BlAInded by Fluency: How Idiomatic Machine Translation Outputs Affect Student Post-Editors' Edit Types

Valentin Scourneau
University of Mons, Belgium & Polytechnic University of Hauts-de-France, France
valentin.scourneau@umons.ac.be & valentin.scourneau@uphf.fr

Loïc De Faria Pires
University of Mons, Belgium
loic.defariapires@umons.ac.be

Abstract

This study explores the influence of two prompting strategies on the lexical and syntactic metrics of the large language model-based (LLM-based) machine translations (MTs) of a corpus of 18 British editorials into French as well as their impact on the edit types made by Master's translation students post-editing (PE) from a representative editorial of the corpus, as evaluated using the machine translation post-editing annotation system (MTPEAS) taxonomy. Quantitatively, the prompt specifically requesting more syntactic and lexical variety leads to significantly higher syntactic and lexical metrics scores in the MTs, but differences remain significant only for lexical metrics in the post-edited versions of the representative editorial. Qualitatively, we show that students post-editing from an MT featuring more idiomatic rephrasings and fewer syntactic calques (as opposed to an MT that is structurally closer to the source text) seem to make fewer edits overall, leave more MT errors unaddressed, and make fewer successful edits.

1 Introduction

The translation performance of large language models (LLMs) has increased tremendously in the last few years, especially in high-resource language pairs (Kocmi et al., 2024; Kocmi et al., 2025). However, despite improvements in accuracy and fluency (Alvarez-Vidal et al., 2025), non-English LLM-based MT outputs (MTs generated using an LLM) may often sound robotic or unnatural (Chen et al., 2024; Kunilovskaya et al., 2024; Kocmi et al., 2025), a phenomenon known as "translationese" (Gellerstam, 1986). Related effects had previously been shown in neural MT, which can result in decreased lexical variety as well as higher syntactic equivalence with the source text (Hansen and Esperança-Rodier, 2022; Loock, 2018; Toral, 2019; Vanmassenhove et al., 2019; Vanmassenhove et al., 2021). The latter two results have also been observed in PE in the form of "post-editese", but results are less clear-cut in that modality, with cases of conflicting results, sometimes even in a same study (Castilho et al., 2019; Castilho and Resende, 2022; Daems et al., 2017; Farrell, 2023a; Toral, 2019; Volkart et al., 2022; Volkart and Bouillon, 2023; Volkart and Bouillon, 2024). Also, since priming from the MT output seems to occur in PE (Bangalore et al., 2015; Green et al., 2013), structures from the source text are ultimately more likely to be mirrored in target post-edited texts. When it comes to translation students in particular, previous research has revealed several interesting trends. Volkart et al. (2022) found that students leave around half the MT mistakes unaddressed while post-editing and that when they are not forced to consult the source text first, they almost do monolingual PE instead of bilingual PE, disregarding the source text. More recently, Schumacher (2025) also showed that texts post-edited by students had lower lexical variety and higher syntactic equivalence with the source text than those translated from scratch – hinting at post-editese –, while lexical density and average sentence length did not differ significantly.

As the use of MT, PE, and artifical intelligence (AI) is rapidly growing in the language industry

(ELIS, 2026), researchers have looked into means of increasing naturalness as well as lexical and syntactic variety in outputs from pipelines relying on LLMs, such as LLM-based MT and automatic PE methods (Chen et al., 2024; Li et al., 2025; Scourneau, 2025). That research avenue is gaining momentum as it is increasingly seen as a means to avoid target language impoverishment at a larger scale (Farrell, 2018; Volkart and Bouillon, 2024).

However, it is not clear how PE lexical and syntactic variety and the type of edits made by post-editors are influenced by the level of lexical variety in the MT and that of syntactic equivalence between the source text and the MT underlying the PE process. In this case study, we aim to gain insight into those questions. To that end, we gathered 18 editorials from online British national newspapers. Two LLM-based MTs (GPT-5.0) into French were then generated for each editorial using a straightforward prompt and a prompt designed to counteract MT artifacts, i.e. reduced lexical variety and higher syntactic equivalence. The effectiveness of the prompts in doing so was assessed with a series of lexical and syntactic metrics. As we could only have one editorial post-edited by the students due to class organization constraints, we selected the most representative according to a series of linguistic metrics computed on the source editorials. Twelve Master's students in translation were randomly assigned one of the two machine-translated versions of the selected editorial and asked to post-edit it. As a final step, the post-edited versions were evaluated qualitatively and quantitatively, using the MTPEAS taxonomy (Bodart et al., 2024) and linguistic metrics of lexical variety and syntactic equivalence.

2 Methodology

In this section, we describe 2.1) how the source corpus was compiled and how the editorial to be post-edited was selected; 2.2) the GPT version and the prompts used for the MT step as well as the PE setting; 2.3) the metrics and statistical testing methods employed for the quantitative assessment of the MTs and post-edited versions; and 2.4) the setting and the taxonomy used for the qualitative assessment of the post-edited editorials.

2.1 Source corpus and source text selection

We created a small corpus of 18 editorials published in British national newspapers (six from each of *The Guardian*, *The Telegraph*, and *The Independent*), from which we selected the most representative editorial to be post-edited by the students. We chose editorials published between mid-July and mid-September 2025 in order to avoid any training of the GPT model on the texts. The editorials were selected based on both their length (± 400–700 words) and their topic, i.e. recent geopolitical events or discussions, avoiding UK-specific events. This was to ensure the student post-editors would both manage to post-edit the attributed editorial during a two-hour class and have at least general knowledge of the matter at hand. Editorials are especially interesting to investigate as part of a study examining lexical diversity and syntactic equivalence, since they are not purely informative as opposed to news pieces (Poibeau, 2022) and are considered to allow more freedom in the writing style (Marques and Mont'Alverne, 2021).

The English editorials were left untouched, apart from the removal of "The Guardian view on" in editorials from *The Guardian* so as not to skew lexical metrics. They were saved in UTF-8 .txt format after punctuation was standardised across all texts.

