Variety and inventiveness are not always desirable features in every kind of text. For example, excessive lexical variation might make a smartphone user's guide more difficult to follow.

> Nevertheless, there are various other kinds where lexical uniformity would make the text less interesting to read and less intellectually stimulating (marketing, advertising, literature, journalism, education, entertainment, and creative writing in general).” (2018: 58)

The editorials were mostly left untouched, apart from the removal of “The Guardian view on” from the titles of articles from *The Guardian* to avoid skewing the lexical metrics. As the editorials were saved in UTF-8 .txt format, we standardised the punctuation across all texts, but refrained from making other changes as the editorials were to be machine-translated and automatically post-edited.

When selecting the source texts, we aimed to maximise representativeness across a series of linguistic dimensions while enabling a certain level of comparison between the editorials. We selected the five most representative editorials, based on their Euclidean distance from the median values for word count, syntactic simplicity, TTR (type-token ratio), and lexical density, as computed using Coh-Metrix (McNamara et al., 2014), and for MATTR-100 (moving-average type-token ratio, 100-word window; see Section 3.3.2 MATTR), as computed using the R package `koRpus` and its built-in tokeniser. The Euclidean distances were computed using an ad hoc Python script. Accordingly, five editorials with word counts ranging from 569 to 601 words were identified as the most representative texts among the 48. Their representativeness enabled us to use bootstrap resampling for statistical testing, especially in the few cases where the normality hypothesis was violated. The source corpus data are available in the Appendix.

3.2 Machine translation, automatic post-editing, and prompts

In early February 2025, the five most representative source editorials were translated using two commercial NMT systems, Google Translate (GTrans) and DeepL, and one LLM, GPT-4o. The prompt used for the raw LLM-generated translations was based on Jiao et al. (2023), Peng et al. (2023), and Wang et al. (2023), but did not include domain information to avoid giving GPT-4o an unfair advantage. Note that in all the prompts presented below, [SRC] and [MT] correspond to the entire source editorial and the entire MT, without sentential boundaries. The translation prompt was as follows:

> You are a machine translation system. Please translate this document from English to French:
> [SRC]

The 15 raw translations were then automatically post-edited using each of three increasingly constrained prompts (P1, P2, and P3) in GPT-4o (default temperature), as described below.

P1 was the most effective iterative prompt in Chen et al. (2024), as tested on WMT22 datasets. “English” was prepended to “source”, and unlike P2 and P3, the prompt was run twice (first using the MT as input, then once more using the first round’s output as input). That number of iterations was chosen based on the results of Chen et al. (2024), where it gave the highest COMETKiwi scores in the English-to-German direction, with a third iteration leading to a decrease in performance. Another reason we did not run a third iteration is that it would have increased environmental costs. In contrast to the other two prompts, P1 did not include any instructions regarding syntactic or lexical variety and stylistic aspects:

> English source:
> [SRC]
>
> Translation:
> [MT]
>
> Please give me a better French translation without any explanation.

P2 was slightly adapted from Farrell (2023b). It mentioned syntactic variety in addition to lexical variety and specified "in English" instead of "in original language", so that all prompts clearly mentioned both the source and target language:

> Please post-edit the text below, which was machine-translated into French, keeping in mind that lexical and syntactic variety is required for a human-quality final text.
> Here is the source text in English:
> [SRC]
>
> Here is the text to post-edit:
> [MT]

Finally, P3, the most constrained prompt, was adapted from Macken (2024) as it was to be applied to editorials rather than literary texts. The prompt instructed an expert post-editor, imagined as a native French speaker with strong English proficiency, to refine French translations of English texts. The task balanced two priorities: (1) faithfulness to meaning (ensuring no omissions, additions, or distortions of the source text's content) and (2) faithfulness to style (preserving the author's tone and rhetorical effects while allowing creative solutions when needed). The post-editor was to correct errors in meaning or style only when necessary, adhering to strict rules: no unsupported additions, correct capitalisation and determiners, no unwarranted punctuation changes, and no assumptions about typos in the source. If a translation was very poor, the post-editor was allowed to rewrite it entirely, but was otherwise to make minimal edits to improve accuracy, fluency, and style (see the full prompt in the Appendix).

3.3 Linguistic metrics

Several linguistic metrics used in studies on machine translationese and post-editese enable us to quantify the lexical and syntactic differences between machine-translated and automatically post-edited versions of texts. Lexical density and lexical diversity have notably been used as proxies for the translation universal of simplification (Laviosa-Braithwaite, 1996; Niu & Jiang, 2024).

In this section, we present the metrics we computed, as well as results from studies in which they have been used. To analyse lexical differences, we considered lexical density and MATTR, the latter of which measures lexical diversity based on the degree of word repetition. Meanwhile, SACr (syntactically aware cross), ASTrED (aligned syntactic tree edit distance), and PoS (part-of-speech) changes provided insight into the level of syntactic equivalence between source and target text. We also computed the expanding ratio between source and target.

3.3.1 Lexical density

Lexical density is a measure of information density (Johansson, 2008) and is calculated as the ratio between the number of content words and the total number of words (Toral, 2019), as shown in Formula (1). Lemmatisation and tagging were performed using the `spaCy` library.

$$Lexical\ density = \frac{\#content\ words}{\#words} \quad (1)$$

Studies on post-editese have obtained different results regarding lexical density. Toral (2019) and Volkart & Bouillon (2022) reported increased lexical density in HT as compared to HPE, while other studies yielded mixed results depending on the corpus or language pair involved (Castilho et al., 2019; Castilho & Resende, 2022; Volkart & Bouillon, 2023).

3.3.2 MATTR

MATTR (moving-average type-token ratio) (Covington & McFall, 2010) was chosen as a metric of lexical diversity. Unlike TTR (type-token ratio), MATTR is insensitive to text length (Bestgen, 2024) since it uses a sliding window of *n* words, computing TTR for the words 1 to *n*, 2 to *n*+1, and so on until the end of a text of *w* words. One limitation of MATTR, though, is that the *n*-1 first and last words have a lower weight in the computation as they occur in fewer windows (Bestgen, 2024). MATTR is calculated as shown in Formula (2).

$$MATTR = \frac{\sum_{i=1}^{w-n+1} TTR_i}{(w-n+1)}\ where\ TTR_i = \frac{\#types\ in\ window\ i}{n} \quad (2)$$

In this work, we computed the MATTR-100 (*n* = 100 words) for each source and target text using the R package `koRpus` and its built-in tokeniser. We also calculated the TTR as it is commonly reported in the literature, but we did not overly rely on it due to its limitations. The computation was case-insensitive for both metrics.

Studies in the field of post-editese have generally relied on TTR-based metrics. MATTR was used by Hansen & Esperança-Rodier (2022), who reported a higher MATTR in human references as compared to MTs, as well as by Macken (2024) (see 2. Previous work).

3.3.3 SACr

SACr uses alignment crossings to quantify the degree of reordering between source and target word sequences (Vanroy et al., 2021). The metric is an improved version of seq_cross (Vanroy et al., 2019) that considers relations in the dependency trees to create linguistically motivated alignments. However, it remains a shallow-level metric of structural changes, which is why it is complemented by ASTrED. SACr is computed as the ratio between the number of alignment crossings and the total number of alignments (Vanroy et al., 2021), as shown in Formula (3).

$$SACr = \frac{\#crossings_{alignment}}{\#alignments} \quad (3)$$

We computed SACr between all source and target texts using the ASTrED Python library (Vanroy et al., 2021) in its fully automated mode, which is based on the Stanza tokeniser and parser (Qi et al., 2020) and a modified version of awesome-align (Dou & Neubig, 2021). The same method was employed to compute ASTrED and PoS changes (Sections 3.3.4 ASTrED and 3.3.5 Part-of-speech changes). We also report the unnormalised scores for the three metrics in order to provide a more comprehensive picture, especially since the alignment and part-of-speech tagging were performed automatically.

HTs have been observed to have a higher absolute number of alignment crossings than MTs (Hansen & Esperança-Rodier, 2022) and HPE (Schumacher, 2025). Volkart & Bouillon (2024), who worked on professional HT and HPE corpora, obtained mixed results, with HPE obtaining a significantly higher SACr than HTs in one corpus and an insignificantly lower SACr in two other corpora.

3.3.4 ASTrED

ASTrED is based on the same alignment method as that used for SACr or PoS changes, but it can capture deeper levels of syntactic reorganisation. The metric draws on the number of edits that are needed for the source dependency tree to match the target dependency tree instead of solely considering alignment crossings as SACr does. The number of edits is then normalised by the average number of source and target words (Vanroy et al., 2021), as shown in Formula (4).

$$ASTrED = \frac{\#edits_{source\_tree \rightarrow target\_tree}}{\left(\frac{\#words_{source} + \#words_{target}}{2}\right)} \quad (4)$$

Results from studies using ASTrED are more consistent. Both Hansen & Esperança-Rodier (2022) and Schumacher (2025) obtained higher unnormalised ASTrED in HT as compared to MT and HPE. Likewise, Volkart & Bouillon (2024) reported that normalised ASTrED was higher in HT than in HPE for the three corpora they used, even though the difference was not significant in one case.

