Metaphors in Literary Post-Editing: Opening Pandora's Box?

Aletta G. Dorst, Mayra O. Nas and Katinka Zeven
Leiden University Centre for Linguistics
{a.g.dorst, m.o.nas, k.zeven}@hum.leidenuniv.nl

Abstract

This paper investigates how post-editors of literary texts react and respond to the way metaphors have been translated by Neural Machine Translation (NMT) and Large Language Models (LLMs). The results show that one in three metaphors in the output were changed by the post-editors, demonstrating that the translation of figurative language is indeed problematic in literary MT (LitMT). The responses indicate that the post-editors were aware of overly literal translations, though mostly for multiword expressions. Moreover, at times they found it difficult to determine whether solutions were acceptable. They rated the overall quality of the MT output as quite poor and stated that the post-editing was more work and more effort than it would have been translating from scratch. This supports previous studies arguing that post-editing constrains translators in their creativity and diminishes their sense of text ownership.

1 Introduction

LitMT is no longer a fantasy, a perhaps equally fascinating and frightening scenario of science fiction. In 2023, Reedz was heralded as "the world's first fully AI powered publisher" (Reedz, 2023) and Nuanxed has been making a name for itself since 2021 by turning AI post-editing into its standard workflow to "give books the audience they deserve" by making "quality book translations [...] easy, fast and affordable"[1]. The same positive, almost idealistic promise of providing authors with fair and affordable access to new audiences and helping them fulfill their dreams of fame and fortune was found in Amazon's launch of Kindle Translate, "an AI-powered translation service for authors to reach global readers" creating opportunities for authors "to reach new audiences and earn more"[2].

Despite this alluring promise, literary authors and translators alike remain skeptical about the suitability of machine translation (MT) and post-editing (PE) for literary texts (Moorkens et al., 2018; Way et al., 2023). While research shows that the quality of MT output has improved considerably since the introduction of neural techniques (Bentivogli et al., 2016; Castilho et al., 2017), it has mostly been a matter of marketing hype to state that neural MT (NMT) has "bridged the gap" between human and machine translation (Wu et al., 2016) or achieved "human parity" (Hassan et al., 2018), as shown by, amongst others Läubli et al. (2018) and Toral and Way (2018). Even state-of-the-art NMT and LLM translation produces output that contains errors and cannot be published without human revision. As pointed out by Way et al. (2023, p. 89), "claims that human translators are doomed as their jobs will be taken over completely by machines" and "[o]verhyping the capability of the technology does nobody any good". At the same time, the authors argue that the time has come to challenge the assumption that Computer-Assisted Translation (CAT), including MT, is "too crude and wayward to be of service" (*ibid*) in the translation of literary texts.

[1] https://www.nuanxed.com/
[2] https://www.aboutamazon.com/news/books-and-authors/amazon-kindle-translate-books-authors

The current study advances the debate surrounding literary MT and post-editing by focusing specifically on how post-editors react and respond to the way metaphors have been translated by NMT and LLMs in three short excerpts from a spy novel. We focus on metaphors because they are a well-known problem in translation and previous research indicates that MT has a tendency to translate metaphors too literally (Zajdel, 2022; Dorst, 2023; Karakanta et al., 2025), "leading to nonsensical, word-for-word translations" (Zajdel, 2022, p. 134). Metaphors are thus a persistent problem in MT that requires human post-editing. Our investigation aims to shed more light on which metaphors are perceived as problematic and how post-editors handled the way machines translate them.

