# Machine Translation Meets Large Language Models: Evaluating ChatGPT's Ability to Automatically Post-Edit Literary Texts

Lieve Macken LT<sup>3</sup>, Language and Translation Technology Team Ghent University Belgium lieve.macken@ugent.be

#### Abstract

Large language models such as GPT-4 have been trained on vast corpora, giving them excellent language understanding. This study explores the use of Chat-GPT for post-editing machine translations of literary texts. Three short stories, machine translated from English into Dutch, were post-edited by 7-8 professional translators and ChatGPT. Automatic metrics were used to evaluate the number and type of edits made, and semantic and syntactic similarity between the machine translation and the corresponding post-edited versions. A manual analysis classified errors in the machine translation and changes made by the post-editors. The results show that ChatGPT made more changes than the average post-editor. ChatGPT improved lexical richness over machine translation The analysis of editing for all texts. types showed that ChatGPT replaced more words with synonyms, corrected fewer machine errors and introduced more problems than professionals.

#### **1** Introduction

In recent years, there has been a noticeable shift in the perception of the use of computerassisted translation technologies for literary translation. Advances in the quality of machine translation (MT) and the development of sophisticated computer-assisted translation (CAT) tools have contributed to this changing landscape (Rothwell et al., 2023).

© 2024 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

Following the emergence of neural machine translation, a growing body of research has examined the use of Machine Translation (MT) and Post-Editing (PE) for literary texts. Researchers have looked at various aspects related to the use of MT and PE in the context of literary translation, such as perceived usefulness of MT (Moorkens et al., 2018; Şahin and Gürses, 2021; Daems, 2022; Ruffo, 2022), ethical issues (Taivalkoski-Shilov, 2019; Kenny and Winters, 2020; Li, 2023), translation quality (Webster et al., 2020; Macken et al., 2022; Castilho and Resende, 2022), the impact on the translation process (Toral et al., 2018; Kolb, 2023), and the reader's reception of the final product (Guerberof-Arenas and Toral, 2020).

Further automation of the translation process can be accomplished by implementing Automatic Post-Editing (APE), which refers to methods that improve the output of machine translation systems by applying automatic editing operations (do Carmo et al., 2021). Technological advances are rapidly evolving and the potential of using AI systems based on large language models (e.g. Chat-GPT) is currently being explored for a variety of applications (Guimarães et al., 2024), one of which is the post-editing of machine translation output (Raunak et al., 2023).

This study explores ChatGPT's ability to automatically post-edit literary texts that were machine-translated from English into Dutch by a neural machine translation system. We evaluate ChatGPT's performance by comparing its automatically post-edited texts to versions that were post-edited by professional literary translators.

#### 2 Related research

Over the past decade, a number of studies have been carried out to examine the usefulness and suitability of machine translation and post-editing for literary translation. Researchers often compare raw (unedited) machine translations of literary texts with their (published) human-translated counterparts. The MT systems used are either generic systems (Webster et al., 2020; Hu and Li, 2023) or MT systems adapted specially for literary translation (Toral et al., 2024; Matusov, 2019; Toral et al., 2024).

In order to gain valuable insights into the strengths and limitations of MT for literary translation, error classification schemes such as MQM (Lommel et al., 2014) or SCATE (Tezcan et al., 2017) are often used. These classification schemes typically distinguish between accuracy and fluency errors. Accuracy errors refer to the failure to transfer meaning correctly from source to target, whereas fluency errors refer to the failure to produce grammatically correct, idiomatic and fluent translations. Existing error classification schemes have been adapted to suit the specific characteristics of literary texts (Tezcan et al., 2019; Matusov, 2019).

Despite the high quality of the current generation of transformer-based neural MT systems, they still produce errors in both accuracy and fluency. This is certainly the case with more creative use of language, which is typical of literary texts. In addition, machine-translated texts exhibit different linguistic characteristics (e.g. less lexical variety, less cohesion, syntactically less diverse texts) than human translations (Vanmassenhove et al., 2019; Webster et al., 2020). The involvement of professional translators in the translation of literary texts is therefore essential.

In the context of literary translation, post-editing can be applied by having human translators work on the raw machine translation suggestions (Toral et al., 2018; Şahin and Gürses, 2019; Guerberof-Arenas and Toral, 2020; Castilho and Resende, 2022; Kolb, 2023). Human translators then correct the errors and polish the machine's raw output, transforming it into a high-quality, publishable literary translation by ensuring that the translated texts capture the nuances, cultural references, and literary techniques present in the original work.

In their study, Macken et al. (2022) compared three successive versions of a Dutch translation of an English novel: the raw MT output, the postedited version and the revision of the post-edited text. They manually annotated the errors in the MT and categorised the editing changes in accordance with a linguistic typology. The study showed that most MT errors were corrected in the postediting process, and that the post-editor mainly made lexico-semantic and stylistic changes. Fortyfour percent of the post-editing changes involved the correction of MT errors, 24% were preferred changes and 9% were labelled as 'undesirable'.

They also used different automatic metrics to measure the (dis)similarity between the different versions, focusing on different aspects. The amount of editing was assessed by Translation Edit Rate (Snover et al., 2006) and CharCut (Lardilleux and Lepage, 2017). Semantic similarity was measured by the neural metrics COMET (Rei et al., 2020) and BERTScore (Zhang et al., 2019), which calculate the distance between vector representations of sentences and tokens. ASTrED (Vanroy et al., 2021), a metric that compares the edit distance between the dependency structures of two sentences, taking into account word alignment information, was used to quantify syntactic changes.