3.3.5 Part-of-speech changes

Part-of-speech (PoS) changes are a way of quantifying how parts of speech (such as nouns, verbs or adjectives) differ between the aligned source-target pairs. The metric can be used as a shallow indicator of syntactic equivalence (Hansen & Esperança-Rodier, 2022). In sentence pairs aligned at the word level, it is computed as the number of instances in which the target PoS differs from the source PoS, normalised by the number of word-level alignments. Initially, we also calculated label changes, a metric calculated using the same library (Vanroy et al., 2021), but only PoS changes are reported in the results as they are strongly correlated with label changes in this analysis (Spearman's ρ = 0.76). For further details on the metrics of syntactic equivalence, see Vanroy et al. (2021).

The results obtained by Hansen & Esperança-Rodier (2022) and Schumacher (2025) converge here. They concluded that there are higher absolute numbers of PoS changes in HT in comparison to MT and HPE. However, they did not report this metric normalised by the number of word-level alignments or words in source and target.

3.3.6 Expanding ratio

The expanding ratio (ER) measures the difference in word count between the target and source text, normalised by the word count of the source text (Cochrane, 1995), as shown in Formula (5). In comparison with the length ratio (Toral, 2019), Volkart & Bouillon (2023) mention that the ER has the advantage of considering the source text length. Translation agencies generally expect ERs of between 20 and 25% in the English-to-French translation direction (De Clercq et al., 2021).

$$Expanding\ ratio = \frac{\#words_{TT} - \#words_{ST}}{\#words_{ST}} \times 100 \quad (5)$$

Studies comparing HT and HPE have obtained different results, finding higher (Martikainen & Kübler, 2016; Schumacher, 2025), lower (Volkart & Bouillon, 2022), and either higher or lower — depending on the corpus — (Volkart & Bouillon, 2023) ERs in HT than in HPE. Conclusions are similar when comparing MT and other translation modalities, with ER being either higher (Hansen & Esperança-Rodier, 2022) or lower (Martikainen & Kübler, 2016) in MTs as compared to HTs.

3.4 Statistical testing

A Shapiro-Wilk test (α = 0.05) was used to assess the normality of each feature across editorials, depending on the MT engine and prompt. Only MATTR-100 was non-normal for APE using P3, with all other scores being normally distributed.

In order to compare MT to APE and pairs of APE prompts or MT systems, we ran Holm-corrected Student's paired *t*-tests where applicable, i.e. when the normality of the differences was confirmed. In the few instances where that was not the case, a Wilcoxon signed-rank test was used. We also ran bootstrap resampling (*n* = 5,000 iterations, 95% confidence interval in all cases) to support the results of the significance tests. To compare the ER between MT and APE across prompts and underlying MT systems, we had to rely on bootstrap resampling only as there were not enough pairs for a signed-rank Wilcoxon test to establish significance.

We report effect size using Hedges' g and not Cohen's d as it is more accurate when the sample size is small (Hedges & Olkin, 2014). For reference, values of 0.2, 0.5, and 0.8 are generally considered to indicate small, medium, and large effects respectively (Cohen, 2013).

4. Results

In this section, we report the differences for each linguistic metric i) between MT and APE across prompts and the underlying MT systems; ii) between MT and APE by underlying MT system; and iii) between APE prompts across MT systems.

4.1 Lexical density

Considering only the translation stage, lexical density increased significantly between MT and APE according to a Wilcoxon signed-rank test ($p < 0.002$). When also taking the MT system underlying APE into account, Holm-corrected Student's paired t-tests indicate all prompts consistently led to a significant increase in lexical density for MTs produced using DeepL ($p < 0.01$) and GPT4o ($p < 0.02$), with a very large effect size (Hedges' $g \approx$ 2.7 and 1.8 respectively). On the other hand, there was no significant difference when APE was based on MTs from GTrans, which we attribute to their higher lexical density before the APE step. Bootstrap resampling corroborated these results.

	Lexical density		
MT system	Raw MT	APE	Δ vs raw MT
DeepL	0.5225	0.56	+7.2%$^{\circ}$
GPT-4o	0.5369	0.5525	+2.9%†
GTrans	0.5481	0.5555	+1.3%

Table 1: Mean lexical densities in raw MT and APE (all prompts) and the relative percentage difference between MT and APE, by MT system underlying APE. $^{\circ}$ *indicates significance at* $p < 0.01$ *and* † *indicates significance at* $p < 0.05$*. Holm-corrected Student's paired t-tests were used for all pairs.*

Between pairs of prompts across MT systems, the lexical density differences were largest between P1 and P3 (P1 > P3): mean difference $\approx$ 0.024 (mean = 0.568 in P1 vs 0.544 in P3) (bootstrap resampling with 95% CI [0.014, 0.035]) and large effect size (Hedges' g = 1.06). Lexical density was also significantly higher in P2 than in P3, albeit to a lesser extent (mean difference $\approx$ 0.12, Hedges' g = 0.63). There was no significant difference between P1 and P2. Thus, the starkest difference was between the least constrained prompt (P1) and the most constrained prompt (P3).

4.2 MATTR

Similarly to the case of lexical density, there was a significant increase (Holm-corrected Student's paired *t*-test, $p < 0.001$) in MATTR-100 between the raw MTs and the automatically post-edited texts across all prompts and underlying MT systems. However, when considering the MT system underlying APE, MATTR-100 significantly increased ($p < 0.02$) between the MT and APE steps only when DeepL underlay APE, while no significant difference was found when using GPT4o and GTrans according to the Holm-corrected Student's paired *t*-test. However, looking at bootstrapping results, statistical significance was achieved for TTR and MATTR-100 in all cases. Effect sizes (0.78 for GPT4o and 0.89 for GTrans) also suggest the Holm-corrected Student's paired *t*-tests failed to reach significance as they were quite conservative and the sample size was small.

	TTR	∆ vs MT	MATTR-100	∆ vs MT
MT	0.4907	/	0.7507	/
P1	0.5247	+6.9%*	0.7707	+2.7%*
P2	0.5213	+6.2%*	0.768	+2.3%*
P3	0.5087	+3.7%◦	0.7573	+0.9%◦

Table 2: Mean TTR and MATTR-100 and the relative percentage difference between MT and each APE prompt across underlying MT systems. * *indicates significance at* $p < 0.001$ *and* ◦ *indicates significance at* $p < 0.01$*. Holm-corrected Student's paired t-tests were used for all pairs except MATTR-100 (MT vs P3), for which a Wilcoxon signed-rank test was used as the assumption of the normality of the differences was not met (Shapiro-Wilk,* $\alpha = 0.05$*).*

As seen in Table 2, all prompts significantly increased TTR and MATTR-100 across MT systems, but the less constrained P1 and P2 did so to a larger extent. The results of Student's paired *t*-tests (Wilcoxon signed-rank test for TTR in P1 vs P2) mirror those obtained for lexical density: both P1 and P2 led to significantly higher TTR ($p < 0.02$) and MATTR-100 ($p < 0.001$) than P3, with large effect sizes (Hedges' $g > 0.8$), while there was no significant difference between P1 and P2. This was confirmed by bootstrapping results. Looking at Figure 1, a pattern emerges: the less constrained the prompt, the higher the lexical metrics.

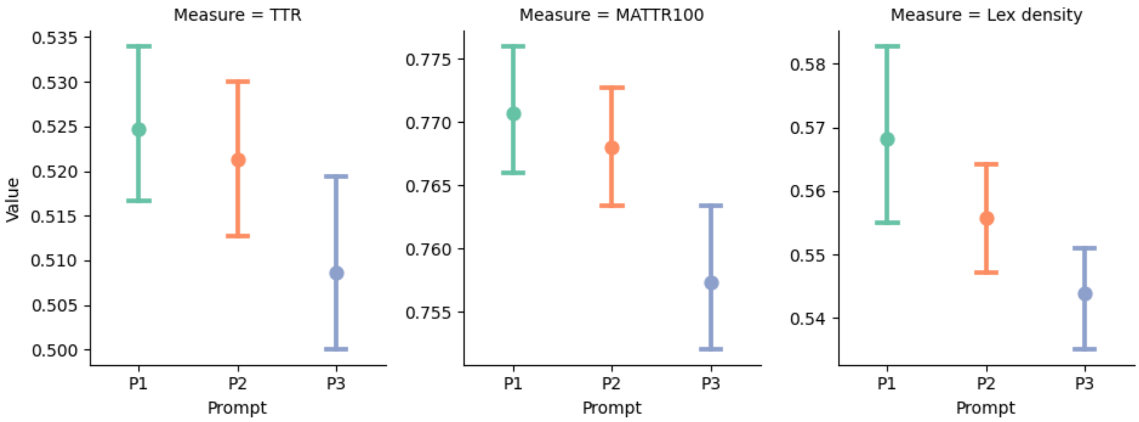

Figure 1: Lexical metrics in the automatically post-edited editorials by prompt

4.3 SACr

The Student's paired *t*-tests with Holm correction did not reveal a significant difference in normalised SACr between MT and APE across the underlying systems and prompts. However, the effect size was high (Hedges' g = 1.01) and bootstrap resampling suggested significant differences (95% CI, [0.014, 0.0637]). Sample size could therefore be the reason for the insignificant results of the *t*-test. Subsequently, even though Holm-corrected Student's paired *t*-tests and Wilcoxon signed-rank tests did not show any significant effect of the underlying MT system, bootstrap resampling showed that APE based on GPT4o led to higher normalised SACr than APE based on GTrans (95% CI, [0.014, 0.034]). The effect size was large (Hedges' g = 1.39).