For each editorial, we then computed the word count, the lexical density, the TTR, and the MATTR-50 (Covington and McFall, 2010) using the Python SpaCy library[1], as well as the mean sentence length, the syntactic simplicity, and the sentence syntax similarity using Coh-Metrix (McNamara et al., 2014). In order to select the most representative editorial in the corpus, we computed the Mahalanobis distance for the 18 editorials based on the linguistic measures above. Since some measures (e.g. TTR and MATTR-50) are correlated, the Mahalanobis distance was chosen over the Euclidean distance, as the former accounts for the correlations between the measures. The first two Mahalanobis distances were extremely close, so we qualitatively selected the one editorial that was less factual and whose topic was likely better known by the students, i.e. the future of the European Union.

2.2 Machine translation and post-editing

LLM-based MT The 18 editorials from the corpus, including the selected article, were machine translated in late September 2025 using the paid

[1] https://spacy.io

version of GPT-5.0. We relied on a temporary chat, in automatic reflection mode and with memory disabled so as to minimize biases. Two prompts were used to generate the MTs. Both of them were followed by a line break and the full editorial. The first one, hereunder referred to as P1, is a straightforward translation prompt from Wang et al. (2023, p. 4): "Translate this document from English to French.", to which we added the instruction "Output only the translation.". For the second one, P2, we started from a prompt employed in Farrell (2023b, p. 110): "Please translate the text below into Italian, keeping in mind that lexical variety is required for a human-quality final text". We edited it including the source language so the language pair would appear in both P1 and P2, we added information on the text type, and gave further instructions related to syntactic equivalence between the source and the target text. This resulted in the following prompt P2: "Translate this editorial from English to French. Please keep in mind that lexical and syntactic variety is required for a human-quality final text. Calques from English must be avoided and natural-sounding French structures and expressions should be used where needed. Output only the translation."

Post-editing setting The 594-word editorial was post-edited as part of a regular post-editing class session by twelve second-year Master's students in multidisciplinary translation, all native speakers of French, in mid-October 2025. They had been following this class for one month and had prior extensive training in translation but little PE training. After filling in an informed consent form to participate, they were instructed to perform a full PE of the text in fr_FR (French from France) to the best of their ability. The students did not have access to any MT tools, but were allowed to use the Internet for any other purpose. Two groups of six students were created by randomly assigning them one of the machine-translated versions of the selected editorial: MT1 (MT generated using P1) or MT2 (MT generated using P2). For the sake of clarity, the post-edited versions of the editorial that are based on MT1 will be referred to as PE1 and those based on MT2 will be referred to as PE2. The students were not informed of the MT origin nor of the aims of the study to limit biases. After the post-editing brief, they were given 110 minutes to complete the PE task in class.

2.3 Linguistic metrics

To assess the prompt effectiveness in increasing lexical variety and decreasing syntactic equivalence, a series of linguistic metrics (described below) were computed on the MTs of the 18 editorials generated using P1 and P2. These metrics were also used for comparing the post-edited versions of the selected editorial depending on the MT version underlying the PE step (MT1 or MT2).

For statistical testing, we tested the normality of differences between the paired metrics in the machine-translated editorials and the normality of each distribution in the post-edited editorials using the Shapiro-Wilk test (α = 0.05). In cases where the normality assumption was met, we relied on Student's paired t-tests to assess whether P1 and P2 led to significantly different linguistic metrics in the MTs, and on Welch's t-tests (MT1 or MT2) to assess whether the MT version underlying PE led to significantly different linguistic metrics in the post-edited versions. As the normality assumption was not met for SACr between MTs and for PoS changes between post-edited editorials, a Wilcoxon signed-rank test and a Mann-Whitney U-test were used, respectively. To obtain a more comprehensive view, especially because sample sizes are quite low, we also computed effect sizes relying on Hedges' g since it has a higher accuracy than Cohen's d when sample sizes are low (Hedges and Olkin, 2014). For reference, values of 0.2, 0.5, 0.8, 1.2, and 2.0 are considered to indicate small, medium, large, very large, and huge effects (Sawilowsky, 2009), with a g of 1 indicating a difference of one standard deviation. In the comparison between post-edited versions, bootstrap resampling (n = 5,000 iterations, 95% confidence interval) was run to complement the Welch's t-tests results as the sample size was particularly low. The risk of false positives due to multiple comparisons of correlated metrics was addressed by applying Bonferonni corrections to all p-values related to lexical metrics on the one hand, and to syntactic metrics on the other hand.

TTR & MATTR The type-token ratio (TTR) and moving-average type-token ratio (MATTR) (Covington and McFall, 2010) are two closely related measures of lexical diversity based on the ratio of word types to the number of words. One of the main advantages of MATTR over TTR is its really low sensitivity to text length (Bestgen, 2024), but we still calculated the TTR as it is widely used

in the literature. While TTR is computed as the ratio between the number of word types and the total number of words in a text, MATTR is based on 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 the text. One drawback of MATTR, though, is that the *n*-1 first and last words of a text have a lower weight than the others as they appear in fewer windows (Bestgen, 2024). In the present study, we relied on MATTR-50, i.e. using a 50-word window. It should also be noted that these metrics measure repetition, which does not give access to dimensions such as vocabulary sophistication.

Lexical density Lexical density is an indicator of information density (Johansson, 2008) as the ratio between content words and the total number of words in a text (Toral, 2019). Texts with lower lexical density are considered to be easier to process, as the proportion of grammatical words to content words is higher (Baker, 1996).

Metrics of syntactic equivalence In order to assess the level of syntactic equivalence between the source text and the MTs, we used the AS-TrED Python library (Vanroy et al., 2021) in its fully automated mode.[2] The library comprises four metrics: SACr (syntactically aware cross), PoS changes (part-of-speech changes), label changes, and ASTrED (aligned syntactic tree edit distance). Based on aligned and parsed source-target sentence pairs, SACr, PoS changes, and label changes are related to the number of alignment crossings, differences in parts of speech, and differences in dependency labels between source and target, respectively. While the former three are rather shallow indicators of syntactic reorganization in a translation, ASTrED identifies deeper structural changes, since it takes into account both the alignment and the distance between the parsing trees of the source and target sentences (Hansen and Esperança-Rodier, 2022). We relied on the normalized scores as they allow fairer comparisons. For instance, this limits the influence of differences in text length between MTs or between post-edited versions. Higher values suggest lower syntactic equivalence between the source text and the target text, i.e. fewer syntactic calques, with SACr/ASTrED increasing in line with alignment/parsing trees differences between the source and target text. For more comprehensive explanations on the metrics of syntactic equivalence, we refer the reader to Vanroy et al. (2021).