Another feature that has been widely studied in previous research on literary machine translation is lexical richness. Vanmassenhove et al. (2019) showed that MT systems are not able to achieve the same level of lexical richness as human translated texts. Webster et al. (2020) also observed a decrease in lexical richness from human translation to machine translation, suggesting a certain homogenisation of the lexicon used by NMT systems. Macken et al. (2022) investigated whether the level of lexical richness increases during postediting and revision, but in their study they found similar levels of lexical richness in the MT, PE and revised translation.

Large language models such as GPT-4 or LLaMA are trained on unprecedentedly large corpora. LLaMA-3 for example has been pre-trained on approximately 15 trillion tokens of text gathered from publicly available sources<sup>1</sup>. Due to the size of the training set, they have a comprehensive understanding of language. As they are trained on a much larger data set than other end-user applications such as automatic speech recognition or machine translation, researchers propose a combination of the two. Radhakrishnan et al. (2023) used LLaMA to correct errors produced by the Whisper automatic speech recognition system (Radford et al., 2023), a task similar to the post-editing of

<sup>&</sup>lt;sup>1</sup>https://ai.meta.com/blog/meta-llama-3/

machine translation output.

Raunak et al. (2023) explore the use of GPT-4 for automatic post-editing of NMT output in different language pairs. They experimented with WMT-22 General MT translation task datasets and WMT-20 and WMT-21 News translation task submissions annotated with MQM. Translation quality was assessed using neural evaluation metrics. Their results show that GPT-4 effectively improves translation quality compared to the best systems from WMT-22 across a number of language pairs and generates meaningful edits to translations. But they also show that GPT-4 can produce hallucinated edits, suggesting caution in its use as an expert translation post-editor.

Research on the use of ChatGPT for automatic post-editing is very scarce and has not yet been applied to challenging text types such as literary texts. In this study, we extend the work of Raunak et al. (2023) and use ChatGPT 4.0 to automatically post-edit more creative texts. We are not only interested in whether automatic post-editing with ChatGPT improves the quality of the neural machine translation output. We also want to know how ChatGPT's post-editing ability compares with that of professional literary translators. Using automatic and manual evaluation methods we seek an answer to the following research questions

- RQ1: Does ChatGPT make more or less changes to the machine-translated texts than professional literary translators?
- RQ2: To what extent does ChatGPT preserve the meaning of the text compared to professional literary translators?
- RQ3: Does ChatGPT make different types of changes to the machine-translated texts than professional literary translators?
- RQ4: Does ChatGPT solve all the errors present in the machine-translated texts? Does it introduce new problems?

#### 3 Methodology

#### 3.1 Data

We use part of the data set collected in the DUAL-T project (Ruffo et al., 2023; Ruffo et al., 2024), which compares three different literary translation conditions: the conventional method using a word processing tool (Microsoft Word),

translation within a computer-assisted translation environment (Trados Studio 2022), and postediting of machine translation output. Three short stories were selected from the short story collection 'One More Thing' by the American writer B. J. Novak<sup>2</sup>.

A total of twenty-three professional literary translators (8 male, 15 female) participated in the DUAL-T experiments. The translators were contacted through professional translator associations in Flanders and The Netherlands and they were paid to take part in the study. Years of experience in translating literary texts ranged from 1 year to 43 years. Eight participants had made use of post-editing in their professional translation work. Each translator translated each of the three texts into Dutch in a different condition. They were instructed to produce translations of publishable quality. Each combination of text and condition appeared the same number of times in the entire data set.

In this study, we only use the post-edited versions of the three texts. The machine translations of the three texts were generated in July 2023 using a commercial neural machine translation system (DeepL). The professional literary translators worked in a proprietary web-based platform that displayed the source and the machine-translated target text side by side. During translation, the translators could consult online resources when they felt it was appropriate. The source text characteristics and the number of post-edited versions of the three texts are presented in Table 1.

	Words	Sentences	Post-edited versions
T1	306	30	7
T2	349	27	8
Т3	290	30	8

Table 1: Source text characteristics of the three short stories and number of texts post-edited by professional literary translators

We slightly adapted the system and user prompts of Raunak et al. (2023) to generate the post-edited versions of ChatGPT 4.0. The system prompt contains the initial instruction to ChatGPT to complete the post-editing task. The user prompts were given three times, one for each text. The prompts we used for the experiments are presented in Appen-

<sup>&</sup>lt;sup>2</sup>A published translation of this collection is available in Dutch, but only as a printed book. It is therefore very unlikely that this Dutch translation was used to train chatGPT.

dices A and B.

#### 3.2 Automatic evaluation

We use various automatic metrics to evaluate and compare all post-edited versions of each text. Before calculating the automatic metrics, all texts were tokenized using the Stanza toolkit (Qi et al., 2020) and manually aligned at sentence level.

To quantify the amount of editing done by each post-editor, we compare the machine-translated texts with the post-edited versions of each professional translator and ChatGPT. We use Translation Edit Rate (TER) (Snover et al., 2006) and Char-Cut (Lardilleux and Lepage, 2017). TER quantifies editing operations at the token level, while CharCut works at the character level. As such Charcut is more lenient and penalises the use of different word forms (e.g. the Dutch word *stad* (En: *town*) had been changed to the diminutive *stadje* (En: *small town*)) to a lesser extent than TER. TER scores were obtained via the MATEO platform<sup>3</sup> (Vanroy et al., 2023). For CharCut, we used the Python code available on GitHub<sup>4</sup>.