	Unnormalised SACr	∆ vs MT	Normalised SACr	∆ vs MT
MT	2,172	/	0.1637	/
P1	2,990	+37.7%$^{\circ}$	0.2134	+30.4%$^{\circ}$
P2	2,790	+28.5%$^{\circ}$	0.2017	+23.2%$^{\circ}$
P3	2,408	+10.9%†	0.1859	+13.6%†

Table 3: Sums of raw SACr, mean normalised SACr, and relative percentage differences between MT and each APE prompt across underlying MT systems (15 texts). $^{\circ}$ *indicates significance at $p < 0.01$ and* † *indicates significance at $p < 0.05$. Holm-corrected Student's paired t-tests were used for all MT-APE pairs.*

As can be seen in Table 3, between prompts and across underlying MT systems, normalised SACr differences were highest between P1 and P3, but neither Holm-corrected Student's and Wilcoxon tests nor bootstrapping revealed significant differences between prompts. However, in terms of raw SACr, differences were significant in P1 vs P3 and P2 vs P3 according to both Holm-corrected Student's paired *t*-tests (p < 0.05) and bootstrapping results, but not in P1 vs P2. As with lexical metrics, the less constrained prompts led to lower normalised SACr and fewer alignment crosses (P1 > P2 > P3). We hypothesise that the lack of significance for the normalised SACr is due to the alignment method potentially flattening out differences between prompts in the normalised SACr, as described in Section 3.3.3 SACr.

4.4 ASTrED

In the editorials, normalised ASTrED scores were significantly higher in APE than in MT across underlying MT systems and prompts, according to Holm-corrected Student's paired *t*-tests ($p < 0.05$) and bootstrap resampling (95% CI, [0.043, 0.101]). According to Student's and Wilcoxon tests, the underlying MT system did not affect ASTrED scores. However, this may once again be due to the small sample size as bootstrapping showed DeepL-based APE to obtain significantly lower ASTrED than GPT4o-based and GTrans-based APE, with close-to-large and large effect sizes (mean difference = 0.028 and 0.05; Hedges' g = 0.7 and 1.74 respectively). This may be because GPT4o and GTrans MTs had slightly higher normalised ASTrED (0.618 and 0.612) than DeepL MTs (0.601). GPT4o MTs, despite having higher ASTrED, led to lower ASTrED than GTrans MTs after the APE step (mean difference = -0.022; Hedges' g = 0.68).

	Unnormalised ASTrED	Δ vs MT	Normalised ASTrED	Δ vs MT
MT	5,613	/	0.5366	/
P1	6,180	+10.1%*	0.6236	+16.2%*
P2	6,267	+11.7%*	0.6266	+16.8%*
P3	5,674	+1.1%	0.5743	+7%°

*Table 4: Sums of raw ASTrED, mean normalised ASTrED, and relative percentage differences between MT and each APE prompt across underlying MT systems (15 texts). * indicates significance at p < 0.001 and ° indicates significance at p < 0.01. Holm-corrected Student's paired t-tests were used for all MT-APE pairs except ASTrED (MT vs P3), for which a Wilcoxon signed-rank test was used as the normality of differences was violated (Shapiro-Wilk, α = 0.05).*

Similarly to the case of lexical metrics, Holm-corrected Student's paired *t*-tests revealed that P1 and P2 resulted in significantly higher normalised and unnormalised ASTrED than P3 ($p < 0.05$), with large effect sizes (Hedges' $g > 0.8$). No significant difference was found between P1 and P2 in terms of ASTrED.

4.5 Part-of-speech changes

As with ASTrED, the Holm-corrected Student's *t*-tests and bootstrapping showed that APE led to more PoS changes than MT across underlying MT systems and prompts ($p < 0.05$; 95% CI, [0.005, 0.014]). Student's paired *t*-tests and bootstrapping did not reveal any significant difference in normalised or unnormalised PoS changes according to the MT system underlying APE.

	Unnormalised PoS	Δ vs MT	Normalised PoS	Δ vs MT
MT	1,790	/	0.1371	/
P1	1,851	+3.4%	0.1403	+2.4%
P2	2,032	+13.5%*	0.1527	+11.3%*
P3	1,901	+6.2%◦	0.1465	+6.9%*

*Table 5: Sums of PoS changes, mean normalised PoS changes, and relative percentage differences between MT and each APE prompt across underlying MT systems (15 texts). * indicates significance at p < 0.001 and ◦ indicates significance at p < 0.01. Holm-corrected Student's paired t-tests were used for all pairs.*

In contrast to all other lexical and syntactic metrics, P2 achieved the highest normalised and unnormalised PoS changes, while P1 obtained the lowest scores. Differences in normalised PoS changes were significant only between P1 and P2 according to Holm-corrected *t*-tests (*p* < 0.001) and bootstrapping between the pair (95% CI, [-0.017, -0.008]).

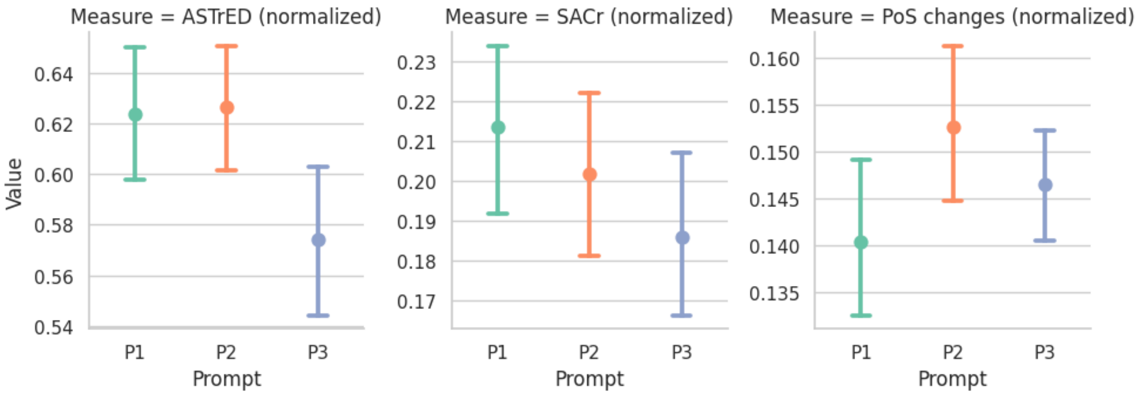

Figure 2: APE prompt effect on metrics of syntactic equivalence across MT systems underlying APE

4.6 Expanding ratio

Bootstrap resampling results showed the ER to be significantly lower in APE when compared to MT (i.e. the APE step shortened the raw MT) (95% CI, [-9.41, -4.97]), but it should be noted that the ER varied more in APE than in MT. According to Holm-corrected Student's paired *t*-tests, the MT system underlying APE did not affect the ER significantly, but bootstrapping results hinted at the output of DeepL-based APE being longer than that of GPT4o-based APE on average (95% CI, [0.91, 10.82]).

	Mean expanding ratio

MT	+14%
P1	+7.72%
P2	+11.98%
P3	+7.97%

Table 6: Mean expanding ratio in MTs and APE across MT systems

Although the mean ER was lowest in P1, Student's *t*-tests and bootstrapping results on pairs of prompts only showed a significant difference in P2 vs P3 (paired *t*-test: $p < 0.05$; bootstrapping mean difference = 3.95) because of the higher variance in ERs for P1. The least constrained prompt, P1, reduced the ER the most in comparison to the raw MTs, but the most constrained prompt, P3, lowered the ER to a similar extent. Therefore, the main takeaway is that, on average, all prompts reduced the ER significantly in the APE step.

5. Automatic evaluation

As well as calculating differences in linguistic metrics, we investigated the impact of the APE step in terms of document-level COMETKiwi (Rei et al., 2022b) scores, computed as the mean of the sentence-level scores for each text, and whether any prompt led to significantly higher COMETKiwi scores than the others. Before computing the COMETKiwi scores using `wmt22-cometkiwi-da`, we manually realigned the source editorials and the translations at the sentence level to account for sentence splits and merges during the APE step as accurately as possible. After confirming the normality of the differences between COMETKiwi in the raw MTs and in the automatically post-edited texts using a Shapiro-Wilk test, we ran Holm-corrected Student's paired *t*-tests between the raw outputs and the post-edited outputs.

	Mean COMETKiwi
MT	87.35
P1	84.79
P2	86.17
P3	86.90

Table 7: Mean COMETKiwi scores in the raw MT and by prompt across MT systems

Even though the power of the test was low, differences in COMETKiwi scores between MT and APE were borderline significant (p = 0.057). As the effect size was large (Hedges' g = 0.95), we ran bootstrap resampling, which supported significantly higher COMETKiwi scores in MT as compared to APE (95% CI, [0.59, 2.41]). This is quite surprising as the APE step increased COMETKiwi scores in Chen et al. (2024) (from English) and Raunak et al. (2023). The difference may be due to the language pair, as the English-to-French direction was not investigated in those studies. It should also be noted that despite this difference, the mean COMETKiwi score for APE exceeds 85 and still indicates high content preservation in the automatically post-edited editorials.