2 Related Work

2.1 LitMT

Although literary translation was for a long time considered "the last bastion of human translation" and a challenge to "the perceived wisdom [...] that MT is of no use for the translation of literature" (Toral and Way, 2015b, p. 123), the examples in the Introduction show that the publishing industry is ready to embark full speed ahead. And while researchers warn against overstating MT's capabilities (e.g. Läubli et al., 2018; Toral et al., 2018; Way et al., 2023), both in industry and academia, a growing number of studies has explored the potential of statistical and neural systems in handling different literary genres across different language pairs (e.g. Besacier, 2014; Genzel et al., 2010; Greene et al., 2010; Jones and Irvine, 2013; Ploeger et al., 2024; Thai et al., 2022; Toral et al., 2023; Toral and Way, 2015a,b, 2018; Voigt and Jurafsky, 2012).

Studies assessing the potential of MT for particular genres and domains typically measure the quality of the MT output using automatic metrics, such as BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and COMET (Rei et al., 2020). As pointed out by Do Carmo (2022), such measurements often start from a misleading conceptualization of the notion of "quality"; moreover, the ensuing misplaced confidence that these metrics enable them to reliably measure quality prevents scholars from reflecting critically on their usefulness. In a similar vein, Van Egdom et al. (2023) argue that "metrics and algorithms cover only parts of the notion of "quality", and that a more fine-grained approach is needed if potential literary quality of machine translation is to be captured" (p. 129).

At present, most studies assessing the quality of literary MT output employ a combination of automatic metrics and human evaluation; in those human evaluations, usually carried out by native speakers, readers have been reported to rate a considerably high number of MT sentences as acceptable, error-free or equivalent to human translation: 60% for Spanish to Catalan (Toral and Way, 2018); 34% for English to Catalan (Toral et al., 2018); ˜20% for English into Russian and German (Matusov, 2019); and 44% for English into Dutch (Fonteyne et al., 2020). However, as Van Egdom et al. (2023) point out: "evaluative judgments are often passed by individuals who lack the required expertise to judge the quality [...] Evaluation is often performed by people that were simply available and willing to help out" (p. 131). Rather tellingly, a recent multilingual study involving 20 language pairs reported that professional translators preferred human translations to MT 85% of the time (Thai et al., 2022).

While recent research convincingly demonstrates that the quality of LitMT output can be significantly improved by techniques including domain adaptation (Toral et al., 2023), author-tailored adaptation (Kuzman et al., 2019; Oliver, 2023) and restoration of lexical richness (Ploeger et al., 2024), there is still a sizeable gap between the output and publishable translations, especially in terms of adequacy, style, tone, cohesion and the translation of figurative language (Tezcan et al., 2019; Matusov, 2019; Hansen and Esperanca-Rodier, 2022). Of the different forms of figurative language, linguistic metaphors - especially idioms and multi-word expressions - continue to be a problematic characteristic of literary texts and notoriously hard to translate for machines (Dorst, 2023; Karakanta et al., 2025; Zajdel, 2022). Even state-of-the-art NMT and LLMs have a tendency to translate metaphors literally (i.e. word-by-word), resulting in unidiomatic, illogical, even non-sensical translations (Dorst, 2023; Karakanta et al., 2025; Zajdel, 2022; Tezcan et al., 2019). In the current study, we therefore focus on whether literary post-editors notice such overly literal metaphor translations and how they react and respond to them.

2.2 Post-editing literary texts

Surveys on literary translators' perceptions and use of technology – ranging from simple online dictionaries and word processing to CAT tools and MT – such as those conducted by Ruffo (2018, 2022, 2024) and Daems (2021, 2022a,b), consistently indicate that while most literary translators embrace technologies such as the Internet, glossaries and terminology tools, they are often reluctant to use CAT tools and MT for multiple reasons: they have often not received any training in how to use such tools, they are unaware of the latest developments, they consider such tools unsuitable for creative texts, and they are opposed to "tools that threaten to steal the essence of their translation activity, ignoring the peculiarly human aspects of it" (Ruffo, 2018, p. 130).

Interestingly, the study by Moorkens et al. (2018) on literary post-editing versus translation 'from scratch' found that while all participants "were faster when post-editing NMT, [...] they all still stated a preference for translation from scratch, as they felt less constrained and could be more creative" (p. 256). The participants in the study "complained that the MT systems 'conditioned' them to produce a literal translation" (*ibid*).