We used BERTScore (Zhang et al., 2019) to measure the semantic similarity between the machine-translated texts and each of their postedited versions. BERTScore is an automatic evaluation metric for text generation, which uses contextual embeddings to compute a similarity score for two given sentences. As such, it can capture semantic similarity of synonyms and will give a higher score to sentences that are semantically similar (e.g. van de plank – van een rek (En: from the shelf – from a rack) than sentences in which the content has been changed, (e.g. van de plank – uit een kast (En: from the shelf – from a cupboard). BERTScores were also obtained via the MATEO platform.

Webster et al. (2020) observed that MT systems tend to follow the syntactic structure of the source text more closely than human translators. We assume that the post-editors will therefore adapt the syntactic structure to bring it closer to the norms of the target language. We use AS-TrED<sup>5</sup> (Vanroy et al., 2021) to quantify the degree of similarity between the syntactic structure of the machine-translated texts and each of their post-edited versions. ASTrED computes the edit distance between the dependency structures of two sentences, taking into account word alignment information. Under the hood, ASTrED uses the stanza parser (Qi et al., 2020) for the creation of universal dependency trees and AWESOMEalign (Dou and Neubig, 2021) for word alignment. It assigns a lower score to sentences with a more similar dependency structure than to sentences where the structure has changed more. In the example in Figure 1, the human post-editor made only minimal changes to the structure, whereas ChatGPT made more changes to the structure. The resulting ASTrED scores are resp. 0,13 for the human post-edited sentence and 0,26 for ChatGPT's version.

Finally, to assess the lexical diversity of each post-edited text, we calculated the Moving Average Type-Token Ratio with a window size of 50 (MATTR-50)<sup>6</sup>. MATTR calculates the ratio of different unique words (types) to the total number of words (tokens) using a moving window of predefined word length and is therefore not sensitive to differences in text length. To obtain more accurate results, we lower-cased all texts before calculating MATTR.

#### **3.3 Manual evaluation**

For the manual evaluation, we largely follow the methodology of Macken et al. (2022). We annotate all errors in the MT output and classify all post-editing changes in the subsequent post-edited translations. As the manual annotation of postediting changes is very time-consuming, we only annotated the ChatGPT version and the post-edited texts produced by the three most experienced professional translators for each text. The translators' years of experience were 24, 21 and 10 for text 1, 43, 20 and 15 for text 2 and 28, 8 and 8 for text 3.

To evaluate the quality of the machine translation, the three machine-translated texts were annotated according to an adapted version of the SCATE error taxonomy (Tezcan et al., 2019). We used the same reduced set of labels as in Macken et al. (2022).

We further classified all post-editing changes in the 4 post-edited versions per text from a linguistic perspective. We made minor adaptations to the categorisation scheme of Macken et al. (2022), which includes four main categories (*lexico-semantic*, *syntax & morphology, style* and *spelling & punc*-

<sup>&</sup>lt;sup>3</sup>https://mateo.ivdnt.org/

<sup>&</sup>lt;sup>4</sup>https://github.com/alardill/charcut

<sup>&</sup>lt;sup>5</sup>https://github.com/BramVanroy/ASTrED

<sup>&</sup>lt;sup>6</sup>https://github.com/kristopherkyle/

lexical\\_diversity

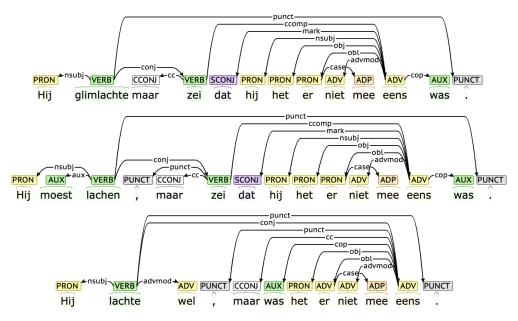


Figure 1: Universal Dependency Trees for the machine-translated sentence, and two post-edited sentences, resp. by a human post-editor and by ChatGPT. English source sentence: *He smiled but said he didn't agree*.

*tuation*), which are subdivided into subcategories (see Table 4 and Appendix C for more details). We also labelled each post-editing change in terms of correctness and necessity by using the following labels (*MT error correction, consistency, preferential* and *undesirable change*). Undesirable changes are edits that clearly degrade the quality of the translation. In the final translation we also identified any MT errors that were not fixed.

All annotations were done in Excel by the author of the paper. To facilitate the labelling of postediting changes, we used Charcut (Lardilleux and Lepage, 2017), which produces an HTML document visualizing the differences between the MT output and the PE version, see Figure 2. The annotation guidelines are given in Appendix C.

T02\_EN.tok.sent.txt *He smiled but said that he didn't agree*. T02\_MT.tok.sent.txt

Hij glimlachte maar zei dat hij het er niet mee eens was . T02\_ChatGPT4.tok.sent.txt Hij lachte wel , maar was het er niet mee eens .

```
28/106= 26%
```

Figure 2: Example of Charcut visualizations (MT-APE)

#### 4 Results

#### 4.1 Automatic evaluation

Table 2 shows the results of the four automatic metrics quantifying the amount of editing that took place (CharCut and TER), semantic similarity (BERTScore) and syntactic similarity (AS-TrED). The metrics were calculated on all translations available to us. The table summarises the results per condition: APE represents the results of the automatically post-edited text by ChatGPT, while PE is the average of the 7 or 8 human post-edited versions.