Regarding differences between the prompts, we used a Friedmann test and bootstrap resampling since the normality of difference scores was not met for the P2 vs P3 pair and the sample size was too low for a Wilcoxon test to assess statistical significance. The Friedmann test revealed that the prompt used significantly affected COMETKiwi scores ($p < 0.01$). Bootstrap resampling on the three pairs of prompts indicated a trend for COMETKiwi scores opposite to that of lexical metrics and SACr: the more constrained the prompt, the higher the COMETKiwi scores (P1 < P2 < P3).

Kocmi et al. (2024) calculated that a difference of 1.18 in COMETKiwi-22 scores is also perceived by human evaluators in > 95% of cases. This rule of thumb implies that the difference between P1 and P3 (95% CI, [-2.94, -1.38]) is such that human evaluators would nearly always agree that P3 is better. That may not apply to APE, though, as COMETKiwi was developed for MT evaluation and might not be able to take into account more complex changes occurring in APE (see next section).

6. Qualitative differences in MT and APE

In order to see how the differences in metrics translate into qualitative changes, here we take a closer look at two excerpts from machine-translated and automatically post-edited editorials with mean metrics scores differing by a large margin. In the tables, we report the stage (source, MT or APE), the MT system, the prompt, and the text-level metric involved.

	Source / /	**There are already signs** *of this happening* to the *EU reset*. [...] **There is also lingering suspicion** that the underlying trend of *Euroscepticism in* British political culture [...] When *even compromise* was treated as aggression **there was** less ***reason to make concessions***.
a.	MT GTrans / ASTrED = 0.530	**Il y a déjà des signes** que *cela se produit* dans le cadre de la réinitialisation de l'UE. [...] **Il existe également un soupçon persistant** selon lequel la tendance sous-jacente à l'euroscepticisme dans la culture politique britannique [...] Lorsque même un compromis était traité comme une agression, **il y avait moins de raisons de faire des concessions**.
b.	APE GTrans P3 ASTrED = 0.622	**Des signes montrent** déjà *cette dérive* dans la *réinitialisation des relations avec l'UE*. [...] **Un soupçon persistant demeure** : la tendance sous-jacente à l'euroscepticisme dans la culture politique britannique [...] Lorsque même un compromis était perçu comme une agression, **les *incitations aux concessions* disparaissaient**.
c.	APE GTrans P2 ASTrED = 0.735	**Des signes précurseurs** de *cette dynamique* **se manifestent déjà** dans le cadre de la *réinitialisation des relations avec l'UE*. [...] De plus, **une méfiance persistante subsiste** quant à l'influence durable de l'*euroscepticisme enraciné* dans la culture politique britannique [...] Lorsqu'*un simple compromis* est perçu comme une capitulation, **les incitations à faire des concessions s'amenuisent**.

Table 8: Source text, machine translation (GTrans), and two automatically post-edited versions (P3 and P2)

The editorial to which Table 8 refers features three instances of existential constructions (there is/are), which Farrell (2018) identified as an MT marker. All three are translated literally in the MT (a.), while all of them are rephrased in APE (in bold), which is in line with Farrell's (2023) results and consistent with normalised ASTrED values: APE led to fewer literal translations of the existential constructions, among others.

Results from Macken (2024) also seem to apply. While the automatically post-edited versions correct some MT errors or improve the style (italics), they also introduce undesirable corrections or errors (underlined). For instance, in b., "*incitation aux concessions*" flows better than "*raisons de faire des concessions*" but, when linked to "*disparaissaient*", does not render the meaning correctly. In c., "*euroscepticisme enraciné*" suits the editorial style well, but "*influence*" does not have the same meaning as "*tendance*". On the one hand, in both b. and c., "*réinitialisation des relations avec l'UE*" is clear, whereas the MT output "*réinitialisation de l'UE*" is unclear and misleading. On the other hand, in the last sentence of c., the simple past is replaced by the present tense and "aggression" is translated as "*capitulation*", which is not appropriate. Also, while avoiding existential constructions can be considered appropriate, "[*u*]*n soupçon persistant demeure*" (b.) and "*une méfiance persistante subsiste*" (c.) are both pleonastic as the verb replacing the existential construction already refers to a long duration and makes "*persistant*[*e*]" superfluous.

	Source / / /	Italy's divided opposition: a Five Star revolution **can help unite** the left [...] Mr Grillo's **fall from grace**, rejected by the movement he **set up to such remarkable effect**, carries **an element of pathos**. [...] A reinvigorated M5S, minus Beppe Grillo, could **begin to redress the balance**.
a.	MT DeepL / MATTR = 0.74 ASTrED = 0.52	L'opposition italienne divisée : la révolution des Cinq Étoiles **peut *aider* à unir** la gauche [...] La ***déchéance*** de M. Grillo, rejeté par le mouvement qu'il a ***créé*** **avec un effet si remarquable**, *comporte* ***un élément de*** **pathos**. [...] Un M5S ***revigoré***, **sans** Beppe Grillo, pourrait **commencer à rétablir l'équilibre**.
b.	APE DeepL P3 MATTR = 0.75 ASTrED = 0.53	L'opposition italienne divisée : la révolution des Cinq Étoiles **pourrait *contribuer* à unir** la gauche [...] Le ***déclin*** de M. Grillo, rejeté par le mouvement qu'il a lui-même ***fondé*** **avec un succès remarquable**, *contient* ***une part de*** **pathos**. [...] Un M5S ***rajeuni***, **sans** Beppe Grillo, pourrait **commencer à rétablir l'équilibre**.
c.	APE DeepL P1 MATTR = 0.79 ASTrED = 0.59	L'opposition italienne divisée : une révolution des Cinq Étoiles **pour unir** la gauche [...] La **disgrâce** de Grillo, rejeté par le mouvement qu'il avait **propulsé au sommet**, a **une dimension tragique**. [...] Un M5S **renouvelé**, **débarrassé de l'ombre** de Beppe Grillo, pourrait **contribuer à rééquilibrer le paysage politique italien**.

Table 9: Source text, machine translation (DeepL), and two automatically post-edited versions (P3 and P1)

In Table 9, a. and b. are literal translations almost throughout (bold). Macken's (2024) observation that GPT-APE tends to feature a lot of synonyms and preferential changes seems to apply in b. (italics), an editorial generated using a prompt close to the one Macken herself used. When looking at the differences between a. and b., nearly all changes are replacements with synonyms.

In contrast, c. was generated using the least constrained APE prompt and most of the changes made seem appropriate. In the title, the prepositional clause "*pour unir*" has more impact and flows better than the more literal translations in a. and b. The tone and style are better suited to editorials ("*débarrassé de l'ombre de*"). Rephrasing is more idiomatic ("*dimension tragique*", "*propulsé au sommet*"). Lengthening out "redress the balance" to "*rééquilibrer le paysage politique italien*" (back translation: "rebalance the Italian political landscape") in the last sentence seems an appropriate way to conclude the editorial. Lastly, "*renouvelé*" best renders the meaning of "reinvigorated" in context, i.e. the need for new voices in the party.

7. Discussion

In this paper, we computed automatic metrics of lexical diversity and syntactic equivalence for machine-translated and automatically post-edited editorials, before conducting quality estimation using COMETKiwi. Regarding RQ1, our results seem to show that APE leads to an increase in linguistic metrics across the board (and a decrease in ER) compared to raw MT, especially when using less constrained prompts. In particular, as reported in Macken (2024), there was a general increase in MATTR in GPT-APE as compared to MT. The differences observed between the human reference and MT in Hansen & Esperança-Rodier (2022) appear similar to those found between APE and MT, with metrics of syntactic equivalence and MATTR all increasing in APE relative to MT. This may suggest that APE brings MT outputs closer to HT in terms of linguistic metrics and supports a positive answer to RQ2. The co-occurring increase in lexical density and decrease in ER might also indicate more conciseness in APE. However, additional qualitative research, which we will leave for future work, is needed to assess whether the lower ER in APE results from omissions or from more concise rephrasing without information loss.

As for differences between prompts (RQ3), our results indicate that P1 > P2 > P3 for both lexical density and lexical diversity. To a degree, two of the three metrics of syntactic equivalence follow a similar pattern: P1 > P3 and P2 > P3, with no significant difference between P1 and P2. PoS changes seem to behave differently to ASTrED and SACr in the automatically post-edited versions, as P2 leads to the most PoS changes. This might be because SACr and ASTrED are more strongly affected by sentence reordering, while PoS changes also capture lexical transformations such as nominalisation, which do not necessarily involve an extensive reorganisation of the sentence structure and might be encouraged by the specific request for lexical and syntactic variety in P2. The most constrained prompt generally yields lower linguistic metrics, which we hypothesise might be due to additional information in the prompts hindering the editing freedom of the model, even though P2 specifically asks for lexical and syntactic variety. This mirrors the results obtained by Du et al. (2025), whose most generic prompt led to more creative GPT-generated MTs than a prompt explicitly requesting creativity.