This sense of being constrained by the MT output and being lured into accepting MT output was also found in a series of experiments conducted by Guerberof-Arenas and colleagues (Guerberof-Arenas and Toral, 2020, 2022, 2024), focusing specifically on creativity, as well as the real-life case studies conducted by Kenny and Winters (2020), Winters and Kenny (2023) and Kolb (2023a,b). These findings are not only relevant from a stylistic perspective, but also raise legal and ethical questions. Taivalkoski-Shilov and Koponen (2023), for example, write about the ongoing debate regarding copyright, authorship and intellectual property in translation. They note how post-editing poses a serious challenge to traditional copyright laws as authorship will depend on how much of the raw MT output is kept and how much "originality" is added through the post-editor's unique changes.

In the study by Guerberof-Arenas and Toral (2020) human translation scored higher on creativity than MT and PE and ranked higher in narrative engagement and translation reception. Guerberof-Arenas and Toral therefore argue that "professional translators, by providing solutions that are both novel and acceptable, add the creativity factor that MT is lacking at present" (p. 277). Moreover, Guerberof-Arenas and Toral (2022) found not only that their literary-adapted neural MT system did not have the "necessary capabilities for a creative translation" but also that "using MT to post-edit raw output constrains the creativity of translators, resulting in a poorer translation often not fit for publication, according to experts" (p. 184). In the same study, the reviewers were unanimous in their verdict that the human translations were good, the MTs were bad, and the post-edited versions were neither good nor bad.

In the current study we investigate how literary post-editors react and respond to the way metaphors have been translated by MT systems, both NMT and LLM, while post-editing short excerpts from a novel. We are particularly interested in how often metaphor translations were changed by the post-editors and how aware they were of overly literal translations that require more creative solutions.

3 Methodology

3.1 Machine Translations

Three systems were used to generate machine translations for the literary post-editing task: two commercial NMT engines – Google Translate and DeepL – and one general-purpose LLM, ChatGPT (GPT4; accessed via the ChatGPT user interface). All systems were accessed via their online web interfaces [3], reflecting the way most literary translators use MT, if at all. Most professional literary translators do not build or customize their own MT systems, and when they do post-editing assignments, they often receive the MTs as editable Word files (Ruffo, 2024; Macken et al., 2025). ChatGPT received a bare prompt ("Translate the following English text into Dutch for the Netherlands"). No in-domain training, customization or finetuning took place; all translations were generated "out of the box". These three systems were chosen because they are widely used, freely accessible online, and representative of the engines most translators are familiar with. Three different versions of the PE task were created so each engine was used for one fragment and the post-editors saw one output from each of the three engines.

[3] https://translate.google.com/; https://www.deepl.com/nl/translator; https://chatgpt.com/

3.2 Source Text: The Lucy Ghosts (Eddy Shah)

The novel used as the source text (ST) in this study was selected from the Fiction subcorpus of the annotated VU Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010b). The annotated VUAMC is available for downloading for non-commercial use via the Oxford Text Archives website [4]. All texts in the VUAMC were sampled from the BNC Baby corpus [5] with permission under NWO Vici grant 277-30-001 [6]. All sampled excerpts were manually annotated for linguistic metaphor by a team of experts using the Metaphor Identification Procedure VU or MIPVU (Steen et al., 2010). The VUAMC has been widely used as the gold standard in NLP studies on metaphor processing. Three short excerpts were selected (798 words in total), offering a range of different types of metaphor (i.e. single vs multiword metaphors; different word classes; different degrees of creativity). Since the post-editors were revising three short excerpts instead of the whole text, a synopsis and some background information were provided for contextualization along with the ST excerpts.

