Literary translation as a three-stage process: machine translation, post-editing and revision

Lieve Macken, Bram Vanroy, Luca Desmet and Arda Tezcan
LT[3], Language and Translation Technology Team
Ghent University
Belgium
{firstname.lastname}@ugent.be

Abstract

This study focuses on English-Dutch literary translations that were created in a professional environment using an MT-enhanced workflow consisting of a three-stage process of automatic translation followed by post-editing and (mainly) monolingual revision. We compare the three successive versions of the target texts. We used different automatic metrics to measure the (dis)similarity between the consecutive versions and analyzed the linguistic characteristics of the three translation variants. Additionally, on a subset of 200 segments, we manually annotated all errors in the machine translation output and classified the different editing actions that were carried out. The results show that more editing occurred during revision than during post-editing and that the types of editing actions were different.

1 Introduction

With the current quality of neural machine translation (NMT) systems, the question arises whether post-editing NMT output is a viable alternative to human translation for real large-scale translation tasks. In this paper, we present the results of a case study on literary translations. We collaborated with Nuanxed, a book translation company, which uses an MT-enhanced workflow consisting of a three-stage process of automatic translation followed by post-editing and revision.

In this case study, we compare three successive versions of the target texts as they proceed through the translation process: the machine translation, the post-edited and the (mainly) monolingually revised translation. We used different automatic metrics to measure the (dis)similarity between the consecutive versions and to analyze the linguistic characteristics of the three translation variants. To assess the quality of the MT output and to get an insight into the editing actions that were carried out, a fine-grained manual annotation was carried out on a subset of 200 segments.

2 Related research

Although employing Machine Translation (MT) for more creative text types such as literature may not seem to be a natural fit, several researchers looked into the feasibility of using MT for literary texts, first with statistical (Besacier and Schwartz, 2015; Toral and Way, 2015) and later with neural machine translation systems (Toral and Way, 2018; Kuzman et al., 2019; Toral et al., 2020).

To assess the usefulness of MT for literary texts, researchers often compare raw (unedited) machine translations of literary texts with their human-translated (HT) counterparts. Three successive studies were conducted to assess the quality of generic NMT systems for English-Dutch literary texts, the language pair we also focus on in this study (Tezcan et al., 2019; Fonteyne et al., 2020; Webster et al., 2020). According to these studies, the main issues found in literary NMT are different types of mistranslations, coherence issues, and style & register problems. The percentage of NMT sentences that were free of errors varied and averages ranged from 44% to 25% in different studies, with a notable exception of the NMT version of Jane Austen's *Sense and Sensibility* in which only 5% of all machine-translated sentences were error-free. It thus seems that NMT quality is highly de-

pendent on the source text and that some literary texts are more challenging for automatic translation systems than others. When comparing linguistic characteristics of NMT and HT, the machine translations were less lexically rich, showed a lower level of lexical and semantic cohesion and tended to follow the structures of the source sentences more closely, whereas the human translations showed the ability to deviate from the source structure (Webster et al., 2020). It is thus clear that in order to obtain high-quality literary translations, human intervention in the form of a post-editing (PE) step is needed.

Daems and colleagues investigated whether post-edited MT output differs from HT in English-to-Dutch texts (2017), and called this (dis)similarity between PE and HT 'post-editese'. The authors did not find proof of this. Neither humans nor computer systems were able to distinguish between the two types of translation, although the authors note that this may be due to a rather limited dataset size. They considered features such as average word and sentence length, average tf-idf, perplexity, type-token ratio, number of verb phrases/passives, parse tree depth, and so on. Working with different language combinations and architectures, Toral (2019) came to a different conclusion. He found that PE is indeed notably different from HT in terms of a limited set of features, namely lower lexical variety (type-token ratio) and density (content words ratio), sentence length inference of ST, and POS sequence perplexity. It must be noted however, that not only the language pairs differed in the studies of Daems et al. (2017) and Toral (2019), and hence the MT quality, but also the proficiency level and the degree of postediting that was requested (light or full). It is thus difficult to draw conclusions about the existence of post-editese.

Neither Daems and colleagues nor Toral investigated post-editese in literary texts. Castilho and Resende (2022), however, found some evidence for post-editese in literary translation of English into Brazilian Portuguese but note that such observations depend on the literary genre. Statistical differences between HT and PE were found, especially in the thriller genre (The Girl on the Train; TGOTT) and only barely in children's literature (Alice's Adventures in Wonderland; AW), which is explained by the emphasis on the author's figurative style in the latter book. Post-editese effects for lexical density (simplification), length ratio (text length of PE vs HT; explicitation), personal pronoun ratio (explicitation), and convergence (translated texts are more similar to each other than original texts are to each other) (partially) were found for TGOTT, but only evidence for convergence was discovered in AW.