Overall, COMETKiwi scores seem to show an inverse relationship to linguistic metrics in APE (RQ4), as the simplest prompt, which generally leads to higher lexical and syntactic metrics, obtains markedly lower COMETKiwi scores. However, at the text level, we noticed a few cases of APE in which the semantic content from one source sentence was split across two target sentences, which probably resulted in lower COMETKiwi scores as they are calculated at the sentence level. We also hypothesise that even state-of-the-art quality estimation metrics such as COMETKiwi can be biased against extensive lexical or syntactic changes possibly brought about in the APE step, even at the sentence level. In addition to the undesirable edits described in Section 6. Qualitative differences in MT and APE, this could be another reason for the lower COMETKiwi scores in the automatically post-edited texts. It could also partly explain why P1, which ranks highest for lexical metrics and SACr and statistically ties for the highest ASTrED, achieves the lowest COMETKiwi scores, while the opposite is true for P3. An example supporting that hypothesis can be found in the COMETKiwi scores for the automatically post-edited sentence from Table 9 "Mr Grillo's fall from grace, rejected by the movement he set up to such remarkable effect, carries an element of pathos": APE in response to P1, although correct and arguably more idiomatic and stylistically appropriate, results in a much lower COMETKiwi score (77.48) than APE in response to P3 (84.64), which can sound overly literal.

Kocmi & Federman (2023a) found that less constrained prompts yielded better results as part of translation evaluation using LLMs. Puppel & Borg (2024) reported similar results when using ChatGPT in creative text MT: the simplest prompt achieved the best overall quality score as measured with an adapted DQF-MQM score (Lommel, 2018), but more constrained prompts reduced fluency and style errors. This somewhat contrasts with our results, as COMETKiwi scores are lowest for P1 while the prompt seems to lead to more adequate stylistic changes in a few qualitative examples. However, LLMs may behave differently when used for MT or APE, and evaluation methods differ in both studies, preventing any fair comparison. Similarly, language pairs differ, which can substantially impact performance.

In the few qualitative examples we presented, differences in metrics of syntactic equivalence between MT and APE translated into more idiomaticity and fluency in the French editorials (RQ5). This is consistent with prior findings that APE increases naturalness and fluency (Chen et al., 2024; Kunilovskaya et al., 2024; Li et al., 2025), but a thorough, larger-scale analysis is needed to provide more conclusive evidence. Many instances of undesirable syntactic calques in MT were adapted into more natural sentences in APE, but the output still needs human attention as unidiomatic parts remain. Furthermore, as Macken (2024) concluded, while APE solved some MT errors, other changes introduced errors where the MT was correct. One could also argue that some of the replacements with synonyms have little to no added value.

8. Conclusion

In this case study, we computed metrics of lexical diversity and syntactic equivalence for editorials machine-translated from English to French and automatically post-edited, using three MT systems and three increasingly constrained prompts. We also computed COMETKiwi scores and examined some qualitative examples.

According to our results, the APE step consistently increases lexical and syntactic diversity, but prompting strategies lead to significant differences: for most linguistic metrics, the simpler the prompt, the higher the score. Our qualitative analysis shows that the increased lexical or syntactic metrics in the automatically post-edited excerpts seem to be linked to fewer syntactic calques and more appropriate word choice. This reinforces the idea that, when it comes to prompting, "less can be more". Nevertheless, it remains to be seen whether the same trends also apply to other text types and language pairs in APE.

Conversely, COMETKiwi scores are lower in APE in all the cases analysed, with less constrained prompts leading to lower scores. More qualitative research is needed to assess whether the similar behaviour of the linguistic metrics translates into similar editing patterns in APE and HPE or HT, and whether the lower COMETKiwi scores are due to lower content preservation or to a bias against lexical and syntactic variety.

As Kunilovskaya et al. (2024) and Chen et al. (2024) have pointed out, readers may prefer translations that better reflect target language patterns and are more natural, i.e. with lower levels of translationese, which emphasises the need for research in translationese mitigation. Amidst the growth of post-editing, MT, and AI on the market (ELIS, 2025), maintaining naturalness and variety to some extent may become a concern and help avoid any target language impoverishment on a broader scale (Farrell, 2018; Niu & Jiang, 2024; Volkart & Bouillon, 2024). This is all the more important since translations can be reused for training purposes, reinforcing translationese patterns.

We leave two research avenues for future work. The first, much like in Macken's (2024) case, consists of a thorough qualitative analysis of the changes made in HPE and APE, in another language pair and domain, complemented with a human evaluation of fluency and idiomaticity. The second is a comparison of HPE from MT and from APE to investigate how post-editing effort and changes made differ in both methods, as well as whether any particular method leads to more effective machine translationese mitigation while preserving content.

Our methodology has some limitations. We aimed to avoid any training of the GPT model on our source texts by working on recent editorials. Since we did not use an existing parallel corpus, no reference translation was available, which would have been interesting for comparison purposes. This is not the focus of this work, though. Also, as GPT is a black-box proprietary model that is regularly updated, there is no guarantee our results can be replicated, even using exactly the same data and prompts. Furthermore, Volkart & Bouillon (2023) have challenged the idea that post-editese features are universal. They have argued that such features may depend on many more factors (language pair and text type, among others) than just the translation mode. The texts used in this study, even though they were checked for representativeness within the text type, are all editorials. Therefore, we assume our results only apply to APE of this text type, in the English-to-French direction, using the same MT systems and prompts.

Acknowledgements

We warmly thank Prof. Pierrette Bouillon for her insightful comments and Dr Loïc De Faria Pires for the translation of the abstract into Spanish. We are grateful to our anonymous reviewers for their comprehensive feedback.

Bibliography

Alvarez-Vidal, Sergi; do Campo, Maria; Olalla-Soler, Christian; Sánchez-Gijón, Pilar (2025). Using Translation Techniques to Characterize MT Outputs. In: Bouillon, Pierrette; Gerlach, Johanna; Girletti, Sabrina; Volkart, Lise; Rubino, Raphael; Sennrich, Rico; Farinha, Ana C.; Gaido, Marco; Daems, Joke; Kenny, Dorothy; Moniz, Helena; Szoc, Sara (eds.). *Proceedings of Machine Translation Summit XX, Volume 1*, pp. 619-627. <https://aclanthology.org/volumes/2025.mtsummit-1/>. [Accessed: 20251217].

Baker, Mona (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In: Baker, Mona; Gill, Francis; Tognini-Bonelli, Elena; Sinclair, John (eds.). *Text and Technology*. Philadelphia: J. Benjamins Pub, pp. 233-250. <https://doi.org/10.1075/z.64.15bak>. [Accessed: 20251217].

Baker, Mona (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target: International Journal of Translation Studies*, v. 7 n. 2., pp. 223-243. <https://doi.org/10.1075/target.7.2.03bak>. [Accessed: 20251217].

Bestgen, Yves (2024). Diversité lexicale et longueur du texte en évaluation du langage. In: Dister, Anne; Longrée, Dominique (eds.). *Actes des 17es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 89-98. <https://perso.uclouvain.be/yves.bestgen/images/JADT24.pdf>. [Accessed: 20251217].

Briva-Iglesias, Vicent (2025). Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication. In: Bouillon, Pierrette; Gerlach, Johanna; Girletti, Sabrina; Volkart, Lise; Rubino, Raphael; Sennrich, Rico; Farinha, Ana C.; Gaido, Marco; Daems, Joke; Kenny, Dorothy; Moniz, Helena; Szoc, Sara (eds.). *Proceedings of Machine Translation Summit XX, Volume 1*, pp. 365-377. <https://aclanthology.org/volumes/2025.mtsummit-1/>. [Accessed: 20251217].

Castilho, Sheila; Resende, Natalia (2022). Post-Editese in Literary Translations. *Information*, v. 13 n. 2. <https://doi.org/10.3390/info13020066>. [Accessed: 20251217].

Castilho, Sheila; Resende, Natalia; Mitkov, Ruslan (2019). What Influences the Features of Post-editese? A Preliminary Study. *Proceedings of the Human-Informed Translation* *and Interpreting Technology Workshop (HiT-IT 2019)*, pp. 19-27. <https://doi.org/10.26615/issn.2683-0078.2019_003>. [Accessed: 20251217].

Chan, Venus; Tang, William Ko-Wai (2024). GPT and Translation: A Systematic Review. *2024 International Symposium on Educational Technology (ISET)*, pp. 59-63. <https://doi.org/10.1109/ISET61814.2024.00021>. [Accessed: 20251217].

Chen, Pinzhen; Guo, Zhicheng; Haddow, Barry; Heafield, Kenneth (2024). *Iterative Translation Refinement with Large Language Models.* <https://doi.org/10.48550/arXiv.2306.03856>. [Accessed: 20251217].

Cochrane, Guylaine (1995). Le foisonnement, phénomène complexe. *TTR : traduction, terminologie, rédaction*, v. 8 n. 2, pp. 175-193. <https://doi.org/10.7202/037222ar>. [Accessed: 20251217].

Cohen, Jacob (2013). *Statistical Power Analysis for the Behavioral Sciences.* NewYork: Routledge. <https://doi.org/10.4324/9780203771587>. [Accessed: 20251217].

Covington, Michael A.; McFall, Joe D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, v. 17 n. 2, pp. 94-100. <https://doi.org/10.1080/09296171003643098>. [Accessed: 20251217].

Čulo, Oliver; Nitzke, Jean (2016). Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation. *Baltic J. Modern Computing*, v. 4 n. 2, pp. 106-114. <https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/4_2_4_Culo.pdf>. [Accessed: 20251217].