3.3 Post-editors

In total, six post-editors (PEs) were invited to participate in the task and subsequent interview. All six had recently graduated from the Master's in Translation at Leiden University in the Netherlands and completed the specialization in Literary Translation as well as a course in Translation Technology. This ensured a relatively homogeneous background and a reasonable familiarity with translation technology, including MT. We invited graduates who we knew received high grades for their literary translations and who are currently working as professional translators, both in literary and non-literary translation, with no more than 5 years of professional experience. The participants were told they would complete a literary post-editing task, and would be invited to share their thoughts on the suitability of MT for literary texts, but they were not informed of the authors' particular interest in metaphors. The participants signed an Informed Consent Form before the task and were debriefed during the interview. They received financial compensation for their participation.

3.4 Data Collection - PE task and Interview

Each of the PEs came to Leiden to complete the task and interview individually. They were encouraged to bring their own laptop, which three of the six participants did. The other three worked on a laptop provided by the authors. The PEs received two Word documents: one with the three ST fragments (F1, F2, F3) and some information on the novel, the plot and the setting of the excerpts; and one with three MT outputs. There were three versions: PE1 and PE2 received version 1, with F1=DeepL, F2=ChatGPT, F3=GT; PE3 and PE4 received version 2, with F1=GT, F2=DeepL, F3=ChatGPT; and PE5 and PE6 received version 3, with F1=ChatGPT, F2=GT, F3=DeepL). The document also included instructions for the PE task: participants were instructed to work with Tracked Changes and Comments and use any resources they required. A subset of the data and materials associated with this study is publicly available in a GitHub repository [7].

During the post-task interview, two of the three authors were present. The interviews were semi-structured and started from a predefined list of questions. Our interest in metaphors was brought up towards the end of the interview, so we could see whether problems with specific metaphors or figurative language more generally were raised by the post-editors themselves. The interviews were recorded using a smartphone and subsequently transcribed by one of the authors. All interviews were conducted in Dutch; quotations from the PEs were translated into English by the authors for the sake of readability.

Transcribed interviews were uploaded into ATLAS.ti to allow for a more systematic and quantitative analysis of the post-editors' attitudes toward literary PE and the suitability of MT for literary translation. The coding process followed a bottom-up, inductive thematic analysis in line with the approach outlined by Braun and Clarke (2006). This iterative process allowed the analysis to evolve from general themes towards more concrete conceptual groupings (Skjøtt Linneberg and Korsgaard, 2019, p. 12-13). To prevent the unconscious falsification of data by coding it according

[4] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2541?show=full

[5] https://www.natcorp.ox.ac.uk/corpus/babyinfo.html

[6] https://www.nwo.nl/projecten/277-30-001-0

[7] https://github.com/dorstag/

to theoretically desirable patterns (Schuyt, 2019, p. 127), we annotated and reviewed the dataset both separately as well as together and did so multiple times.

Reflections often included statements that fit multiple categories, and each relevant statement was coded separately. These characteristics should be understood as themes, meaning the post-editors did not have to use the exact wording of the category labels (i.e., if a post-editor said "*aftrap* refers to football, we don't say that in Dutch", a code was added that the PE noticed the metaphor and that they thought it was a bad translation).

4 Results

This section presents the results of the post-editing task and follow-up interviews with the PEs (hereafter referred to as PE1 to PE6). Our aim was to determine how the PEs reacted to MT metaphors - that is, whether they retained or changed the output - and how they responded to them, focusing on whether their reported perceptions correspond to their actual post-editing behavior.

Engine	F1	F2	F3	Total (mean)
Google Translate	181	184	158	523 (174)
DeepL	170	214	169	553 (184)
ChatGPT	130	205	190	525 (175)
Total (mean)	481 (160)	603 (201)	517 (172)	1601

Table 1: Number of edits per MT system and fragment.

To determine whether the MT system influenced editing behavior, the number of edits was analysed per system (Table 1).