Guerberof-Arenas and Toral (2020) focused on creativity, one of the distinguishing features of literary texts. They analyzed both creativity and acceptability in MT, PE and HT texts. The translation and post-edited version were created by two professional translators specialized in literary translation. To quantify acceptability they counted the number of errors in the different translations. Interestingly, they found that the HT translations contained slightly more errors than the PE translations, with HT having lowest number fluency errors and PE having the lowest number of accuracy errors. To measure translational creativity they selected 48 English source sentences that contained units of high creativity potential (in which translators most likely depart from the source text structure): metaphorical expressions, imagery and abstraction, idioms, comparisons, verbal phrases or complex syntactic structures. They quantified creativity by investigating creative shifts, which can be defined as "abstracting, modifying or concretising source text ideas in the target text" (Bayer-Hohenwarter, 2011, p. 663). When comparing the three types of shifts in the HT and PE condition, no major differences were found for abstractness and modification, but the HT contained more instances of concretisation.

The work of Daems et al. (2017) mentioned above built on earlier work on 'translationese' (Gellerstam, 1986). In the field of translation studies, it is generally accepted that a translated text is different from an original text in the same language, almost as if it is a genre on its own. Baker (1993, p. 243-245) discusses six "universal features of translation" that may mark translated texts: explicitness, disambiguation and simplification, a focus on grammaticality (especially in interpreting), avoiding repetitions by omission or rewording, exaggeration of target language features, and finally unexpected distributions of certain language features with respect to the source text (ST) and original texts in the target language. This phenomenon where translation is considered different from original text is often referred to as 'transla-

tionese', and researchers have investigated its existence, both via human perception and computer models.

Kruger (2017), however, made an interesting point by suggesting that some of these translationese features might also be the consequence of the editorial intervention subsequent to translation. Evidence for features commonly denoted as translationese such as increased explicitness, simplification and normalisation were also found in a parallel corpus of monolingual edited texts and their unedited counterparts. It thus seems that translation and linguistic editing share certain similarities.

In the publishing sector, it is quite common that many actors play a role in the production of a translation. For example, Moe and colleagues (2021) explain that in Slovenia language revisors correct the grammar and style of translations, usually without having access to the source texts. They may change the text's structure, syntax and word order and replace words and phrases to make the text more suitable. Different terms are used to refer to this process: linguistic revision/editing, copy-editing and translation revision. Mossop states that both editing and revision "involve checking linguistic correctness as well as the suitability of a text's style for its future readers and for the use they will make of it" (Mossop et al., 2020, p. 1). Translation revision can be considered the broader term as it also comprises a bilingual component, although different revision procedures exist (Ipsen and Dam, 2016) and the process can be predominantly monolingual (the revisor focuses on the target text and only refers back to the source text if a passage is problematic) or bilingual (the revisor systematically compares the source and target text).

3 Method

3.1 Data

The data we received from the company consists of an English novel (68,762 source words) and three Dutch translations: the machine translation generated by DeepL[1], the post-edited (PE) version and the revised (REV) version. An NDA was signed between the researchers and the company. The post-editor worked in a standard CAT tool that divides the text in sentences and displays both the source and target segments side-by-side. The post-editor thus worked on a segment-by-segment basis to edit the machine translation suggestions. The revisor received the post-edited translation in Microsoft Word. Revision in this case is mainly a monolingual process, which aims at improving the reading experience or, in the case of audio-books, the listening experience. The revisor could consult the source text whenever there was a need. The post-editor was Flemish, the revisor Dutch. Both the post-editor and the revisor were paid by the hour, so there was no real time pressure. For this study, we used the first chapter of the book. We used YouAlign[2] to align all versions at sentence level and manually verified the sentence alignments. The data set consists of 578 aligned segments (7,921 source words; 9,419 source tokens).

3.2 Automated evaluation

Automatic evaluation metrics for MT play a central role in rapid assessment of MT quality. A key characteristic of almost all automatic MT evaluation metrics is that they assess MT quality by calculating the similarity between the MT output to a reference translation. We use automatic MT evaluation metrics with a different goal, namely to measure the (dis)similarity between the consecutive versions of the texts produced in the target language, i.e. the machine translations (MT), the post-edited (PE) and the revised translations (REV).