Daems, Joke; De Clercq, Orphée; Macken, Lieve (2017). Translationese and Post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series: Themes in Translation Studies*. v. 16. <https://doi.org/10.52034/lanstts.v16i0.434>. [Accessed: 20251217].

De Clercq, Orphée; De Sutter, Gert; Loock, Rudy; Cappelle, Bert; Plevoets, Koen (2021). Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French. *Translation Quarterly,* n. 101, pp. 21-45. <http://hdl.handle.net/1854/LU-8725139>. [Accessed: 20251217].

Directorate-General for Translation (2025). *European Language Industry Survey 2025.* <https://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf>. [Accessed: 20251217].

Do Carmo, Félix; Shterionov, Dimitar; Moorkens, Joss; Wagner, Joachim; Hossari, Murhaf; Paquin, Eric; Schmidtke, Dag; Groves, Declan; Way, Andy (2021). A review of the state-of-the-art in automatic post-editing. *Machine Translation*, v. 35 n. 2, pp. 101-143. <https://doi.org/10.1007/s10590-020-09252-y>. [Accessed: 20251217].

Dou, Zi-Yi; Neubig, Graham (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In: Merlo, Paola; Tiedemann, Jorg; Tsarfaty, Reut (eds.). *Proceedings of the 16th Conference of the European Chapter of the Association for*

Computational Linguistics: Main Volume, pp. 2112-2128. <https://doi.org/10.18653/v1/2021.eacl-main.181>. [Accessed: 20251217].

Du, Shuxiang; Guerberof Arenas, Ana; Toral, Antonio; Gerrits, Kyo; Borillo, Josep Marco (2025). Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish. In: Bouillon, Pierrette; Gerlach, Johanna; Girletti, Sabrina; Volkart, Lise; Rubino, Raphael; Sennrich, Rico; Farinha, Ana C.; Gaido, Marco; Daems, Joke; Kenny, Dorothy; Moniz, Helena; Szoc, Sara (eds.). *Proceedings of Machine Translation Summit XX, Volume 1*, pp. 578-591. <https://aclanthology.org/volumes/2025.mtsummit-1/>. [Accessed: 20251217].

Farrell, Michael (2018). Machine Translation Markers in Post-Edited Machine Translation Output. *Proceedings of the 40th Conference Translating and the Computer*, pp. 50-59. <https://apeiron.iulm.it/handle/10808/47325>. [Accessed: 20251217].

Farrell, Michael (2023a). Current evidence of post-editese: differences between post-edited neural machine translation output and human translation revealed through human evaluation. *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2023)*, pp. 52-63. <https://doi.org/10.26615/issn.2683-0078.2023_005>. [Accessed: 20251217].

Farrell, Michael (2023b). Preliminary evaluation of ChatGPT as a machine translation engine and as an automatic post-editor of raw machine translation output from other machine translation engines. *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2023)*, pp. 108-113. <https://doi.org/10.26615/issn.2683-0078.2023_009>. [Accessed: 20251217].

Fernandes, Patrick; Deutsch, Daniel; Finkelstein, Mara; Riley, Parker; Martins, André; Neubig, Graham; Garg, Ankush; Clark, Jonathan; Freitag, Markus; Firat, Orhan (2023). The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In: Koehn, Philipp; Haddow, Barry; Kocmi, Tom; Monz, Christof (eds.). *Proceedings of the Eighth Conference on Machine Translation*, pp. 1066-1083. <https://doi.org/10.18653/v1/2023.wmt-1.100>. [Accessed: 20251217].

Frawley, William (1984). Prolegomenon to a theory of translation. In: Frawley, William (ed.). *Translation: Literary, Linguistic and Philosophical Perspectives*. Newark: University of Delaware Press, pp. 159-175.

Gellerstam, Martin (1986). Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, v. 1, pp. 88-95.

Guo, Jiaxin; Yang, Hao; Li, Zongyao; Wei, Daimeng; Shang, Hengchao; Chen, Xiaoyu (2024). *A Novel Paradigm Boosting Translation Capabilities of Large Language Models*. <https://doi.org/10.48550/arXiv.2403.11430>. [Accessed: 20251217].

Hansen, Damien; Esperança-Rodier, Emmanuelle (2022). Human-Adapted MT for Literary Texts: Reality or Fantasy? *NeTTT 2022,* pp. 178-190. <https://hal.science/hal-04038025>. [Accessed: 20251217].

He, Zhiwei; Liang, Tian; Jiao, Wenxiang; Zhang, Zhuosheng; Yang, Yujiu; Wang, Rui; Tu, Zhaopeng; Shi, Shuming; Wang, Xing (2024). Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, v. 12, pp. 229-246. <https://doi.org/10.1162/tacl_a_00642>. [Accessed: 20251217].

Hedges, Larry V.; Olkin, Ingram (2014). *Statistical Methods for Meta-Analysis*. Saint-Louis: Elsevier Science.

Hendy, Amr; Abdelrehim, Mohamed; Sharaf, Amr; Raunak, Vikas; Gabr, Mohamed; Matsushita, Hitokazu; Kim, Young Jin; Afify, Mohamed; Awadalla, Hany Hassan (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.* <https://doi.org/10.48550/arXiv.2302.09210>. [Accessed: 20251217].

Jiao, Wenxiang; Wang, Wenxuan; Huang, Jen-Tse; Wang, Xing; Shi, Shuming; Tu, Zhaopeng (2023). *Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine.* <https://doi.org/10.48550/arXiv.2301.08745>. [Accessed: 20251217].

Jiménez-Crespo, Miguel A. (2023). “Translationese” (and “post-editese”?) no more: on importing fuzzy conceptual tools from Translation Studies in MT research. In: Nurminen, Mary; Brenner, Judith; Koponen, Maarit; Latomaa, Sirkku; Mikhailov, Mikhail; Schierl, Frederike; Ranasinghe, Tharindu; Vanmassenhove, Eva; Alvarez Vidal, Sergi; Aranberri, Nora; Nunziatini, Mara; Parra Escartín, Carla; Forcada, Mikel; Popovic, Maja; Scarton, Carolina; Moniz, Helena (eds.). *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 261-268 <https://aclanthology.org/2023.eamt-1.25/>. [Accessed: 20251217].

Johansson, Victoria (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. In: Lund University, Department of Linguistics and Phonetics (eds.). *Working Papers*, v. 53., pp. 61-79.

Ki, Dayeon; Carpuat, Marine (2024). Guiding Large Language Models to Post-Edit Machine Translation with Error Annotations. In: Duh, Kevin; Gomez, Helena; Bethard, Steven (eds.). *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4253-4273. <https://doi.org/10.18653/v1/2024.findings-naacl.265>. [Accessed: 20251217].

Kocmi, Tom; Avramidis, Eleftherios; Bawden, Rachel; Bojar, Ondřej; Dvorkovich, Anton; Federmann, Christian; Fishel, Mark; Freitag, Markus; Gowda, Thamme; Grundkiewicz, Roman; Haddow, Barry; Karpinska, Marzena; Koehn, Philipp; Marie, Benjamin; Monz, Christof; Murray, Kenton; Nagata, Masaaki; Popel, Martin; Popović, Maja; Shmatova, Mariya; Steingrímsson, Steinthór; Zouhar, Vilém (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In: Haddow, Barry; Kocmi, Tom; Koehn, Philipp; Monz, Christof (eds.). *Proceedings of the Ninth Conference on Machine Translation*, pp. 1-46. <https://doi.org/10.18653/v1/2024.wmt-1.1>. [Accessed: 20251217].

Kocmi, Tom; Avramidis, Eleftherios; Bawden, Rachel; Bojar, Ondřej; Dvorkovich, Anton; Federmann, Christian; Fishel, Mark; Freitag, Markus; Gowda, Thamme; Grundkiewicz, Roman; Haddow, Barry; Koehn, Philipp; Marie, Benjamin; Monz, Christof; Morishita, Makoto; Murray, Kenton; Nagata, Masaaki; Nakazawa, Toshiaki; Popel, Martin; Popović, Maja; Shmatova, Mariya (2023). Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In: Koehn, Philipp; Haddow, Barry; Kocmi, Tom; Monz, Christof (eds.). *Proceedings of the Eighth Conference on Machine Translation*, pp. 1-42. <https://doi.org/10.18653/v1/2023.wmt-1.1>. [Accessed: 20251217].

Kocmi, Tom; Zouhar, Vilém; Federmann, Christian; Post, Matt (2024). Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. In: Ku, Lun-Wei; Martins, Andre; Srikumar, Vivek (eds.). *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1999-2014. <https://doi.org/10.18653/v1/2024.acl-long.110>. [Accessed: 20251217].

Kunilovskaya, Maria; Dutta Chowdhury, Koel; Przybyl, Heike; España-Bonet, Cristina; Genabith, Josef (2024). Mitigating Translationese with GPT-4: Strategies and Performance. In: Scarton, Carolina; Prescott, Charlotte; Bayliss, Chris; Oakley, Chris; Wright, Joanna; Wrigley, Stuart; Song, Xingyi; Gow-Smith, Edward; Bawden, Rachel; Sánchez-Cartagena, Víctor M.; Cadwell, Patrick; Lapshinova-Koltunski, Ekaterina; Cabarrão, Vera; Chatzitheodorou, Konstantinos; Nurminen, Mary; Kanojia, Diptesh; Moniz, Helena. *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pp. 411-430. <https://aclanthology.org/2024.eamt-1.35/>. [Accessed: 20251217].