The differences between the three systems are relatively small: DeepL required the highest total number of edits (553), followed by ChatGPT (525) and Google Translate (523). This suggests that there were no clear quality differences between the engines used in the task. How many edits were made seemed to depend more on the ST characteristics (fragment variation) and post-editing style (PE variation) than on which engine was used to generate the output. The strongest PE variation was observed between PE1, who had 389 edits, and PE2, who made 144 edits. The other four fell between these two extremes (see Table 2).

The interviews indicate that these differences do indeed reflect individual post-editing styles, as well as varying levels of experience with professional post-editing, views on the suitability of MT and PE for literature, and personal quality thresholds. Participants reported varying degrees of experience with both MT and PE. While all participants were familiar with MT, either from their studies or their professional work, only half reported having professional post-editing experience. These varying levels of familiarity with MT and (literary) PE may in part explain the observed variation in PE behavior.

4.1 Reactions to metaphor in MT: Changes made during the post-editing task

This section focuses on the post-editors' reactions to metaphor by determining whether they reacted by retaining or changing the MT output.

Table 3 gives the total number of changes made to the metaphor translations per fragment for each of the six PEs, regardless of engine.

For all PEs, more than a third of the machine-translated metaphors were problematic, in all three fragments. This indicates that metaphor translation is not a "solved problem" yet, with at least one out of every three metaphors requiring a revision. Considering that corpus-based metaphor studies by Steen and colleagues (2010; 2010b; 2010; 2010a) have shown that, on average, one in eight words is metaphorical in authentic discourse (including fiction, journalism and academic texts), this entails that post-editors will need to pay careful attention to metaphor translation when working with machine-translated texts.

While there is some variation between the PEs, all of them changed more than 30% of the metaphors in the fragments (so at least 21 of the 69 metaphors). PE1 has, again, consistently made considerably more changes than all other PEs, changing more than half of the metaphor translations in each fragment. Comparing these findings for changes to metaphor to the previously established overall patterns, we can now see that F2 was problematic for other reasons besides the presence of (difficult) metaphors.

While all PEs made considerably more revisions in F2, focusing only on how often the metaphors were changed shows more variation across fragments and post-editors: F3's metaphor translations were considered most problematic by PE1 (65% changed), PE2 (45% changed), PE4 (50% changed) and PE6 (40% changed) while PE3 and PE5 made more changes to the metaphors in F2 (48% each). However, PE4 and PE6 made the fewest changes to F2, and PE3 made the fewest

Category	PE1	PE2	PE3	PE4	PE5	PE6	Total (mean)
Google Translate	122	36	99	82	90	94	523 (87)
DeepL	120	50	122	92	89	80	553 (92)
ChatGPT	147	58	99	91	54	76	525 (88)
F1	120	50	99	82	54	76	481 (80)
F2	147	58	122	92	90	94	603 (101)
F3	122	36	99	91	89	80	517 (86)
Total (mean)	389 (130)	144 (48)	320 (107)	265 (88)	233 (78)	250 (83)	1601

Table 2: Overview of edits across systems, fragments, and post-editors, including totals per category.

Fragment	PE1	PE2	PE3	PE4	PE5	PE6
F1 (20)	11 (55%)	6 (30%)	9 (45%)	7 (35%)	8 (40%)	7 (35%)
F2 (29)	17 (59%)	10 (34%)	14 (48%)	9 (31%)	14 (48%)	10 (34%)
F3 (20)	13 (65%)	9 (45%)	6 (30%)	10 (50%)	9 (45%)	8 (40%)
Total (69)	41 (59%)	25 (36%)	29 (42%)	26 (38%)	31 (45%)	25 (36%)

Table 3: Number of changes to machine-translated metaphors per fragment and post-editor.

changes to F3. This suggests that not all post-editors find the same metaphor translations problematic.