In literature, we can find various metrics that differ with regard to the approach they take to measure the similarity between two texts. To obtain a nuanced picture, we use a variety of MT evaluation metrics, which focus on different dimensions, such as Translation Edit Rate (TER) (Snover et al., 2006), CharCut (Lardilleux and Lepage, 2017), COMET (Rei et al., 2020) and BERTScore (Zhang et al., 2019). While CharCut and TER measure the amount of editing required to transform one text into another in terms of character- and token-level edit operations respectively, COMET and BERTScore target the semantic aspect of translation quality by calculating the distance between vector representations of sentences and tokens, respectively. Additionally, we use ASTrED (Vanroy, 2021), which has been originally proposed to quantify syntactic similarity between a source sentence and its (human) translation. By working on a deeper linguistic level, ASTrED compares the

[1] https://www.deepl.com/, translations created end 2021

[2] https://youalign.com/

edit distance between the dependency structures of two sentences, while also taking word alignment information into account. Word alignments were automatically created with AwesomeAlign (Dou and Neubig, 2021). For this metric, we only used sentences that were translated as single sentences, without splitting or merging (156 in total of the manually verified subset, see below).

Besides analysing the degree of similarity between the different versions of the target texts, we were also interested in how well the lexical richness of the original novel was captured in the three versions. With the assumption that an increase in number of types with respect to number of tokens indicates a greater lexical richness in a given text, we calculated type-token ratio (TTR) and Mass index (Mass, 1972), which, unlike TTR, is not sensitive to text variations in text length. We calculated TTR and Mass index values of each document separately.

Word translation entropy, finally, is a formula to measure lexical variation by taking into account for each source word how many translations it has or can have in a given corpus based on its word alignments, and the distribution of those translations (Schaeffer et al., 2016). Put differently, it quantifies how certain or unambiguous the translation of a token is. A higher value indicates more uncertainty, i.e., a less straightforward lexical choice. In this study, we use this formula to measure average word translation entropy (AWTE) on document level, by measuring entropy for each source word (English) of the first chapter of the novel taking into account the three different translations in Dutch.

All data sets were tokenized prior to performing automatic measurements, using the Stanza Toolkit (Qi et al., 2020). While the MT metrics were calculated using the data with the original casing, to obtain more accurate results, we used the lowercased version of each document to measure lexical richness.

3.3 Manual evaluation

The first 200 segments (3,222 source tokens) of the data set were manually annotated. The manual annotation task consists of three separate sub-tasks: labelling of errors in the MT output, labelling of PE and REV actions and labelling of remaining errors in the final translation. The first sub-task allows us to assess the quality of the NMT system on the literary text; the second and third sub-tasks give us insights in the post-editing and revision actions and allow us to assess the quality of the final translation.

To label the MT errors we used the SCATE MT error taxonomy tailored to the annotation of literary MT on document level (Tezcan et al., 2019). This taxonomy is based on the well-known distinction between accuracy and fluency and is hierarchical in nature. According to this taxonomy accuracy and fluency errors can be annotated on the same text span, e.g. when a mistranslation error (accuracy error) causes a logical problem (fluency error). However, to minimize the annotation workload, we decided to only label the accuracy errors in this case. We also reduced the number of error labels by merging a number of error categories that were present in the original taxonomy.

To classify the PE and REV actions from a linguistic perspective, a classification scheme was devised based on the work of Desmet (2021) and Vandevoorde et al. (2021). The categorization scheme contains four main categories (*lexico-semantic*, *syntax & morphology*, *style* and *spelling & punctuation*), which are further subdivided in subcategories (see Table 1). All PE and REV actions were also labelled from a translation quality perspective. We distinguished the following four categories to label a post-editing action for its correctness and necessity: *MT error correction*, *consistency*, *preferential* and *undesirable* change. When labelling revision actions, the label *PE error correction* was added to this list to indicate undesirable changes made by the post-editor that were corrected by the revisor. In the final translation we also labelled all MT and PE errors that were not corrected.

Detailed annotation guidelines were drafted to ensure consistency between annotators. To facilitate the manual annotation process, the WebAnno[3] annotation tool was used. Figure 1 shows a full example of the annotation process. Two errors were labelled in the MT version: the phrase *met een opgewonden glimlach* (*with an excited smile*) is placed in a wrong position in the clause and *glinsteren* is a wrong translation for *glimpse*. The post-editor corrected these two MT errors and made two preferential changes: *zojuist* was replaced by a synonym (*net*) and *the red of Rudolph's nose* is changed into *Rudolph's red nose*. The revisor

[3] https://webanno.github.io/webanno/

Lexico-semantic	Syntax & morphology
Addition	Agreement
Coherence marker	Number
Explicitation	Diminutive
Implicitation	Comparison
Deletion	Tense
Synonym	Other
Collocation & idiom	
Specific	**Spelling & punctuation**
Vague	Capitalization
Other	Compound
	Linking word punctuation
Style	Punctuation linking word
Word order	Punctuation added
Structural change	Punctuation deleted
Shorter	Other
Split sentence	
Merged sentence	
Other	

Table 1: Linguistic typology

made additional changes: *glimlach* was replaced by a diminutive *lachje*, the proper name *Rudolph* was spelled in Dutch, and the preposition *tussen* was replaced by another preposition *door*. The revisor also made some structural changes and split the long sentence and rephrased the last clause making it a less literal translation.