		$\mathbf{CharCut} \downarrow$	TER $\downarrow$	<b>BERTScore †</b>	ASTrED↓
T1	APE	0,31	0,42	89,33	0,22
T1	PE	0,26	0,36	90,57	0,24
T2	APE	0,37	0,47	88,58	0,23
T2	PE	0,24	0,34	91,46	0,18
T3	APE	0,31	0,39	88,68	0,27
Т3	PE	0,17	0,24	93,07	0,18

Table 2: Overview of automatic evaluation results per text.

 Up arrow: higher value means more similar; down arrow:

 lower value means more similar.

If we compare ChatGPT with the 'average human post-editor', we see that ChatGPT makes more changes to the machine-translated texts than the average human post-editor. The results show a higher degree of editing, both in terms of CharCut and TER, a lower semantic similarity and, for two texts, also a lower syntactic similarity.

Figure 3 presents the CharCut and ASTrED

scores per text and per participant. For texts 2 and 3, ChatGPT obtains the highest CharCut scores. For text 1, two professional literary translators (P07 and P11) make more changes to the machine translation. ChatGPT's ASTrED score for texts 2 and 3 is the second highest; for text 1 it is in the middle range.

Figure 4 presents the MATTR-50 scores for the English source texts, the machine-translated texts and all post-edited versions. While English and Dutch MATTR-50 values cannot be directly compared due to different word formation rules (compounds are written as one word in Dutch), the MATTR-50 values of the English source texts can be used as benchmark to interpret the other values. In most cases, post-editing results in higher MATTR-50 values for all texts, increasing the level of lexical richness compared to the machine-translated texts.

#### 4.2 Manual evaluation

All errors were manually annotated in the machine-translated texts. Overall, the quality of the commercial neural machine translation system is relatively good, with only 18 accuracy errors and 29 fluency errors. Table 3 gives an overview of all the errors found in the machine-translated texts.

In terms of accuracy, mistranslations make up the largest group of errors. Examples of accuracy errors are wrong translations of single words (e.g. *scraggly* is translated as *schamele* (En: *poor*, *scanty*)) or expressions (e.g. *on the last day the rain cleared* is translated literally as *op de laatste dag klaarde de regen op*, which is not idiomatic in Dutch).

The other major group of problems are style problems, and in particular disfluent sentences (belonging to the 'fluency' category). Disfluent sentence constructions are most often the result of copying the structure of the source sentences too literally, as is the case in the following example: *De plek waar mensen hun hele leven voor gespaard hebben om naartoe te gaan*, which is a rather literal translation of *The place that people saved up to visit their whole lives*.

Table 4 gives an overview of all post-editing changes in the three texts. In total, 751 post-editing changes were annotated. Most post-editing changes are lexico-semantic (69%) or stylistic (20%) changes.

Accuracy	18	Fluency	29
Mistranslation	16	Coherence	2
Multiword	5	Discourse marker	0
Word sense	2	Coreference	1
Other	9	Tense	0
Addition	0	Other	1
Omission	0	Lexicon	7
Untranslated	2	Grammar & syntax	1
Do not translate	0	Style	16
		Disfluent	12
		Repetition	1
		Other	3
		Spelling & punctuation	3
		Capitalisation	0
		Compound	1
		Punctuation	2
		Other	0

 
 Table 3: Accuracy and fluency errors in the three machinetranslated texts

Post-editors often replace words with synonyms (boekwinkel  $\rightarrow$  boekhandel (En: bookstore)), make words or phrases more explicit or specific (plek  $\rightarrow$  stad (En: place  $\rightarrow$  town)), make them more implicit or vague (hij tekende zelfs een diagram  $\rightarrow$  hij tekende het zelfs uit (En: He even drew a diagram  $\rightarrow$  he even drew it); in de stad  $\rightarrow$  in de buurt (En: in the city  $\rightarrow$  in the neigbourhood)), or replace words or phrases from the MT output with a better collocation or more idiomatic expression (klaarde de regen op  $\rightarrow$  klaarde het op (En: the rain cleared); op zijn laatst om  $\rightarrow$  of uiterlijk (En: at the latest)).

Post-editors also often make improvements to the structure of the machine translation, e.g. (*De* plek waar mensen hun hele leven voor gespaard hebben om naartoe te gaan  $\rightarrow$  Sommige mensen spaarden hun hele leven om er een keer naartoe te gaan (En: The place that people saved up to visit their whole lives  $\rightarrow$  Some people saved their whole lives to go there once)). They also often prefer another word order (om bloemen te kopen voor zijn vrouw  $\rightarrow$  om bloemen voor zijn vrouw te kopen (En: to pick up flowers for his wife)) or make other stylistisc changes (een vage en opgeblazen en maffe lach  $\rightarrow$  een vage opgeblazen gekke glimlach (En: a vague and bloated and goofy smile)).

Most of the spelling and punctuation changes are related to changing double quotes by single quotes.