[2]It relies on the Stanza tokenizer and parser (Qi et al., 2020) and a modified version of awesome-align (Dou and Neubig, 2021)

2.4 Qualitative analysis of the post-edited texts

The second main focus of the study is a qualitative analysis of the post-edited texts. Such human evaluation was deemed necessary in the framework of this article, since automatic evaluation struggles to detect "language nuances and context" (Mukherjee and Shrivastava, 2025, p. 11). At the same time, a well-known limitation of human evaluation is the subjectivity it entails, which explains why "recent developments in MT have seen the development of several human assessment methodologies" (Mukherjee and Shrivastava, 2025, p. 11).

Among the most commonly used methodologies is the MQM (Multidimentional Quality Metrics) typology, but it is not specifically tailored for MTPE quality assessment. We therefore relied on the MTPEAS taxonomy, which was specifically created to classify the types of errors made by students when post-editing (Lefer et al., 2022) and seems to achieve high inter-annotator agreements (Bodart et al., 2024).

The MTPEAS taxonomy contains seven categories that depend on the types of edits made by the post-editors in the MT (Bodart et al., 2024). On the one hand, it provides for edits in parts of the raw MT that are considered as erroneous by the annotators according to whether the post-editor fully, partially, or hardly corrects the tagged error (successful, incomplete, or unsuccessful edit), or leaves it unaddressed (missing edit). On the other hand, it classifies edits in parts of the raw MT where the annotators did not tag any error according to whether the edit improves, does not affect, or reduces (value-adding, unnecessary, or error-introducing edit) the MT quality.

The seven type of edits (or absence thereof) from the MTPEAS taxonomy were subclassified into three categories: positive "+" (value-adding and successful edits), neutral "=" (unnecessary edits), and negative "-" (incomplete, unsuccessful, error-introducing, and missing edits). We also derived additional measures from the MTPEAS annotations: we defined the "missed error rate" as the ratio between the number of MT errors tagged by the annotators and that of missing edits, and the "successful edit rate" as the ratio between the num-

ber of MT errors tagged by the annotators and that of successful edits.

The intended two-step MTPEAS workflow was followed. First, two annotators with extensive professional experience in both PE and translation flagged the errors appearing in MT1 and MT2, before discussing them to obtain a single gold standard of errors in each of the machine-translated texts. Both these annotated versions were used as a gold standard for the consecutive MTPEAS evaluation. After the collection and anonymisation of the 12 post-edited texts produced by the students, said texts were submitted to the annotators, who individually tagged each edit made in the post-edited texts using the appropriate category from the MTPEAS taxonomy. Then, the outcomes of the individual MTPEAS tagging were compared to verify whether the annotation results obtained by the annotators were coherent (though the number of annotators prevented any meaningful measurement of inter-annotator agreement).

3 Results

This section will be divided into two parts. First, quantitative results based on the computed linguistic metrics will be presented. We will report the statistical differences in metrics between the 36 MTs of the 18 editorials depending on the prompt used (P1 vs P2) as well as between the 12 post-edited versions of the selected editorial (PE1 vs PE2).

The second subsection will be dedicated to the qualitative PE analysis carried out using the MTPEAS taxonomy. We will start by highlighting the similarities and differences between the tagging figures obtained by both annotators. Then, we will report the number of positive, neutral and negative edits performed in PE1 as opposed to PE2. Finally, we will present the number of errors tagged by the annotators in both gold standards as well as the successful edit rate and missed error rate in PE1 and PE2.

3.1 Linguistic metrics

When it comes to differences between the 18 MTs produced using P1 and the 18 MTs produced using P2 in terms of lexical and syntactic metrics, the Bonferonni-corrected Student's paired *t*-tests (and the Bonferonni-corrected Wilcoxon signed-rank test for SACr) indicated significant differences at low significance thresholds ($p < 0.0005$). The effect size was medium for lexical density (Hedges's $g > 0.5$), large for TTR, MATTR-50, and label changes ($g > 0.8$), and very large for the normalized measures of ASTrED, SACr, and PoS changes ($g > 1.2$). The mean metrics scores and statistical test results are shown in Table 1. Metrics scores are rounded to three decimals in the tables.

Metric	w/ P1	w/ P2	Δ %	*g*
Lex. density	0.544	0.554	+ 1.91	0.62
TTR	0.509	0.541	+ 6.30	0.82
MATTR-50	0.846	0.860	+ 1.54	0.85
ASTrED	0.454	0.531	+ 16.80	1.85
SACr	0.131	0.189	+ 44.31	1.34
PoS changes	0.164	0.186	+ 13.44	1.17
Label changes	0.207	0.246	+ 18.65	1.76

Table 1. Mean metrics scores in the MTs produced using P1 and P2, percentage difference, and Hedges' *g*. All results are significant at $p < 0.0005$.

Looking at the MTs of the selected editorial underlying the PE process (MT1 and MT2), linguistic metrics follow the general trend: both the lexical and syntactic metrics are higher in the MT generated using P2, which specifically requested lexical and syntactic variety. The scores and the percentage difference between MT1 and MT2 are presented in Table 2.

Metric	MT1	MT2	Δ %
Lexical density	0.549	0.564	+ 2.74
TTR	0.506	0.534	+ 5.60
MATTR-50	0.847	0.864	+ 1.95
ASTrED	0.437	0.494	+ 13.13
SACr	0.204	0.231	+ 13.03
PoS changes	0.218	0.222	+ 1.55
Label changes	0.206	0.220	+ 6.93

Table 2. Metrics in MT1 and MT2 and percentage difference

Regarding differences in metrics between the 6 texts that were post-edited from MT1 and the 6 texts post-edited from MT2, results were very different between lexical and syntactic metrics. As a matter of fact, the Bonferonni-corrected Welch's *t*-tests (and the Bonferonni-corrected Mann-Whitney U-test for PoS changes) revealed no significant difference in terms of metrics of syntactic equivalence, while differences in lexical metrics remained significant after the PE step, with huge effect sizes ($g > 2.0$). Table 3 contains the mean metrics scores of the texts post-edited from MT1 (PE1) as opposed to from MT2

(PE2), the percentage difference between these scores, and the statistical test results.