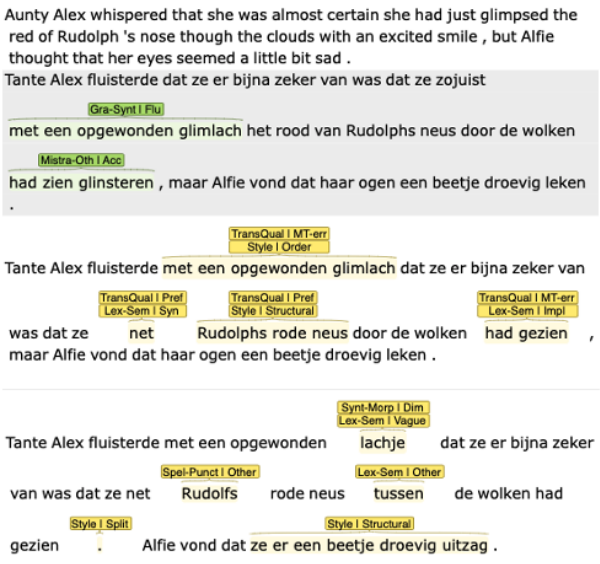

Figure 1: Example of annotations made in Webanno

To help the annotators to spot the differences between the MT output and the PE version or the PE and the REV, we used Charcut (Lardilleux and Lepage, 2017), which creates an HTML document in which differences between two versions are visualized (see Figure 2).

S2_EP1-DeepL.tok.txt

Tante Alex fluisterde dat ze er bijna zeker van was dat ze **zojuist met een opgewonden glimlach het rood van** Rudolphs neus door de wolken had zien **glinsteren** , maar Alfie vond dat haar ogen een beetje droevig leken .

S2_EP1-PE.tok.txt

Tante Alex fluisterde **met een opgewonden glimlach** dat ze er bijna zeker van was dat ze **net** Rudolphs **rode** neus door de wolken had **gezien** , maar Alfie vond dat haar ogen een beetje droevig leken .

Figure 2: Example of Charcut visualizations (MT–PE)

4 Results

4.1 Automated evaluation

First, we use five automatic metrics that target different aspects of (dis)similarity, as described in Section 3.2, between the consecutive versions of the texts produced in the target language. The results are presented in Table 2.

	MT-PE	PE-REV	MT-REV
CharCut ↓	0.126	0.148	0.240
TER ↓	0.215	0.251	0.355
BERTScore ↑	0.941	0.936	0.900
COMET ↑	0.835	0.765	0.620
ASTrED ↓	0.305	0.307	0.332

Table 2: Overview of automated evaluation results. Up arrow: higher value means more similar; down arrow: lower value means more similar.

According to all automatic metrics used in this analysis, each consecutive modification made to the MT output, i.e. post-editing and revision, results in observable differences for all measured aspects, namely the degree of editing (CharCut and TER), semantic (BERTScore and COMET) and syntactic (ASTrED) similarity. Moreover, the level of (dis)similarity between the different document pairs seems to be different. As shown by the results of all five metrics, the similarity between the MT output and post-edited version (MT-PE) is higher compared to the similarity between post-edited and revised translations (PE-REV). Moreover, the similarity between the MT output and the revised translations is the lowest in comparison to the analyses made on other document pairs.

To measure lexical richness, we calculated TTR and Mass index for the chapter in English (SRC) and all three versions of the translated text in Dutch. These results are provided in Table 3, together with the unique and total number of tokens for each text.

These results show that, compared to the original text in English, all three translations in Dutch have a higher number of tokens and unique tokens.

	SRC	MT	PE	REV
# unique tokens	1820	1922	1962	2022
# tokens	9419	9285	9429	9632
TTR	0.182	0.196	0.198	0.199
MASS	0.020	0.020	0.019	0.019

Table 3: Summary of lexical richness measures

Moreover, these numbers increase with a similar ratio after each consecutive modification made on the MT output, resulting in a difference of 347 tokens and 100 unique tokens between the revised translations and the MT output. The post-editing and revision steps also make the translations lexically more rich, as observed by the TTR measurements. TTR is also observed to be higher in all three versions of the target text compared to the original novel. However, these observations are not confirmed by the Mass index scores, which indicate similar levels of lexical richness in all four documents.