Laviosa, Sara (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*: *Journal des traducteurs = Translator's Journal*. v. 43 n. 4, pp. 557-570. <https://doi.org/10.7202/003425ar>. [Accessed: 20251217].

Laviosa-Braithwaite, Sara (1996). Investigating Simplification in English Comparable Corpus of Newspaper Articles. In: Kinga, Klaudy; Janos, Kohn (eds.). *Transferre necesse est: Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting*, pp. 531-540. [Accessed: 20251217].

Li, Yafu; Zhang, Ronghao; Wang, Zhilin; Zhang, Huajian; Cui, Leyang; Yin, Yongjing; Xiao, Tong; Zhang, Yue (2025). *Lost in Literalism: How Supervised Training Shapes Translationese in LLMs*. <https://doi.org/10.48550/arXiv.2503.04369>. [Accessed: 20251217].

Loock, Rudy (2018). Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. *Meta: Journal des traducteurs = Translator's Journal* v. 63 n. 3, pp. 785-805. <https://doi.org/10.7202/1060173ar>. [Accessed: 20251217].

Macken, Lieve (2024). Machine translation meets large language models: evaluating ChatGPT's ability to automatically post-edit literary texts. In: Vanroy, Bram; Lefer, Marie-Aude; Macken, Lieve; Ruffo, Paola (eds.). *Proceedings of the First Workshop on*

Creative-text Translation and Technology, pp. 71-87. <https://aclanthology.org/2024.ctt-1.7/>. [Accessed: 20251217].

Marques, Francisco Paulo Jamil; Mont'Alverne, Camila (2021). What are newspaper editorials interested in? Understanding the idea of criteria of editorial-worthiness. *Journalism*. v. 22 n. 7., pp. 1812-1830. <https://doi.org/10.1177/1464884919828503>. [Accessed: 20251217].

Martikainen, Hanna; Kübler, Natalie (2016). Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions. *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*, n. 27. <https://doi.org/10.4000/ilcea.3863>. [Accessed: 20251217].

McNamara, Danielle S.; Graesser, Arthur C.; McCarthy, Philip M.; Cai, Zhiqiang (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix.* New York: Cambrigde University Press. <https://doi.org/10.1017/CBO9780511894664>. [Accessed: 20251217].

Moslem, Yasmin; Haque, Rejwanul; Kelleher, John D.; Way, Andy (2023). *Adaptive Machine Translation with Large Language Models.* arXiv:2301.13294v3. <https://doi.org/10.48550/arXiv.2301.13294>. [Accessed: 20251217].

Niu, Jiang; Jiang, Yue (2024). Does simplification hold true for machine translations? A corpus-based analysis of lexical diversity in text varieties across genres. *Humanities and Social Sciences Communications*. v. 11 n. 1. <https://doi.org/10.1057/s41599-024-02986-7>. [Accessed: 20251217].

OpenAI, Achiam, Josh; Adler, Steven; Agarwal, Sandhini; … Zoph, Barret (2024). *GPT-4 Technical Report*. arXiv.2303.08774v6. <https://doi.org/10.48550/arXiv.2303.08774>. [Accessed: 20251217].

Peng, Keqin; Ding, Liang; Zhong, Qihuang; Shen, Li; Liu, Xuebo; Zhang, Min; Ouyang, Yuanxin; Tao, Dacheng (2023). Towards Making the Most of ChatGPT for Machine Translation. <https://doi.org/10.48550/arXiv.2303.13780>. [Accessed: 20251217].

Poibeau, Thierry (2022). On "Human Parity" and "Super Human Performance" in Machine Translation Evaluation. In: Calzolari, Nicoletta; Béchet, Frédéric; Blache, Philippe; Choukri, Khalid; Cieri, Christopher; Declerck, Thierry; Goggi, Sara; Isahara, Hitoshi; Maegaard, Bente; Mariani, Joseph; Mazo, Hélène; Odijk, Jan; Piperidis, Stelios (eds.). *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6018-6023. <https://aclanthology.org/2022.lrec-1.647/>. [Accessed: 20251217].

Puppel, Melissa; Borg, Claudine (2024). Evaluating ChatGPT's Performance in Creative Text Translation for Communication: A Case Study from English into German. *Media and Intercultural Communication: A Multidisciplinary Journal,* v. 3, n. 1 (March). *1*<https://doi.org/10.22034/mic.2024.480506.1023>. [Accessed: 20251217].

Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason; Manning, Christopher D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: Celikyilmaz, Asli; Wen, Tsung-Hsien (eds.). *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 101-108. <https://doi.org/10.18653/v1/2020.acl-demos.14>. [Accessed: 20251217].

Raunak, Vikas; Menezes, Arul; Post, Matt; Hassan, Hany (2023). Do GPTs Produce Less Literal Translations? In: Rogers, Anna; Boyd-Graber, Jordan; Okazaki, Naoaki (eds.). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1041-1050. <https://doi.org/10.18653/v1/2023.acl-short.90>. [Accessed: 20251217].

Raunak, Vikas; Sharaf, Amr; Wang, Yiren; Awadallah, Hany Hassan; Menezes, Arul (2023). Leveraging GPT-4 for Automatic Translation Post-Editing. In: Bouamor, Houda; Pino, Juan; Bali, Kalika (eds.). *Findings of the Association for Computational Linguistics: EMNLP 2023.* <10.18653/v1/2023.findings-emnlp.804>. [Accessed: 20251217].

Rei, R.; Stewart, C.; Farinha, A. C.; Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In: Webber, Bonnie; Cohn, Trevor; He, Yulan; Liu, Yang (eds.). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685-2702. <https://doi.org/10.18653/v1/2020.emnlp-main.213>. [Accessed: 20251217].

Rei, Ricardo; C. de Souza, José G.; Alves, Duarte; Zerva, Chrysoula; Farinha, Ana C.; Glushkova, Taisiya; Lavie, Alon; Coheur, Luisa; Martins, André F. T. (2022a). COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In: Koehn, Philipp; Barrault, Loïc; Bojar, Ondřej; Bougares, Fethi; Chatterjee, Rajen; *et al* (eds.). *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578-585. <https://aclanthology.org/2022.wmt-1.52/>. [Accessed: 20251217].

Rei, Ricardo; Treviso, Marcos; Guerreiro, Nuno M.; Zerva, Chrysoula; Farinha, Ana C.; Maroti, Christine; C. de Souza, José G.; Glushkova, Taisiya; Alves, Duarte; Coheur, Luisa; Lavie, Alon; Martins, André F. T. (2022b). CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In: Koehn, Philipp; Barrault, Loïc; Bojar, Ondřej; Bougares, Fethi; Chatterjee; *et al.* (eds.). *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634-645. <https://aclanthology.org/2022.wmt-1.60/>. [Accessed: 20251217].

Schumacher, Perrine (2025). Exploration des répercussions de la TA neuronale sur la langue cible après post-édition en contexte d'apprentissage : qu'en est-il du post-editese ? *Langages*. v. 237 n. 1., pp. 109-130. <https://doi.org/10.3917/lang.237.0109>. [Accessed: 20251217].

Toral, Antonio (2019). Post-editese: an Exacerbated Translationese. In: Forcada, Mikel; Way, Andy; Haddow, Barry; Sennrich, Rico (eds.). *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 273-281. <https://aclanthology.org/W19-6627/>. [Accessed: 20251217]..

Vanmassenhove, Eva; Shterionov, Dimitar; Gwilliam, Matthew (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In: Merlo, Paola; Tiedemann, Jorg; Tsarfaty, Reut (eds.). *Proceedings of* *the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2203-2213. <https://doi.org/10.18653/v1/2021.eacl-main.188>. [Accessed: 20251217].

Vanmassenhove, Eva; Shterionov, Dimitar; Way, Andy (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In: Forcada, Mikel; Way, Andy; Haddow, Barry; Sennrich, Rico (eds.). *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 222-232. <https://aclanthology.org/W19-6622/>. [Accessed: 20251217].

Vanroy, Bram; De Clercq, Orphée; Tezcan, Arda; Daems, Joke; Macken, Lieve (2021). Metrics of syntactic equivalence to assess translation difficulty. In *Explorations in empirical translation process research*. v. 3, pp. 259-294. <https://doi.org/10.1007/978-3-030-69777-8_10>. [Accessed: 20251217].

Vanroy, Bram; Tezcan, Arda; Macken, Lieve (2019). Predicting syntactic equivalence between source and target sentences. *Computational Linguistics in the Netherlands Journal*. v. 9, pp. 101-116. <https://clinjournal.org/clinj/article/view/95>. [Accessed: 20251217].

Vilar, David; Freitag, Markus; Cherry, Colin; Luo, Jiaming; Ratnakar, Viresh; Foster, George (2023). *Prompting PaLM for Translation: Assessing Strategies and Performance*. arXiv:2211.09102v3. <https://doi.org/10.48550/arXiv.2211.09102>. [Accessed: 20251217].