Metric	PE1	PE2	Δ %	*g*
Lex. density	0.548	0.561	+ 2.22†	2.40
TTR	0.503	0.530	+ 5.46*	3.92
MATTR-50	0.852	0.867	+ 1.80*	3.45
ASTrED	0.462	0.477	+ 3.09	0.47
SACr	0.227	0.248	+ 8.96	0.43
PoS changes	0.213	0.211	- 0.74	- 0.10
Label changes	0.220	0.216	- 1.62	- 0.29

Table 3. Mean metrics scores in PE1 and PE2, percentage difference, and Hedges' *g*. * indicates significance at $p < 0.0005$ and † indicates significance at $p < 0.01$.

Bootstrapping (n = 5,000 iterations) results corroborate these results, with Bonferonni-corrected p-values less than 0.001 for the lexical metrics, but no significant result for the syntactic metrics.

3.2 Qualitative analysis

In this second part of the analysis, the qualitative results obtained following the application of the MTPEAS taxonomy on the post-edited texts will be presented. As shown in Section 3.1, P2 had the intended effect, as metrics hint at lower syntactic equivalence and more lexical variety in MT2 as compared to MT1. These quantitative differences are also reflected in qualitative observations, with MT2 generally featuring more fluent and idiomatic turns of phrase. Before presenting the MTPEAS results, we provide an example of this to give a clearer idea of the stylistic differences between MT1 and MT2: the English excerpt "A paradigm shift is needed if the "new Europe" to which Ms von der Leyen alludes is truly to emerge, and if its voice is to be heard in a darkening multipolar world" was rendered by "Un changement de paradigme est nécessaire si la « nouvelle Europe » à laquelle Mme von der Leyen fait allusion doit réellement émerger, et si sa voix doit être entendue dans un monde multipolaire assombri[3]" (MT1) and "Un véritable changement de cap s'impose si « la nouvelle Europe » qu'évoque Mme von der Leyen doit réellement voir le jour et faire entendre sa voix dans un monde multipolaire de plus en plus instable[4]" (MT2).

[3]Back translation: "A paradigm shift is needed if the "new Europe" to which Ms von der Leyen alludes is truly to emerge, and if its voice is to be heard in a darkened multipolar world."

[4]Back translation: "A real change of course is essential if the "new Europe" Ms von der Leyen mentions is truly to come into existence and get itself heard in a increasingly unstable multipolar world."

Annotator 1	PE1	PE2	Total
# PE edits	305	236	541
# PE edits (+)	94	50	144
# PE edits (=)	37	46	83
# PE edits (-)	174	140	314
Annotator 2	**PE1**	**PE2**	**Total**
# PE edits	307	225	532
# PE edits (+)	97	66	163
# PE edits (=)	57	31	88
# PE edits (-)	153	128	281

Table 4. PE tagging results – both annotators

Table 4 shows the results of the tagging performed by each of the annotators, subclassified as described in Section 2.4. First of all, both annotators obtained similar results in terms of number of edits, indicating that the tagging setting and the MTPEAS taxonomy provided coherent results overall. The slight discrepancy in the total number of edits in PE2 is due to rephrasing operations leading to changes elsewhere in the text, as these can be considered as one or several edits. The other discrepancies are related to the neutral (unnecessary) edits, because of the relatively subjective nature of this particular category. Indeed, some edits were deemed "unnecessary" by one annotator, while the other considered them to be "value-adding" or "error-introducing". The same goes for "unsuccessful" and "incomplete edits", where one annotator may have considered that a certain edit corrected part of an MT error, while the other considered that the error was not corrected at all.

Overall, it can also be observed that there were more edits in PE1 as compared to PE2. It is however worth mentioning that 31 MT errors were identified by the annotators in MT1 as opposed to 24 in MT2 (although they were sometimes perceived by the annotators as being more serious in MT2). This can partly influence the results in terms of total number of edits, since the presence of more MT errors requires more edits for them to be corrected. However, this is only partially true, as many edits that are not linked to the raw MT version were found in the post-edited texts.

The numbers presented in Table 5 below show that a higher number of positive edits were generally made in individual PE1 texts in comparison to in PE2 texts.

	PE1						PE2					
Annot. 1	14	15	10	10	20	25	6	8	12	5	8	11
Annot. 2	13	18	9	16	17	24	11	11	15	10	8	11

Table 5. Positive edits by text

Table 6 shows that this was also the case for negative edits, i.e. MT problems that were not solved or errors that were introduced into the MT output by the participants (although the MT was correct).

	PE1						PE2					
Annot. 1	27	33	30	30	25	29	24	23	20	29	21	23
Annot. 2	26	26	27	26	26	22	22	24	17	22	21	22

Table 6. Negative edits by text

When post-editing MT1, which contained more structural calques from the source text, the missed error rate was 49.2% on average, meaning the students failed to identify nearly half of the tagged MT errors. This rate increased to 70.8% when the students post-edited MT2, whose sentence structures were further off the source text structures. By contrast, the successful edit rate was twice higher in PE1 as compared to PE2, with 30.6% of errors tagged by the annotators correctly addressed in PE1 as opposed to 15.2% in PE2. Table 7 shows the number of successful and missing edits per post-edited text as well as the corresponding successful edit and missed error rates. Full MTPEAS results are presented in Table 11 in the Appendix.

4 Discussion

In this section, we will analyse the quantitative data obtained from the MTs generated using P1 and P2 and discuss the effectiveness of the latter prompt in producing the expected outputs. Then, we will look at the data related to the post-edited versions of MT1 and MT2, before bridging the gap between these quantitative insights and the results of the qualitative MTPEAS evaluation described above.

First, regarding the effects of prompting on lexical and syntactic metrics, statistical tests on the 18 editorials show that the MTs produced using P2 led both to greater lexical density and variety, and to less syntactic equivalence with the source text as measured with ASTrED, SACr, PoS changes, and label changes. These results show that the elaborate prompt P2 led to the intended increases in lexical and syntactic metrics as compared to the straightforward translation prompt P1. This somewhat contrasts with previous results (Scourneau, 2025), where an automatic post-editing prompt (GPT-4o) explicitly requesting lexical and syntactic variety led to lower lexical density, TTR, and MATTR-100 than a more generic prompt in terms of lexical metrics, while in terms of syntactic metrics, only PoS changes were higher using the more detailed prompt, with ASTrED and SACr measures not differing significantly.