In a final analysis we measure AWTE by comparing the MT output, the PE and REV translations to the original novel in English. To increase our confidence about the differences between the AWTE values (as word alignment was an automatic process), for each comparison, we use translation options with the minimum probability threshold of 0.01 and we repeat the calculations by increasing the minimum frequency threshold for the set of source words (up to 10, which covers 64% of all source tokens) we take into consideration. While a minimum threshold frequency of 1 covers all the source words in the source text, a threshold of n calculates AWTE only for the subset of source words that appear at least n times in the source text. The AWTE measurements made on the three document pairs are shown in Figure 3.

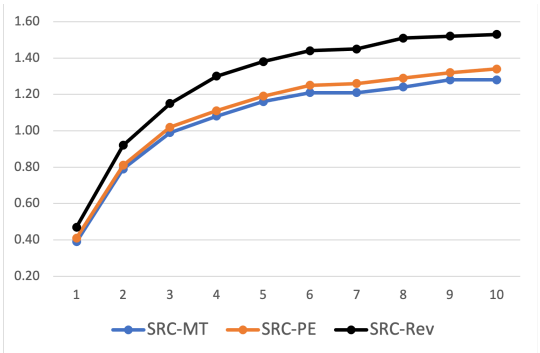

Figure 3: Average word translation entropy values

These results show that, for all minimum frequency thresholds, AWTE increases with each consecutive modification made to the MT output. Furthermore, the revision step increases AWTE to a larger extent, compared to post-editing, resulting in a higher level of uncertainty on average for the lexical choices made for translating source words during this operation.

4.2 Manual evaluation

Given that DeepL is a generic MT system and thus not tailored to literary texts, the overall quality of the machine-translated text can be deemed relatively good. The subset contained 275 MT errors, which is on average 1.38 error per sentence. Fifty-five sentences (27.5%) were free of errors. Table 4 shows the distribution of the 275 MT errors. In terms of accuracy, 152 errors were found, half of which were mistranslations. The NMT system wrongly translated words (e.g. *short crust pastry – korstdeeg*) and tenses (e.g. *was rolling out – rolde . . . uit*), or used a translation of a word or phrase that was incorrect in the given context (word sense e.g. *ports – poorten* (meaning: *porto's*)), which sometimes led to illogical constructions, or even changed the meaning of the entire sentence. The machine moreover appeared to have difficulties translating multiword expressions and idioms as well (e.g. *going to see a man about a dog* was translated literally). The second largest category was capitalization and punctuation errors, which almost solely consisted of missing quotation marks that were not copied from source to target text by the machine. Also quite often, source text information was omitted (e.g. the verb *to sprinkle* was deleted in *as Fergus reminded him* ***to sprinkle*** *– zoals Fergus hem herinnerde)*; additions, on the other hand, did not occur in the subset.

In terms of fluency, the most problematic category was spelling and punctuation. The majority of these errors were related to quotation marks, missing commas and capitalization problems (*kerstman (Santa)* starts with a lowercase letter whereas *Kerstmis (Christmas)* starts with a capital letter in Dutch, which is confusing for the NMT system). Stylistic problems were often-occurring as well, when the MT contained disfluent constructions that are not wrong from a grammatical point of view, but could nonetheless be translated in a more idiomatic and fluent way. These were in most cases very literal translations of English constructions (e.g. *said Fergus with a*

laugh – zei Fergus met een lach). Lastly, a number of lexical problems were found: when a word was not an entirely wrong translation of the source word in the context, but nevertheless did not entirely fit in the Dutch sentence either (e.g. *the* ***glow*** *of his screen – het* ***schijnsel*** *van zijn scherm vs. de* ***gloed*** *van zijn scherm*).

Accuracy	**152**	**Fluency**	**123**
Mistranslation	77	Coherence	13
Multiword	15	Discourse marker	1
Word sense	15	Coreference	2
Other	47	Tense	0
Addition	0	Other	10
Omission	21	Lexicon	18
Untranslated	7	Grammar & syntax	10
Do not translate	1	Style	35
Capitalization & punctuation	46	Disfluent	33
		Repetition	0
		Other	2
		Spelling & punctuation	47
		Capitalisation	13
		Compound	4
		Punctuation	23
		Other	7