Volkart, Lise; Bouillon, Pierrette (2022). Studying Post-Editese in a Professional Context: A Pilot Study. In: Moniz, Helena; Macken, Lieve; Rufener, Andrew; Barrault, Loïc; Costa-jussà, Marta R.; Declercq, Christophe; Koponen, Maarit; Kemp, Ellie; Pilos, Spyridon; Forcada, Mikel L.; Scarton, Carolina; Van den Bogaert, Joachim; Daems, Joke; Tezcan, Arda; Vanroy, Bram; Fonteyne, Margot (eds.). *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 71-79. <https://aclanthology.org/2022.eamt-1.10/>. [Accessed: 20251217].

Volkart, Lise; Bouillon, Pierrette (2023). Are post-editese features really universal? In: Orăsan, Constantin; Mitkov, Ruslan; Corpas Pastor, Gloria; Monti, Johanna (eds.). *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*, pp. 294-304. <https://doi.org/10.26615/issn.2683-0078.2023_025>. [Accessed: 20251217].

Volkart, Lise; Bouillon, Pierrette (2024). Post-editors as Gatekeepers of Lexical and Syntactic Diversity: Comparative Analysis of Human Translation and Post-editing in Professional Settings. In: Scarton, Carolina; Prescott, Charlotte; Bayliss, Chris; Oakley, Chris; Wright, Joanna; Wrigley, Stuart; Song, Xingyi; Gow-Smith, Edward; Bawden, Rachel; Sánchez-Cartagena, Víctor M.; Cadwell, Patrick; Lapshinova-Koltunski, Ekaterina; Cabarrão, Vera; Chatzitheodorou, Konstantinos; Nurminen, Mary; Kanojia, Diptesh; Moniz, Helena. *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, pp. 387-395. <https://aclanthology.org/2024.eamt-1.33/>. [Accessed: 20251217].

Wang, Longyue; Lyu, Chenyang; Ji, Tianbo; Zhang, Zhirui; Yu, Dian; Shi, Shuming; Tu, Zhaopeng (2023). *Document-Level Machine Translation with Large Language Models.* arXiv:2304.02210v2. <https://doi.org/10.48550/arXiv.2304.02210>. [Accessed: 20251217].

Wei, Jason; Wang, Xuezhi; Schuurmans, Dale; Bosma, Maarten; Ichter, Brian; Xia, Fei; Chi, Ed; Le, Quoc; Zhou, Denny (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. <https://doi.org/10.48550/arXiv.2201.11903>. ArXiv:2201.11903v6. [Accessed: 20251217].

Xu, Haoran; Kim, Young Jin; Sharaf, Amr; Awadalla, Hany Hassan (2024). *A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models*. arXiv:2308.01391v2. <https://doi.org/10.48550/arXiv.2309.11674>. [Accessed: 20251217].

Yamada, Masaru (2024). *Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability.* arXiv:2308.01391v2. <https://doi.org/10.48550/arXiv.2308.01391>. [Accessed: 20251217].

Zhang, Biao; Haddow, Barry; Birch, Alexandra (2023). *Prompting Large Language Model for Machine Translation: A Case Study*. arXiv:2301.07069v2. <https://doi.org/10.48550/arXiv.2301.07069>. [Accessed: 20251217].

Zhu, Wenhao; Liu, Hongyi; Dong, Qingxui; Xu, Jingjing; Huang, Shujian; Kong, Lingpeng; Chen, Jiajun; Li, Lei (2024). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In: Duh, Kevin; Gomez, Helena; Bethard, Steven (eds.). *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765-2781. <https://doi.org/10.18653/v1/2024.findings-naacl.176>. [Accessed: 20251217].

Appendix

Full corpus data, computed as described in Section 3.1 Source corpus and selected source texts. The selected editorials are in bold.

Text ID	Word count	TTR	MATTR-100	Lexical density	Syntactic simplicity
1O-EN GV	589	0.58	0.78	0.63	0.25
1O-EN IV	735	0.45	0.69	0.50	-0.49
1O-EN TV	501	0.49	0.75	0.55	0.10
2O-EN GV	575	0.54	0.78	0.62	-0.57
2O-EN IV	574	0.46	0.72	0.52	-0.15
2O-EN TV	596	0.55	0.76	0.54	-0.66
3O-EN GV	**592**	**0.54**	**0.76**	**0.61**	**-0.29**
3O-EN IV	591	0.51	0.73	0.53	-0.40
3O-EN TV	488	0.49	0.73	0.56	-0.16
4O-EN GV	592	0.60	0.80	0.64	0.47
4O-EN IV	544	0.47	0.71	0.54	-0.87
4O-EN TV	466	0.58	0.75	0.60	0.24
5O-EN GV	611	0.55	0.79	0.59	-0.06
5O-EN IV	727	0.45	0.74	0.53	-0.77
5O-EN TV	471	0.53	0.74	0.55	-0.80
6O-EN GV	592	0.53	0.78	0.57	0.20
6O-EN IV	750	0.47	0.74	0.50	-0.89
6O-EN TV	479	0.57	0.77	0.58	-0.86
7O-EN GV	**569**	**0.54**	**0.78**	**0.57**	**-0.06**
7O-EN IV	745	0.52	0.78	0.55	-0.66
7O-EN TV	462	0.55	0.78	0.58	0.17
8O-EN GV	598	0.56	0.79	0.61	-0.10

8O-EN IV	592	0.49	0.73	0.51	-0.92
8O-EN TV	458	0.57	0.79	0.59	-0.39
9O-EN GV	573	0.58	0.79	0.61	0.46
9O-EN IV	765	0.45	0.72	0.53	-0.13
9O-EN TV	507	0.51	0.74	0.57	-0.12
10O-EN GV	574	0.56	0.79	0.61	-0.81
10O-EN IV	626	0.47	0.72	0.49	-0.97
10O-EN TV	448	0.54	0.78	0.59	-0.23
11O-EN GV	**581**	**0.52**	**0.76**	**0.61**	**-0.43**
11O-EN IV	579	0.49	0.74	0.53	-0.18
11O-EN TV	721	0.52	0.78	0.54	0.11
12O-EN GV	**592**	**0.54**	**0.77**	**0.59**	**-0.52**
12O-EN IV	564	0.56	0.78	0.61	0.28
12O-EN TV	489	0.56	0.75	0.56	-0.11
13O-EN GV	547	0.59	0.82	0.66	0.58
13O-EN IV	602	0.46	0.76	0.54	-0.51
13O-EN TV	493	0.54	0.80	0.59	0.64
14O-EN GV	594	0.54	0.77	0.59	-0.12
14O-EN IV	602	0.47	0.74	0.55	-0.32
14O-EN TV	476	0.51	0.73	0.57	-0.88
15O-EN GV	568	0.57	0.79	0.66	-0.98
15O-EN IV	789	0.50	0.78	0.57	-0.87
15O-EN TV	470	0.53	0.77	0.59	-0.17
16O-EN GV	**601**	**0.55**	**0.77**	**0.61**	**-0.13**
16O-EN IV	605	0.50	0.73	0.53	-0.48
16O-EN TV	480	0.45	0.75	0.54	-0.70

Full prompt 3 (P3) used for the automatic post-editing:

System prompt:

You are a native French speaker with a good working knowledge of English. You are also an experienced post-editor of editorial translations from English into French.

You know that every translation is a compromise between two goals: faithfulness to the meaning of the source text and faithfulness to the style of the original author.

"Faithfulness to the meaning of the source text" means that the meaning of the target text must not differ from that of the source text. In other words, no meaningful elements of the source text should be arbitrarily omitted, added or distorted in the French translation.

Therefore, you will notice any deviations in the French translations, including the following issues that make the given French translation not optimal:

1. Meaningful words in the English source text that are not rendered in the French translation
2. Meaningful words in the French translation that are not supported in the input
3. Words in the French translation that do not convey the specific meaning of the corresponding word in the English source text

You will identify and correct the above problems in the French translation, if present, in a way that improves the fluency of the translation.

"Faithfulness to the style of the original author" in editorial translation implies that you sometimes have to think creatively to find solutions that are out of the ordinary, that go beyond the routine, while preserving the argumentative intentions or effects that are evident in the source text.

You will identify any stylistic deviations in the French translation, if present, in a way that improves the style of the translation.

Furthermore, as an expert translation post-editor, you will make sure that the following principles are followed when making improvements to the French translation:

1. Do not edit the translation if the translation is faithful to the meaning of the source text and faithful to the style of the original author.
2. If the translation is very poor, generate an improved translation from scratch.
3. No corrections are made that add words or phrases in the translation that are not supported in the English source text.
4. Capitalization in the translation strictly follows capitalization in the input.
5. The translation contains the appropriate articles and determiners to follow the specifics in the input.
6. No meaningful words are left untranslated in the final, improved translation.
7. Do not add any extraneous words, phrases, clauses or sentences to the translation that are not supported by the input.

8. If the input begins with a non-capitalized word, the translation will begin with a non-capitalized word.
9. Do not add end punctuations or full stops if they are not present in the source text.
10. Do not assume that the source text contains “typos”; always err on the side of assuming that the presented input words are not typos.
11. If the translation fails to convey the meaning of a large part of the input sentence, you include the translation for the missing part.

User prompt:

As an expert translation post editor, your task is to improve the French translation for the below English text.

English text:
[SRC]

French translation:
[MT]

Say "Improved Translation:". Then output the French translation with proposed improvements that increase the faithfulness, fluency and style of the translation.