The MT linguistic metrics of the selected editorial used as the source text to be post-edited as part of the present study followed the general trend. As a matter of fact, all the metrics scores were higher in MT2 than in MT1, which supports i) more information density, ii) less repetition, and iii) sentence structures that are less similar to the source text in MT2 than in MT1.

With respect to the quantitative data extracted from the post-edited texts, syntactic metrics seem to reach a ceiling: looking at the 12 post-edited texts (6 from MT1, 6 from MT2), all the statistically significant differences in terms of syntactic equivalence disappeared. This could be due to the very nature of that kind of metrics, as structural differences between the source and the target language are inherently limited by the typical English and French grammatical structures. For example, given that both English and French are SVO languages using generally similar noun phrase structures, certain alignment crossings are very unlikely even after extensive rephrasing. One hypothesis for the absence of difference after the PE step is therefore that excessively calqued structures in MT1 and MT2 were rephrased during the PE process, which led to an increase in the syntactic metrics up to a soft ceiling. As metrics of syntactic equivalence were higher in MT2, the restructuring effort may have been lower for the students post-editing from that MT, though. However, as far as the lexical metrics are concerned, the initial differences that were observed between MT1 and MT2 remain: the texts post-edited from MT2 present a slightly but significantly higher degree of lexical density and variety than those post-edited from MT1. This could suggest that post-editing an MT characterised by a higher degree of lexical density and variety enables post-editors to produce lexically denser and less repetitive post-edited texts. Of course, it must be emphasised that a small sample of 12 post-edited texts does not allow any de-

	MT1						MT2					
Tagged errors	31	31	31	31	31	31	24	24	24	24	24	24
	PE1						**PE2**					
Annotator 1	PE1-1	PE1-2	PE1-3	PE1-4	PE1-5	PE1-6	PE2-1	PE2-2	PE2-3	PE2-4	PE2-5	PE2-6
Successful edit	9	10	4	6	14	10	2	4	6	2	1	6
Successful edit rate	29.0%	32.3%	12.9%	19.4%	45.2%	32.3%	8.3%	16.7%	19.4%	8.3%	4.2%	25.0%
Mean suc. edit rate	28.5%						13.6%					
Missing edit	18	14	25	14	10	14	20	17	15	15	20	16
Missed error rate	58.1%	45.2%	80.6%	45.2%	32.3%	45.2%	83.3%	70.8%	62.5%	62.5%	83.3%	66.7%
Mean mis. error rate	51.1%						71.5%					
	PE1						**PE2**					
Annotator 2	PE1-1	PE1-2	PE1-3	PE1-4	PE1-5	PE1-6	PE2-1	PE2-2	PE2-3	PE2-4	PE2-5	PE2-6
Successful edit	8	11	6	10	13	13	4	4	8	2	2	4
Successful edit rate	25.8%	35.5%	19.4%	32.3%	41.9%	41.9%	16.7%	16.7%	33.3%	8.3%	8.3%	16.7%
Mean suc. edit rate	32.8%						16.7%					
Missing edit	19	13	23	15	8	10	18	19	15	13	21	15
Missed error rate	61.3%	41.9%	74.2%	48.4%	25.8%	32.3%	75.0%	79.2%	62.5%	54.2%	87.5%	62.5%
Mean mis. error rate	47.3%						70.1%					
Both annotators	**PE1**						**PE2**					
Mean suc. edit rate	30.6%						15.2%					
Mean mis. edit rate	49.2%						70.8%					

Table 7. Tagged errors, successful edits, and missing edits per text and successful edit and missed error rates.

gree of generalisation, and that these trends are only applicable to the present study.

It also remains to be determined whether higher TTR, MATTR and lexical density scores are necessary or even desirable from a qualitative point of view. Some authors have agreed with that view in relation to certain kinds of texts. For example, Farrell (2018, p. 58) argued that "Italians are taught that good writers should avoid unnecessary lexical repetition", which can arguably be true for the French language as well. However, such an assumption is rather subjective and it would be challenging to gather data allowing to reject or reinforce it.

Also, some languages are more accepting of repetitions than others, and the tolerance to repetition also depends on the function of the text: "repeating information serves important communicative and cognitive functions for both speakers and hearers" (Leufkens, 2020, p. 82). This adds a layer of difficulty: for instance, lower lexical density and variety could be desirable in procedural texts as it contributes to making a text easier to process (Baker, 1996), but undesirable in other text types such as editorials, where it "would make the text less interesting to read and less intellectually stimulating" (Farrell, 2018, p. 58). Therefore, more extensive PE quality assessment studies should be carried out to formally establish whether higher lexical density and variety is desirable, and if so, in which text types.

These quantitative metrics can be associated to the qualitative phenomena observed in the second phase of the study. First, the higher number of PE operations in PE1 could be due to the lower lexical variety and greater syntactic equivalence with the source text in MT1 as compared to MT2, since more lexical and structural changes may have been required in MT1 for it to sound natural to French-speaking readers, whereas MT2 required fewer such changes. On the other hand, the students post-editing MT2 may have been misled by its higher lexical variety and density as well as by the more prominent sentence restructuring and rewording, since they spotted fewer errors tagged by the annotators than the students who post-edited MT1, as shown in the Results section. Consistent with this, fewer MT errors were successfully corrected in PE2 than in PE1.

Therefore, a prompt encouraging greater syntactic reordering, such as P2, ultimately seems to largely reduce the students' successful edit rate (30.6% in MT1 vs 15.2% in MT2) and increase their missed error rate (49.2% in MT1 vs 70.8% in MT2).

More specifically, the lower number of repetitions, slightly higher information density, and lower degree of syntactic equivalence (i.e. higher divergence between the source and target structures) in MT2 may have made MT errors more difficult to detect and successfully correct, possibly because of the greater cognitive effort required.