Table 4: MT errors in the manually annotated subset of 200 segments

Table 5 shows the PE and REV quality label distribution. The revisor carried out more editing actions (569) than the post-editor (501), and these in themselves were of a different nature. While the post-editor focused on correcting MT errors (219; 44% of all post-edits), e.g. by adding ST information missing from the MT output, and on making preferential improvements (224), the revisor mainly sought to further improve the overall quality and readability of the text: 492 (86%) of the revisor's edits were preferential changes to make the text more coherent and understandable (by means of explicitations and structural changes as well as splitting of sentences; see Figure 4 for details). Often an MT error was corrected by the post-editor and further improved by the revisor, as can be seen in the example in Figure 1: the post-editor corrected the word order error of the MT and made sure that phrase *met een opgewonden glimlach* correctly modifies the verb. The revisor further improved the translation by replacing *glimlach* by the diminutive *lachje*.

Some MT errors were not spotted by the post-editor but corrected by the revisor, and most of the errors introduced during post-editing were corrected in the revision step as well. A very small number of MT errors (7) seeped through into the final text (e.g. *Christmas play – kerstspel* (*Christmas game*)), and 6 post-editor errors were left uncorrected (e.g. *buddy up – vrienden worden* (*became friends*; ST meaning: *to pair together with someone*)). Finally, 8 revisor changes were deemed undesirable, mostly due to the information presented in the final target text no longer being consistent with the information in the source text. As always some of these are, however, debatable. In the following example the subject of *saw* has been made implicit by the post-editor and was wrongly interpreted by the revisor:

- ST: *Aunty Alex also understood about all the things that Alfie could see and hear,* ***like when he saw the lady*** *who used to live upstairs at their old flat, until she died.*

- PE: *Tante Alex begreep ook alles wat Alfie kon zien en horen, zoals de mevrouw die boven in hun oude flat woonde, tot ze stierf.*
  (*Aunty Alex also understood everything that Alfie could see and hear,* ***like the lady*** *who lived upstairs in their old flat, until she died.)*

- REV: *Bovendien kon tante Alex alles horen en zien wat Alfie kon zien en horen, net zoals de mevrouw die boven hun oude flat gewoond had tot ze doodging.*
  (*Moreover, aunty Alex could hear and see everything that Alfie could see and hear,* ***just like the lady*** *who had lived upstairs from their old flat until she died.)*

As can be seen in Figure 4 both the post-editor and the revisor made lexico-semantic changes for the most part (45% and 44% respectively), of which using synonyms or other words are in the lead. Spelling and punctuation changes represent 24% of all post-edits and were mainly corrections of MT errors; of the revisor changes, 21% were spelling and punctuation changes, although these largely consisted of *mama/papa* being preferentially spelled into *mamma/pappa*. When we look in more detail at the different editing actions, it is clear that the revisor carried out different types of editing actions and made a lot of explicitations, split long sentences, made more structural changes (compared to the post-editor), added more coherence markers and made the translation sometimes more specific and sometimes more vague. These edits greatly improve the readability of the translation and tailor it to the target audience.