Table 5 shows the breakdown of the lexicosemantic and stylistic edits per text and per posteditor. For each of the texts, we can see quite

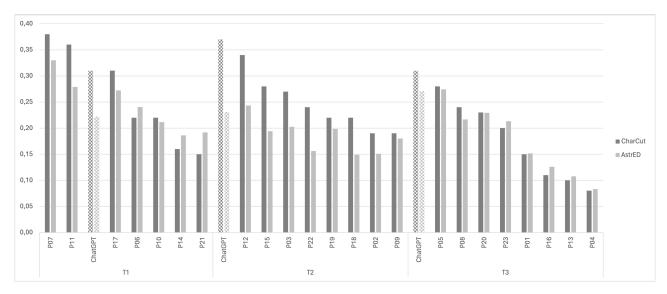


Figure 3: CharCut and ASTrED scores per text per participant, ordered by CharCut scores

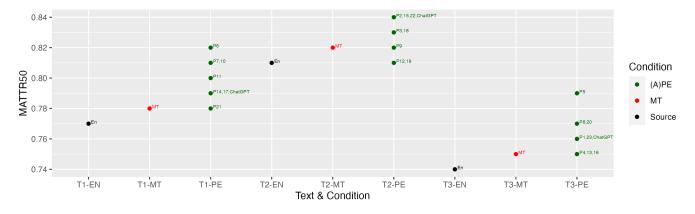


Figure 4: MATTR-50 scores for the English source text, the MT the and (automatically) post-edited texts

Lexico-semantic	517	Syntax & morphology	26
Addition	2	Agreement	2
Coherence marker	39	Number	3
Collocation & idiom	128	Diminutive	8
Deletion	16	Tense	10
Explicitation & specific	68	Other	3
Implicitation & vague	46		
Synonym	134		
Other	84		
Spelling & punctuation	58	Style	150
Capitalization	0	Word order	37
Compound	2	Structural change	50
Linking word & punctuation	9	Shorter	15
Punctuation added	8	Split sentence	0
Punctuation deleted	2	Merged sentence	3
Other	37	Other	45

 Table 4: Categorisation of all post-editing changes in the three texts

a few differences in both the number of changes and the types of changes made by each of the post-editors. The most striking difference between ChatGPT and professional literary translators is that ChatGPT makes more lexico-semantic changes, which can be attributed to the subcategory 'synonym'. ChatGPT thus replaces words with synonyms more often than professional literary translators.

Table 6 presents the quality labels assigned to all post-editing changes by the three post-editors and ChatGPT. The majority of changes (71%) are preferential in nature; 20% of all changes are corrections of MT errors; 5% of the changes were for consistency reasons (e.g. because of adaptations made earlier in the text) and 3% of the changes were labelled as 'undesirable'. These last changes either introduced new errors or made the final target text inconsistent with the information in the source text. Most of the MT errors (80% of all accuracy errors and 88% of all fluency errors) present in the machine-translated texts were solved during post-editing.

The distribution of quality labels is slightly different for ChatGPT compared to the three posteditors, with 74% preferential changes (vs. 70%), 18% MT error corrections (vs. 21%), 6% undesirable changes (vs. 2%) and 2% changes to make the text consistent (vs. 6%). This means that Chat-GPT corrects fewer MT errors and introduces more problems than the human post-editors. An example of problem introduced by ChatGPT is presented in Figure 5. In the example, the MT produces a literal translation of the phrase *In the end*, *this one wasn't for her*, which does not make sense in Dutch. ChatGPT adds the Dutch word *plek* 

TEXT 1	ChatGPT	P6	P11	P17
Lexico-semantic	48	33	40	46
Coherence marker	3	6	4	6
Collocation & idiom	13	10	16	12
Deletion	2	3	2	1
Synonym	15	2	4	9
Explicitation & specific	5	5	8	10
Implicitation & vague	5	5	3	1
Other	5	2	3	7
Style	8	9	14	16
Word order	4	1	3	4
Structural change	1	3	1	1
Shorter	2	1	4	0
Merged sentence	0	1	1	1
Other	1	3	5	10
TEXT 2	ChatGPT	P2	P12	P15
Lexico-semantic	76	32	46	50
Addition	2	0	0	0
Coherence marker	3	3	1	3
Collocation & idiom	11	11	15	12
Deletion	1	0	0	2
Synonym	33	8	7	9
Explicitation & specific	4	4	7	6
Implicitation & vague	8	1	4	5
Other	14	5	12	13
Style	8	16	20	10
Word order	3	5	3	2
Structural change	0	5	10	2
Shorter	1	1	1	1
Other	4	5	6	5
TEXT 3	ChatGPT	P5	P16	P20
Lexico-semantic	46	45	19	36
Coherence marker	3	4	0	3
Collocation & idiom	8	7	5	8
Deletion	3	2	0	0
Synonym	16	11	6	14
Explicitation & specific	7	7	2	3
Implicitation & vague	3	5	2	4
Other	6	9	4	4
Style	19	15	4	11
Word order	5	4	2	1
Structural change	9	9	2	7
		0	0	1
Shorter	3	0	0	1

 Table 5: Overview of the lexico-semantic and stylistic edits

 per text and per post-editor

(*En: place*) so that the meaning of the sentence changes to *In the end, this place was not for her*.

```
T02_EN.tok.sent.txt
In the end, this one wasn't for her.
T02_MT.tok.sent.txt
Uiteindelijk was deze niet voor haar .
T02_ChatGPT4.tok.sent.txt
Op het einde was deze plek toch niet voor
haar bestemd .
```

Figure 5: Example of an undesirable edit by ChatGPT

Quality label	All	Post-editors	ChatGPT
Preferential	536 (71%)	367 (70%)	169 (74%)
MT error correction	153 (20%)	112 (21%)	41 (18%)
Consistency	37 (5%)	33 (6%)	4 (2%)
Undesirable	25 (3%)	10 (2%)	15 (6%)

**Table 6:** Overview of the quality labels assigned to all postediting changes by the three post-editors and ChatGPT

#### 5 Discussion

We conducted an experiment comparing the postediting capabilities of ChatGPT with those of experienced professional literary translators working on English-Dutch literary texts. We used a data set collected in the DUAL-T project, in which 23 professional English-Dutch literary translators post-edited the neural machine translations of three short stories by the same author. We then asked ChatGPT 4.0 to create post-edited versions of the three texts. This collection of post-edited literary translations allows us to compare the results of human and automatic post-editing.