Furthermore, the more natural sentence structures and added idiomaticity in MT2 seem to have led the students who post-edited it to overlook meaning shifts in the MT output, probably because of an excessive trust in the fluent MT output linked to their lack of PE experience: "[c]overt errors in overtly fluent AI translation can be difficult to recognise, sometimes causing serious consequences" (Zhang and Doherty, 2025, p. 237). For instance, the English excerpt "[...] has hobbled the EU's response to a genocidal war that has horrified European populations" was machine-translated into "[...] ont entravé la réponse de l'UE à une guerre génocidaire qui a horrifié les populations européennes[5]" in MT1 and into "[...] ont entravé la réponse européenne face à une guerre génocidaire qui a profondément choqué les opinions publiques[6]" in MT2. The adjective "European" is omitted in MT2, yet 3 students out of 6 failed to restore this information. One possible explanation is that "opinions publiques" is arguably more fluent than "populations européennes" (MT1), making the omission less noticeable. Another hypothesis is that post-editing MT2 was more monotonous than post-editing MT1 for the students, since MT2 already featured a lot of the required syntactic reordering and improved idiomaticity, while more elements need to be manually changed in MT1. As a result, the students may have lowered their vigilance when post-editing MT2. Alternatively, it cannot be totally excluded that despite the random assignment of the PE task, there was a difference in vigilance between students post-editing from MT1 and from MT2. Nevertheless, this limitation cannot be overcome in such an experimental setting, as students cannot be asked to post-edit a same text twice, nor is it advisable to compare the impact of two different prompts on two different editorials.

Finally, although the successful edit rate is twice lower in PE2 than in PE1, it should be noted that it is still very low in both cases (see Results section). The same goes for the missed error rate, where nearly half (MT1) and more than two thirds (MT2) of the MT errors remained unaddressed – although the annotators tagged more errors in MT1 than in MT2 (31 vs 24). It is also interesting to mention that the missed error rate in PE1 (based on a raw MT output that may arguably be closer than MT2 to what a neural MT system would produce) is closely in line with the ratio of uncorrected errors found by Volkart et al. (2022) with students post-editing a neural MT output in a similar translation direction (English – French and English – Italian). Overall, these trends could indicate that students do not have enough professional experience to spot most MT errors or that they tend to excessively trust the MT output, especially when it is less literal, more fluent, and lexically richer.

[5]Back translation: "[...] has hampered the EU's response to a genocidal war that has horrified European populations"

[6]Back translation: "[...] has hampered the European response to a genocidal war that has deeply shocked public opinions"

5 Conclusion and future work

The present study aimed at investigating the influence of prompting strategies on lexical and syntactic metrics in MTs and their impact on the edit types made by Master's translation students in their post-edited texts, as measured using the MTPEAS taxonomy.

We did so by submitting a corpus of 18 British editorials to GPT-5.0 for English-to-French translation using two prompts: a straightforward translation prompt (P1) and a more detailed one encouraging lexical and syntactic variety as well as natural-sounding structures and expressions (P2). The linguistic metrics outlined in the article showed that the latter prompt successfully resulted in higher lexical variety and density as well as lower syntactic equivalence in the form of more idiomatic sentence rephrasings. Both MTs (MT1 and MT2) of a representative editorial that was previously selected according to a series of linguistic metrics were then post-edited by 12 Master's students in a controlled environment, with 6 of them post-editing MT1 the 6 others post-editing MT2.

Following this, a quantitative and qualitative analysis of the post-edited texts was performed, using the same metrics as for the MTs and the MTPEAS taxonomy, respectively. The quantitative data showed that the significant difference in lexical metrics between MT1 and MT2 remained in PE1 vs PE2, implying that gains in lexical variety and density in an MT remain after the PE step. On the other hand, post-editing seems to have a leveling effect in terms of syntactic equivalence, with the syntactic metrics not being significantly different anymore between PE1 and PE2 after the PE step.

On the qualitative side, MTPEAS results show that post-editing from the MT featuring fewer

structural calques and more idiomatic sentence structures caused the student post-editors to make fewer edits overall, to address fewer MT errors, and to properly correct fewer of them.

At a time when developments in LLM-based MT lead translators to work with more and more fluent MT outputs, one could therefore argue that PE tasks are becoming more and more difficult for post-editors (be they students or professionals). A further line of research would be to study whether PE training and experience could improve PE skills, or whether perceived improvements in MT fluency and higher numbers of natural-sounding rephrasings in LLM-based MT outputs and automatic PE (frequently integrated into LSP workflows) do not intrinsically increase the post-editor's cognitive load in a task they already often consider as not intellectually stimulating. In other words, it would be interesting to investigate whether it could be more efficient to post-edit from MT outputs that are closer to the source text and therefore require more editing work, but may also be more intellectually rewarding, or from more fluent and elegant outputs, reducing the number of needed PE operations but requiring the post-editors to be especially careful to properly detect MT errors.

Another avenue of research is a pedagogical one: in an era of swift technology developments and fluent LLM-based MT, students must be trained to be able to face tomorrow's challenges. As a matter of fact, the present study shows that students may have trouble post-editing fluent contents, in that they fail to spot and properly correct many MT errors, misled by natural-sounding MT outputs. More than ever, the impact of prompting strategies, syntactic priming, and MT fluency must be carefully dealt with in PE teaching activities.

To overcome one of the limitations of the present study, i.e. the small 12-participant cohort post-editing in the English-to-French direction, further research on larger sample sizes and in different language pairs is needed. It would indeed make it possible to assess whether the results can be generalized and hold true in other language directions. Finally, it would be interesting to also consider fine-grained aspects related to meaning preservation, which were out of the scope of this work.

References

Alvarez-Vidal, Sergi, Maria Do Campo, Christian Olalla-Soler, and Pilar Sánchez-Gijón. 2025. Using Translation Techniques to Characterize MT Outputs. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 619–627, Geneva, Switzerland, June. European Association for Machine Translation.

Baker, M. 1996. Corpus-based Translation Studies: The Challenges that Lie Ahead. In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pages 175–188. John Benjamins Publishing Company.

Bangalore, Srinivas, Bergljot Behrens, Michael Carl, Maheshwar Gankhot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2015. The role of syntactic variation in translation and post-editing. *Translation Spaces*, 4(1):119–144. John Benjamins.

Bestgen, Yves. 2024. Diversité lexicale et longueur du texte en évaluation du langage. In *Actes des 17es Journées internationales d'Analyse statistique des Données Textuelles*, pages 89–98, Brussels, Belgium, June.

Bodart, Romane, Justine Piette, and Marie-Aude Lefer. 2024. The Machine Translation Post-Editing Annotation System (MTPEAS): A standardized and user-friendly taxonomy for student post-editing quality assessment. *Translation Spaces*, 13(2):265–292. John Benjamins Publishing Company.