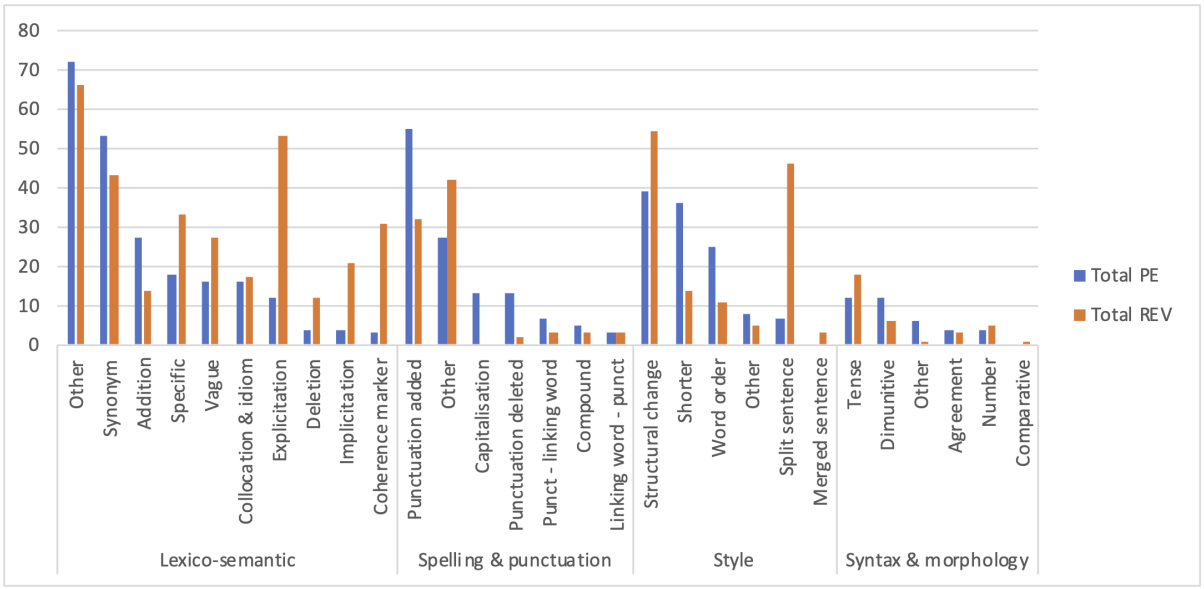

Figure 4: Linguistic classification of the post-editing (PE) and revision (REV) actions

Quality Label	PE	REV
Consistency	13	0
MT error correction	219	32
PE error correction	NA	37
Preferential	224	492
Undesirable	45	8
Total	**501**	**569**

Table 5: Quality labels assigned to the post-editing (PE) and revision (REV) actions

5 Discussion

In this paper, we examined the possibility of using an MT-enhanced translation workflow for the translation of literary texts in a real-life professional translation scenario. We examined three different versions of the target texts as they proceed through the translation process: the MT output, the post-edited version and the revised translation.

DeepL was used as MT engine to translate an English novel into Dutch. MT quality was in line with expectations with 27.5% error-free sentences. The three main error types were various kinds of mistranslations, disfluent sentence constructions and different types of spelling and punctuation problems. DeepL failed to correctly copy quotation marks from source to target, a problem that can potentially be fixed by applying a number of post-processing rules. Mistranslations and disfluent constructions have been reported in earlier research as the main error types and require more attention from the post-editor.

Forty-four percent of all post-editing actions were corrections of MT errors, 24% of all post-edits were preferential changes, 9% of all post-edits were labelled as 'undesirable'. Apart from adding missing punctuation marks, the post-editor mainly carried out lexico-semantic changes (replacing words with better alternatives or synonyms) and stylistic operations (restructuring MT fragments or coming up with shorter translation solutions). Most MT errors were solved in the post-editing step. Only 5.6% of all editing actions during revision were related to MT errors; another 5.5% were corrections of problems introduced during post-editing. The majority of the revisor's edits (86%) were thus preferential in nature. The revisor made slightly more edits than the post-editor. The revisor, just like the post-editor, mainly made lexico-semantic changes, but the subcategories were different. The revisor often made information and relations that the reader might be able to infer from the context explicit as can be seen from subcategories 'explicitation' and 'coherence marker' in Figure 4. The revisor also made a lot of stylistic changes and restructured fragments and even split sentences in 23% of all segments.

Post-editing and revision can be considered two different cognitive processes. Post-editing is by nature a bilingual process in which the post-editor can be primed both by the MT suggestion and the source segment. Moreover, as the post-editor worked in a traditional CAT tool, in which the text is segmented at sentence level, it might be more difficult to focus on the flow of the target text.

Revision was mainly a monolingual process, carried out in Microsoft Word, in which it is easier to focus on the translated text as a standalone product. It is remarkable, however, that the revisor carried out many edits that fall within two subcategories that are often considered as 'translationese', e.g. increased explictness (subcategories 'explicitation' and 'coherence marker') and simplification (subcategory 'split sentence'). We consider this as an indication that monolingual editing and translation indeed share certain similarities as Kruger (2017) suggested.

The automatic evaluation confirmed that more editing took place in the revision step than in the post-editing step. The degree of similarity between the MT, the PE and the REV version was assessed based on the amount of editing, and semantic and syntactic similarity measures. All measures confirmed that the degree of similarity between MT and PE was higher than the degree of similarity between PE and REV. The lowest similarity scores were obtained when comparing the MT with the revised version. As a side note we would like to point out that in MT research it is common practice to use automatic evaluation metrics to compare the MT output with an independent reference translation, often without knowing how this reference translation was created. It might as well be that the reference translation being used is the output of a two-stage process of human translation followed by revision, which, depending on the amount of editing that took place, may have altered the human translation to a large extent.

Another feature that has been widely studied in previous research is lexical richness. In this study, we quantified lexical richness by means of TTR, Mass index and average word translation entropy. Some results were inconclusive (higher TTR values, but lower or similar Mass index values). Average word translation entropy showed a clearer picture, with the revised version having the highest values. It thus seems that the revised version exhibits many characteristics that have been attributed to human translations: a higher degree of explicitation and simplification, more lexical variety and translations that deviate more from the source structure (compared to MT). This study, however, cannot provide a conclusive answer to the question of whether the implemented three-stage process of automatic translation followed by post-editing and revision is a viable alternative to human translation followed by revision. This can only be measured by means of comparative translation reception studies in which the reading (or listening) experience is measured.