We formulated four research questions and used a combination of automatic and manual evaluation methods to compare all post-edited texts. The CharCut and TER results show that ChatGPT makes more changes to the machine-translated texts than the 'average human post-editor' (RQ1). ChatGPT achieved the highest CharCut scores for two texts and made the most lexico-semantic changes in all texts compared to the human posteditors. ChatGPT improved lexical richness over the machine translation for all texts. The obtained BERTScore values indicate that the meaning of the text is less preserved in ChatGPT's versions compared to those of the 'average professional literary translator' (RQ2).

When analysing the types of changes made by post-editors, we clearly see that posteditors mainly make lexico-semantic and stylistic changes, as was the case in Macken et al.'s study (2022). Looking more closely at the types of changes made by individual post-editors, we can see that there is a great deal of variation between different post-editors. A high degree of individual variation between professional translators during revision has been observed in previous studies (Daems and Macken, 2020) and can be attributed to the individual style of professional translators. The only striking difference between ChatGPT and professional literary translators is that it replaces words with synonyms more often than human post-editors (RQ3).

With only 18 accuracy errors and 29 fluency errors, the neural machine translation system did an excellent job. Most errors were solved during post-editing. Looking at the MT quality labels, we can conclude that ChatGPT solves fewer errors in the machine-translated texts and introduces more problems compared to professional literary translators (RQ4).

This study aimed to provide insights into the capabilities and limitations of ChatGPT for automatic post-editing of literary machine translation. Overall, ChatGPT proved to be a more aggressive post-editor than the professionals, making too many changes to the machine-translated text, despite being explicitly instructed not to do so in the prompt ("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"). It also corrected fewer errors, introduced more problems and deviated more from the meaning of the target text. Nevertheless, ChatGPT corrected most of the errors and provided meaningful edits.

While fully automatic post-editing with Chat-GPT is not yet feasible, and probably not desirable from an ethical point of view, AI tools based on large language models can generate highquality post-editing suggestions. As such, they can certainly complement the toolkits of professional translators. A promising direction that deserves further investigation is to have human translators work directly on texts that have been automatically post-edited by AI. This could help to leverage the strengths of both human and machine skills.

#### References

- Castilho, Sheila and Natália Resende. 2022. Posteditese in literary translations. *Information*, 13(2):66.
- Daems, Joke and Lieve Macken. 2020. Post-editing human translations and revising machine translations : impact on efficiency and quality. In Koponen, Maarit, Brian Mossop, Isabelle Robert, and Giovanna Scocchera, editors, *Translation revision and/or post-editing : industry practices and cognitive processes*, pages 50–70. Routledge.
- Daems, Joke. 2022. Dutch literary translators' use and perceived usefulness of technology : the role of awareness and attitude. In Hadley, James Luke, Kristiina Taivalkoski-Shilov, Carlos Teixeira, and Antonio Toral, editors, *Using technologies for creative-text translation*, Routledge Advances in

Translation and Interpreting Studies, pages 40–65. Routledge.

- do Carmo, Félix, Dimitar Shterionov, Joss Moorkens, Joachim Wagner, Murhaf Hossari, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2021. A review of the state-of-the-art in automatic postediting. *Machine Translation*, 35:101–143.
- 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.
- Guerberof-Arenas, Ana and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.
- Guimarães, Nuno, Ricardo Campos, and Alípio Jorge. 2024. Pre-trained language models: What do they know? Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(1):e1518.
- Hu, Kaibao and Xiaoqian Li. 2023. The creativity and limitations of ai neural machine translation: A corpus-based study of deepl's english-tochinese translation of shakespeare's plays. *Babel*, 69(4):546–563.
- Kenny, Dorothy and Marion Winters. 2020. Machine translation, ethics and the literary translator's voice. *Translation Spaces*, 9(1):123–149.
- Kolb, Waltraud. 2023. 'I Am a Bit Surprised': Literary Translation and Post-Editing Processes Compared. In Rothwell, Andrew, Andy Way, and Roy Youdale, editors, *Computer-Assisted Literary Translation*, pages 53–68. Routledge.
- Lardilleux, Adrien and Yves Lepage. 2017. Charcut: Human-targeted character-based mt evaluation with loose differences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 146–153.
- Li, Bo. 2023. Ethical issues for literary translation in the era of artificial intelligence. *Babel*, 69(4):529–545.
- Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455– 463.
- Macken, Lieve, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference* of the European Association for Machine Translation, pages 101–110, Ghent, Belgium, June. European Association for Machine Translation.