Castilho, Sheila and Natália Resende. 2022. Post-Editese in Literary Translations. *Information*, 13(2):66. Multidisciplinary Digital Publishing Institute.

Castilho, Sheila, Natália Resende, and Ruslan Mitkov. 2019. What Influences the Features of Post-editese? A Preliminary Study. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 19–27, Varna, Bulgaria, September.

Chen, Pinzhen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative Translation Refinement with Large Language Models. arXiv:2306.03856 [cs].

Covington, Michael A. and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100. Routledge.

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. Translationese and Post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 16:89–103.

Dou, Zi-Yi and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April. Association for Computational Linguistics.

ELIS, Research. 2026. European Language Industry Survey 2026, March.

Farrell, Michael. 2018. Machine translation markers in post-edited machine translation output. In Chambers, David, Joanna Drugan, Jo ao Esteves-Ferreira, Juliet Margaret Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, editors, *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59, London, UK, November. AsLing.

Farrell, Michael. 2023a. Current evidence of post-editese: differences between post-edited neural machine translation output and human translation revealed through human evaluation. In *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*, pages 52–63, Naples, Italy, July.

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. In *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology 2023*, pages 108–113, Naples, Italy, July.

Gellerstam, M. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 439–448, New York, NY, USA. Association for Computing Machinery.

Hansen, Damien and Emmanuelle Esperança-Rodier. 2022. Human-Adapted MT for Literary Texts: Reality or Fantasy? In *Proceedings of the International Conference New Trends in Translation and Technology 2022*, pages 178–190, Rhodes Island, Greece, July.

Hedges, Larry V. and Ingram Olkin. 2014. *Statistical Methods for Meta-Analysis*. Academic Press, San Diego.

Johansson, Victoria. 2008. Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working papers / Lund University, Department of Linguistics and Phonetics*, 53:61–79.

Kocmi, Tom, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA, November. Association for Computational Linguistics.

Kocmi, Tom, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 462–483, Suzhou, China, November. Association for Computational Linguistics.

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

Lefer, Marie-Aude, Justine Piette, and Romane Bodart. 2022. Machine Translation Post-Editing Annotation System. Technical report, UCLouvain University.

Leufkens, Sterre. 2020. A functionalist typology of redundancy. *Revista da ABRALIN*, 19(3):79–103.

Li, Yafu, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in Literalism: How Supervised Training Shapes Translationese in LLMs, March. arXiv:2503.04369 [cs].

Loock, Rudy. 2018. Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. *Meta: Journal des traducteurs*, 63(3):786–806.

Marques, Francisco Paulo Jamil and Camila Mont'Alverne. 2021. What are newspaper editorials interested in? Understanding the idea of criteria of editorial-worthiness. *Journalism*, 22(7):1812–1830. SAGE Publications.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated*

Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge.

Mukherjee, Ananya and Manish Shrivastava. 2025. Lost in translation? found in evaluation: A comprehensive survey on sentence-level translation evaluation. *ACM Computing Surveys*, 58(1):1–47.

Poibeau, Thierry. 2022. On "Human Parity" and "Super Human Performance" in Machine Translation Evaluation. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6018–6023, Marseille, France, June. European Language Resources Association.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Celikyilmaz, Asli and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.

Sawilowsky, Shlomo S. 2009. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8:597–599.

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*, 237(1):109–130. Publisher: Armand Collin.

Scourneau, Valentin. 2025. Impact of Automatic Post-Editing and Prompting Strategies on the Linguistic Features of English-to-French Editorial Translations. *Tradumàtica tecnologies de la traducció*, 23:65–100.

Toral, Antonio. 2019. Post-editese: an Exacerbated Translationese. In Forcada, Mikel, Andy Way, Barry Haddow, and Rico Sennrich, editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In Forcada, Mikel, Andy Way, Barry Haddow, and Rico Sennrich, editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.

Vanroy, Bram, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken. 2021. Metrics of syntactic equivalence to assess translation difficulty. In *Explorations in empirical translation process research*, volume 3, pages 259–294. Springer.

Volkart, Lise and Pierrette Bouillon. 2023. Are post-editese features really universal? In *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*, pages 294–304, July.

Volkart, Lise and Pierrette Bouillon. 2024. Post-editors as Gatekeepers of Lexical and Syntactic Diversity: Comparative Analysis of Human Translation and Post-editing in Professional Settings. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 387–395, Sheffield, UK, June. European Association for Machine Translation.

Volkart, Lise, Sabrina Girletti, Johanna Gerlach, Jonathan David Mutal, and Pierrette Bouillon. 2022. Source or target first? Comparison of two post-editing strategies with translation students. *Journal of Data Mining and Digital Humanities*, Towards robotic translation? INRIA.

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. arXiv:2304.02210 [cs].

Zhang, Jia and Stephen Doherty. 2025. Investigating novice translation students' ai literacy in translation education. *The Interpreter and Translator Trainer*, 19(3-4):234–253.

A Linguistic metrics per text

The tables below contain the linguistic metrics scores for each text. Table 8 and Table 9 show the full lexical and syntactic metrics scores for the 36 machine-translated editorials from the corpus, depending on the prompt used. Table 10 shows the metrics scores for each of the 12 post-edited texts. We report the normalized scores for the syntactic metrics.

B Full MTPEAS edit results

Table 11 shows the full MTPEAS results. It contains the edit types performed by the students (both according to the MTPEAS taxonomy and to the subclassification described in Section 2.4) as tagged by each annotator in each of the 12 post-edited texts.