One of the major limitations of this study is that we only had data of one post-editor and one revisor. Moreover, the post-editor and the revisor had different experience levels, with the post-editor having less experience in the literary domain. Studying the edits of two different persons most probably changes the distribution of the edits. It would therefore be interesting to replicate this study with more post-editors and more revisors and on different language pairs. In future work we also intend to zoom in on the sentences with high creativity potential as was done by Guerberof-Arenas and Toral (2020) and examine in more detail the creative shifts in the post-edited and revised version.

References

Baker, Mona. 1993. Corpus linguistics and translation studies. *Text and technology*, John Benjamins. 233–250.

Bayer-Hohenwarter, Gerrit. 2011. "Creative Shift" as a Means of Measuring and Promoting Translational Creativity *Meta: Translators' Journal*, 56(3):663–692.

Besacier, Laurent and Lane Schwartz. 2015. Automated Translation of a Literary Work: A Pilot Study. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Association for Computational Linguistics. 114-122.

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

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. Translationese and Post-Editese: How Comparable is Comparable Quality. *Linguistica Antverpiensia*, 16:89–103.

Desmet, Luca. 2021. *An exploratory study of professional post-edits by English-Dutch DGT translators.* Master's thesis, Ghent University.

Dou, Zi-Yi, and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, online. 2112–2128.

Fonteyne, Margot, Arda Tezcan and Lieve Macken. 2020. Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level.

Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France. 3790–3798.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. *Scandinavian Symposium on Translation Theory*, Lund.

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.

Ipsen, A. Helene and Helle V. Dam. 2016 Translation Revision: Correlating Revision Procedure and Error Detection. *HERMES - Journal of Language and Communication in Business*, 55: 143–156.

Kruger, Haidee. 2017. The effects of editorial intervention: Implications for studies of the features of translated language. In G. De Sutter, M.A. Lefer and I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions*, De Gruyter Mouton. 113–156.

Kuzman, Taja, Špela Vintar and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland. 1–9.

Lardilleux, Adrien, and Yves Lepage. 2017. CHARCUT: Human-Targeted Character-Based MT Evaluation with Loose Differences. *Proceedings of the 14th International Workshop on Spoken Language Translation*, Tokyo, Japan. 146–153.

Mass, Heinz-Dieter. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik 2, no. 8:73*.

Moe, Marija Zlatnar, Tamara Mikolič Južnič, and Tanja Žigon. 2021. Who determines the final version? The roles of translators, language revisers and editors in the publishing of a literary translation. *Across Languages and Cultures*, 22 (1):14–44.

Mossop, Brian, Jungmin Hong and Carlos Teixeira. 2020. *Revising and editing for translators* (4th edition). Routledge, Taylor and Francis: London; New York.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *ACL System Demonstrations*.

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

Schaeffer, Moritz, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl. 2016. Word translation entropy: Evidence of early target language activation during reading for translation. *New directions in empirical translation process research*. Springer, Cham, Switzerland. 183–210.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of the 2006 Conference of the Association for Machine Translation in the Americas*, Cambridge, MA, USA. 223–231.

Tezcan, Arda, Joke Daems and Lieve Macken. 2019. When a 'sport' is a person and other issues for NMT of novels. *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland. 40–49.

Toral, Antonio. 2019. Post-editese: An Exacerbated Translationese. *Proceedings of Machine Translation Summit XVII: Research Track*, Dublin, Ireland. 273–281.

Toral, Antonio, Antonio Oliver, and Pau Ribas Ballestí. 2020. Machine Translation of Novels in the Age of Transformer. J. Porsiel (Ed.), *Maschinelle Übersetzung für Übersetzungsprofis*, BDÜ Fachverlag. 276–295.

Toral, Antonio and Andy Way. 2015. Machine-assisted translation of literary text: A case study. *Translation Spaces*, John Benjamins. 4(2):240–267.

Toral, Antonio and Andy Way. 2018. What Level of Quality Can Neural Machine Translation Attain on Literary Text? In J. Moorkens et al. (Eds.), *Translation Quality Assessment: From Principles to Practice*, Springer. 263–287.

Vandevoorde, Lore, Roxana Weintraub and Marta Arabadjieva. 2021. *Sustained quality in Council translations: assessing the importance of human translation actions*. Council of the European Union. Translation Service.

Vanroy, Bram. 2021. *Syntactic Difficulties in Translation.* Ph.D. thesis, Ghent University.

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).

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