- Matusov, Evgeny. 2019. The challenges of using neural machine translation for literature. In Hadley, James, Maja Popović, Haithem Afli, and Andy Way, editors, *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, August. European Association for Machine Translation.
- Moorkens, Joss, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.
- 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.
- Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023.
  Robust speech recognition via large-scale weak supervision. In Krause, Andreas, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul.
- Radhakrishnan, Srijith, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the* 2023 Conference on Empirical Methods in Natural Language Processing, pages 10007–10016, Singapore, December. Association for Computational Linguistics.
- Raunak, Vikas, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009– 12024, Singapore, December. Association for Computational Linguistics.
- Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702.
- Rothwell, Andrew, Andy Way, and Roy Youdale. 2023. *Computer-Assisted Literary Translation*. Taylor & Francis.
- Ruffo, Paola, Joke Daems, and Lieve Macken. 2023. Developing user-centred approaches to technological innovation in literary translation (DUAL-T).

In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 491–492, Tampere, Finland, June. European Association for Machine Translation.

- Ruffo, Paola, Joke Daems, and Lieve Macken. 2024. User testing of three literary translation workflows: measured vs perceived effort when using a word processor, a cat tool, and a machine translation postediting (MTPE) platform. Manuscript submitted for publication.
- Ruffo, Paola. 2022. Collecting literary translators' narratives : towards a new paradigm for technological innovation in literary translation. In Hadley, James Luke, Kristiina Taivalkoski-Shilov, Carlos Teixeira, and Antonio Toral, editors, *Using technologies for creative-text translation*, Routledge Advances in Translation and Interpreting Studies, pages 18–39. Routledge.
- Şahin, Mehmet and Sabri Gürses. 2019. Would MT kill creativity in literary retranslation? In Hadley, James, Maja Popović, Haithem Afli, and Andy Way, editors, *Proceedings of the Qualities of Literary Machine Translation*, pages 26–34, Dublin, Ireland, August. European Association for Machine Translation.
- Şahin, Mehmet and Sabri Gürses. 2021. English-Turkish literary translation through human-machine interaction. *Tradumàtica tecnologies de la traducció*, (19):171–203.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Taivalkoski-Shilov, Kristiina. 2019. Ethical issues regarding machine(-assisted) translation of literary texts. *Perspectives*, 27(5):689–703.
- Tezcan, Arda, Veronique Hoste, and Lieve Macken. 2017. Scate taxonomy and corpus of machine translation errors. In Pastor, Gloria Corpas and Isabel Durán-Muñoz, editors, *Trends in E-tools and resources for translators and interpreters*, volume 45 of *Approaches to Translation Studies*, pages 219– 244. Brill — Rodopi.
- Tezcan, Arda, Joke Daems, and Lieve Macken. 2019. When asport'is a person and other issues for nmt of novels. In *Machine Translation Summit XVII*, pages 40–49. European Association for Machine Translation.
- Toral, Antonio, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5.
- Toral, Antonio, Andreas Van Cranenburgh, and Tia Nutters. 2024. Literary-adapted machine translation in a well-resourced language pair: Explorations with

more data and wider contexts. In *Computer-Assisted Literary Translation*, pages 27–52. Routledge.

- 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.
- Vanroy, Bram, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken. 2021. Metrics of syntactic equivalence to assess translation difficulty. In Carl, Michael, editor, *Explorations in empirical* translation process research, volume 3 of Machine Translation: Technologies and Applications, pages 259–294. Springer.
- Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 499–500, Tampere, Finland, June. European Association for Machine Translation.
- Webster, Rebecca, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics*, 7(3):32.
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

#### Appendix A. System prompt

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

You know that every literary 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 Dutch translation.

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

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

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

"Faithfulness to the style of the original author" in literary 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 aesthetic intentions or effects that are evident in the source text.

You will identify any stylistic deviations in the Dutch 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 Dutch 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 input contains offensive or obscene words, translate them faithfully
- 12. If the translation fails to convey the meaning of a large part of the input sentence, you include the translation for the missing part.

## Appendix B. User prompt

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

English text:

<English text comes here>

Dutch translation:

< Machine translation of the English text comes here>

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

# Guidelines to annotate MT errors and post-editing changes

This manual annotation task consists of three separate subtasks:

- 1. Label all errors in the neural machine translation output according to predefined error taxonomy
- 2. Label all post-editing changes according to a predefined linguistic typology
- 3. Assess the necessity of the post-editing changes

To be able to assess the quality of a neural machine translation system all errors in the MT output will be indicated and labelled.

For each segment you will first label all accuracy and fluency errors according to a predefined error taxonomy.

En: It's fine, I was going to be heading that way anyway — I've been meaning to swing by Rome to get some garden shears, too. Anything else ? I can call you and check when I get to Rome."

MT: Geeft niet, ik was toch al van plan om die kant op te gaan. Ik wilde ook nog even langs Rome om een tuinschaar te halen. Verder nog iets? Ik kan je bellen en het controleren[acc-mistra-ws] als ik in Rome ben."

In a second step you label the post-editing changes. All post-editing changes will be categorized from two independent perspectives. You first classify the post-edits based on a text-linguistic typology. In a second assessment you judge the necessity of the post-edits in terms of translation quality. You decide whether the postedit was (1) necessary to correct an MT error or for consistency reasons, (2) could be considered a preferential change or (3) an undesirable/unneccessary edit.

To help you spot the differences between the MT output and the post-edited version or the post-edited and the revised version, you are provided with an html-document in which the differences are highlighted.