Text ID	TTR		MATTR-50		Lex. dens.	
	w/ P1	w/ P2	w/ P1	w/ P2	w/ P1	w/ P2
1-GV	0.511	0.554	0.841	0.859	0.568	0.562
1-IV	0.478	0.520	0.824	0.852	0.526	0.544
1-TV	0.558	0.579	0.870	0.876	0.563	0.571
2-GV	**0.506**	**0.534**	**0.847**	**0.864**	**0.549**	**0.564**
2-IV	0.441	0.463	0.821	0.837	0.523	0.537
2-TV	0.515	0.560	0.840	0.845	0.535	0.545
3-GV	0.519	0.530	0.845	0.854	0.523	0.527
3-IV	0.447	0.470	0.824	0.838	0.531	0.537
3-TV	0.554	0.598	0.847	0.860	0.531	0.555
4-GV	0.539	0.561	0.864	0.858	0.565	0.564
4-IV	0.457	0.496	0.835	0.859	0.537	0.554
4-TV	0.557	0.600	0.871	0.894	0.557	0.579
5-GV	0.509	0.532	0.835	0.836	0.555	0.561
5-IV	0.497	0.522	0.864	0.867	0.518	0.526
5-TV	0.516	0.553	0.853	0.864	0.563	0.574
6-GV	0.545	0.560	0.855	0.865	0.554	0.555
6-IV	0.476	0.529	0.855	0.870	0.534	0.547
6-TV	0.544	0.585	0.845	0.875	0.558	0.577

Table 8. Comparison of TTR, MATTR-50, and lexical density in the MTs generated using P1 and P2

Text ID	ASTrED		SACr		PoS chgs		Label chgs	
	w/ P1	w/ P2	w/ P1	w/ P2	w/ P1	w/ P2	w/ P1	w/ P2
1-GV	0.595	0.644	0.238	0.415	0.198	0.218	0.243	0.305
1-IV	0.390	0.442	0.108	0.107	0.153	0.176	0.184	0.234
1-TV	0.442	0.557	0.113	0.187	0.178	0.240	0.196	0.271
2-GV	**0.437**	**0.494**	**0.204**	**0.231**	**0.218**	**0.222**	**0.206**	**0.220**
2-IV	0.437	0.505	0.113	0.167	0.128	0.178	0.200	0.261
2-TV	0.498	0.586	0.090	0.109	0.149	0.143	0.235	0.253
3-GV	0.504	0.548	0.107	0.169	0.168	0.206	0.212	0.256
3-IV	0.433	0.494	0.086	0.145	0.135	0.157	0.199	0.233
3-TV	0.396	0.502	0.125	0.228	0.166	0.188	0.193	0.221
4-GV	0.436	0.548	0.117	0.178	0.166	0.189	0.205	0.254
4-IV	0.401	0.573	0.134	0.204	0.159	0.192	0.202	0.269
4-TV	0.529	0.612	0.172	0.254	0.207	0.211	0.223	0.231
5-GV	0.443	0.462	0.114	0.130	0.150	0.157	0.201	0.221
5-IV	0.411	0.515	0.058	0.110	0.192	0.185	0.224	0.233
5-TV	0.485	0.573	0.163	0.240	0.177	0.192	0.236	0.265
6-GV	0.393	0.456	0.113	0.134	0.101	0.129	0.171	0.213
6-IV	0.475	0.567	0.142	0.213	0.158	0.195	0.199	0.259
6-TV	0.471	0.473	0.153	0.170	0.143	0.170	0.200	0.227

Table 9. Comparison of ASTrED, SACr, PoS changes, and label changes in the MTs generated using P1 and P2

Text ID	Lexical density	TTR	MATTR-50	ASTrED	SACr	PoS changes	Label changes
PE1-1	0.557	0.508	0.859	0.440	0.186	0.227	0.221
PE1-2	0.553	0.512	0.848	0.472	0.218	0.228	0.238
PE1-3	0.545	0.510	0.849	0.452	0.266	0.214	0.205
PE1-4	0.543	0.499	0.854	0.481	0.310	0.210	0.222
PE1-5	0.543	0.493	0.854	0.413	0.159	0.181	0.205
PE1-6	0.549	0.498	0.847	0.515	0.227	0.216	0.229
PE2-1	0.560	0.528	0.862	0.488	0.233	0.205	0.220
PE2-2	0.559	0.537	0.865	0.474	0.277	0.233	0.233
PE2-3	0.557	0.527	0.870	0.457	0.232	0.210	0.215
PE2-4	0.560	0.531	0.870	0.503	0.240	0.207	0.214
PE2-5	0.561	0.527	0.866	0.457	0.212	0.210	0.213
PE2-6	0.566	0.536	0.871	0.481	0.293	0.202	0.203

Table 10. Lexical and syntactic metrics in each of the post-edited texts

	TA1						TA2					
Tagged errors	31	31	31	31	31	31	24	24	24	24	24	24
	PE1						**PE2**					
Annotator 1	PE1-1	PE1-2	PE1-3	PE1-4	PE1-5	PE1-6	PE2-1	PE2-2	PE2-3	PE2-4	PE2-5	PE2-6
Value-adding	5	5	6	4	6	15	4	4	6	3	7	5
Successful	9	10	4	6	14	10	2	4	6	2	1	6
Unnecessary	5	7	0	5	12	8	10	10	10	9	4	3
Incomplete	2	7	2	4	6	5	0	1	1	4	0	0
Unsuccessful	1	4	0	8	3	1	2	1	1	1	1	4
Error-introd.	6	8	3	4	6	9	2	4	3	9	0	3
Missing	18	14	25	14	10	14	20	17	15	15	20	16
Total edits	46	55	40	45	57	62	40	41	42	43	33	37
Total positive	14	15	10	10	20	25	6	8	12	5	8	11
Total neutral	5	7	0	5	12	8	10	10	10	9	4	3
Total negative	27	33	30	30	25	29	24	23	20	29	21	23
Annotator 2	PE1-1	PE1-2	PE1-3	PE1-4	PE1-5	PE1-6	PE2-1	PE2-2	PE2-3	PE2-4	PE2-5	PE2-6
Value-adding	5	7	3	6	4	11	7	7	7	8	6	7
Successful	8	11	6	10	13	13	4	4	8	2	2	4
Unnecessary	11	10	5	5	14	12	6	5	6	8	3	3
Incomplete	2	4	0	1	4	3	0	1	0	2	0	1
Unsuccessful	3	3	2	4	5	4	0	0	0	3	0	1
Error-introd.	2	6	2	6	9	5	4	4	2	4	0	5
Missing	19	13	23	15	8	10	18	19	15	13	21	15
Total edits	50	54	41	47	57	58	39	40	38	40	32	36
Total positive	13	18	9	16	17	24	11	11	15	10	8	11
Total neutral	11	10	5	5	14	12	6	5	6	8	3	3
Total negative	26	26	27	26	26	22	22	24	17	22	21	22

Table 11. Full MTPEAS annotations by each annotator