T01\_EN.tok.sent.txt I can call you and check when I get to Rome . " T01\_MT.tok.sent.txt Ik kan je bellen en het controleren als ik in Rome ben . " P10\_WF03\_T01.tok.sent.txt Ik kan je bellen en het nog eens vragen als ik in Rome ben . '

# Error taxonomy to annotate MT errors

## ACCURACY ERRORS

mistra-mw	Mistranslation multiword: incorrect translation (often too literal) of a multi-word expression such as an idiom, a proverb, a collocation, a compound or a phrasal verb.
mistra-ws	Mistranslation word sense: incorrect translation. The MT refers to a different (and a wrong) sense of the source.
mistra-oth	Mistranslation other: incorrect translation (other cases)
dnt	Do not translate: source content is unnecessarily translated into target language
	when it should have been left untranslated.
untra	Untranslated: source text is not translated (but was copied to MT) when it should
	have been translated.
add	Addition: MT text that is not present in the source sentence.
omi	Omission: source content that cannot be found in MT.
cap-punct	Error related to capitalization or punctuation (not transferred correctly to the MT)

# FLUENCY ERRORS

coh-disc	Coherence discourse marker: the conjunction or linking word expresses a strange relationship
coh-tense	Coherence tense: the tense of the verb is wrong/illogical in the context of the rest of the sentence/text
coh-coref	Coherence coreference: mismatch between entities, e.g. feminine pronoun to refer to a male person
coh-oth	Coherence other: other coherence problem (lack of logical structure, confusing relationships)
lex	Lexicon: lexical element does not entirely fit in the Dutch sentence
gra-synt	Grammar & syntax: anything that does not follow the grammatical or structural rules of the Dutch language
styl-disf	Style disfluent construction: The sentence / constituent is not grammatically incorrect, but it is nonetheless very difficult to read, it could be translated in a much more idiomatic way
styl-rep	Style repetition: the same or a very similar word/expression is used more than once
styl-oth	Style other: other stylistic or register-related problems
spel-comp	Spelling compound
spel-cap	Spelling capitalization
spel-oth	Spelling other
punct	Punctuation problem

# Text-linguistic typology to annotate the post-editing/revision actions

# LEXICO-SEMANTIC

add	Addition of a meaningful element (other than Lex-sem- coh or lex-sem-expl/spec)	Uiteindelijk was deze niet voor haar → uiteindelijk was deze plek niet voor haar
coh	A coherence marker has been added, or a pronoun or a conjunction has been replaced by a proper name or a more specific conjunction to improve the coherence of the text.	een bezinestation → zo'n benzinestation
expl-spec	Explicitation: information that could be derived from the source text and/or the MT output has been made explicit or the translation has become more specific, e.g. by using a hyponym	vliegtuigtijdschrift → een tijdschrift aan boord van een vliegtuig; om te halen → om te kopen
impl-	Implication: information has been made more implicit in	om zes uur → om zes; landkaarten → kaarten
vague	the translation or the translation has become vaguer, e.g.	
	by using a hyperonym	
del	Deletion: meaningful element of the source text or a	Hij glimlachte maar zei dat hij het er niet mee
	relation that was present in both the ST and MT output	eens was. → Hij lachte wel, maar was het er niet
	has been deleted	mee eens.
syn	Synonym: word or phrase from the MT output has been substituted for a synonym of that word/phrase	boekwinkel $\rightarrow$ boekhandel
coll-idiom	word or phrase from the MT output has been substituted	klaarde de regen op → trok de regen weg; een
	for a better collocation or more idiomatic expression	of twee boeken $\rightarrow$ een boek of twee
other	Other lexico-semantic change	beter → aantrekkelijker

# SYNTAX & MORPHOLOGY

agree	Agreement: change to solve an agreement problem	het soort dat $\rightarrow$ de soort die
number	Number: change in grammatical number (singular	tuinschaar → tuinscharen
	becomes plural and vice versa)	
dim	Diminutive: change from base form to diminutive or vice	stad → stadje
	versa	_
tense	Tense: change of verb tense	heb gezien → had gezien
other	Other syntactic or morphological change	zou ze schuif → verplaatste ze

### STYLE

order	Word order: the order of words has been altered	erin verdween $\rightarrow$ verdween erin
structural	Structural change: structural change to the MT output	wat vaak genoeg gebeurde → en dat gebeurde nogal eens
short	Shorter: the translation has become shorter / more concise	maakten wandelingen $\rightarrow$ wandelden
split	Split sentence: a MT sentence has been split into several sentences.	Wacht, wat is fauna? → Fauna? Wat is dat?
merge	Merged sentence: MT sentences have been merged into one sentence	Oké. Dat is prima. → Oké, goed hoor.
other	Other stylistic change	Ze keek naar alles → Zij keek naar alles.

# SPELLING & PUNCTUATION

cap	Capitalization: change capitalization of a word	Kerstavond → kerstavond
cmp	Compound: correction of a wrongly spelled compound	$tenminste \rightarrow ten minste$
lw-punct	Linking word & punctuation: replacement of a linking	, → en
	word by a punctuation mark or vice versa	
add	Addition of a punctuation mark	terwijl → , terwijl
del	Deletion of a punctuation mark	prachtig, en → prachtig en
other	Other spelling or punctuation change	Veertig $\rightarrow$ '40; " $\rightarrow$ '

# Categorization of post-editing changes from the perspective of correctness and necessity (translation quality)

MT-err	Edit to solve a problem indicated as an MT error.
Consist	Edit to achieve consistency between the segment and other segments in the text
Pref	Other (preferential) edits
Undes	Edits that deteriorate translation quality