Interestingly, the few instances in which all six PEs changed the metaphor translation were cases of multiword metaphors where all three systems had produced overly literal translations, as illustrated in Table 4 (see Appendix B for literal word-by-word translations into English). However, the revisions show such considerable variation with respect to *how* the PEs solved the problem that a detailed analysis of their solutions and whether the revision was required or optional is unfortunately beyond the scope of this paper.

If we split the data by engine, regardless of fragment, the results per post-editor are as presented in Table 5. This shows that there is not one engine that produces more problematic metaphor translations overall, as far as the number of revisions made by the post-editors is concerned. PE5 and PE6 made the most changes to the metaphor translations in the Google Translate output, PE3 in the DeepL output, and PE1, PE2 and PE4 in the ChatGPT output. One issue for future investigation is whether the engines did differ systematically in what kind of errors they produced for these metaphor translations. Karakanta et al. (2025) showed that NMT made more form errors (fluency) while LLMs made more meaning errors (accuracy). Similarly, the study by Tezcan et al. (2019) suggests that metaphors like 'sport' to refer to a person resulted in errors that were classified both as lexical as well as logical mistakes.

Further investigation is also needed to tell us whether the changes made by the PEs were required or optional, and whether the MT metaphor

Sentence	Source	Translation
'It's important you understand the background,' said Sorge, **not allowing the American to get under his skin.**	ChatGPT	Zonder de Amerikaan onder zijn huid te laten kruipen
	PE1	Die de Amerikaan geen kans gaf om vat op hem te krijgen
	PE2	Die de Amerikaan de kans niet gaf hem op te jagen
	DeepL	Niet toestaand dat de Amerikaan onder zijn huid
	PE3	Die de Amerikaan hem niet op stang liet jagen
	PE4	Die zich niet liet opjagen door de Amerikaan
	GT	Terwijl hij de Amerikaan niet onder zijn huid liet kruipen
	PE5	Ongevoelig voor het gezeur van de Amerikaan
	PE6	Terwijl hij de Amerikaan niet de kans gaf om hem op de zenuwen te werken

Table 4: All versions of the idiom "get under his skin"

System	PE1	PE2	PE3	PE4	PE5	PE6
GoogleTranslate	13 (32%)	9 (36%)	9 (31%)	7 (27%)	14 (45%)	10 (40%)
DeepL	11 (27%)	6 (24%)	14 (48%)	9 (35%)	9 (29%)	8 (32%)
ChatGPT	17 (41%)	10 (40%)	6 (21%)	10 (38%)	8 (26%)	7 (28%)
Total	41	25	29	26	31	25

Table 5: Number of changes to machine-translated metaphors per engine by post-editor.

translations were in fact "correct" or "incorrect" as classified by an error analysis such as the one carried out by Tezcan et al. (2019). Going through the data to catalogue changes and retentions, we noticed instances where the post-editors have retained metaphor translations that are technically "incorrect" or changed "correct" metaphor translations. For example, PE3 retained DeepL's incorrect translation *'breken van cyphers*' for 'breaking cyphers' (lexical error for 'break' and non-existent word for 'cyphers'). Conversely, PE1 changed ChatGPT's correct translation *'aanzienlijk'* ('considerable'/'substantial'/'notable') for 'substantial'

into *'ongelooflijk succesvol'* ('incredibly successful').

Met Type	PE1	PE2	PE3	PE4	PE5	PE6
Single (41)	21 (51%)	10 (24%)	12 (29%)	11 (27%)	15 (37%)	10 (24%)
Multiword (28)	20 (71%)	15 (54%)	17 (61%)	15 (54%)	16 (57%)	15 (54%)
Total (69)	41 (59%)	25 (36%)	29 (42%)	26 (38%)	31 (45%)	25 (36%)

Table 6: Number of changes to machine-translated metaphors per metaphor type by post-editor.

Of the 69 metaphors in the three fragments, 41 were single metaphors (e.g. *'snapped', 'bypass', 'sources'*) and 28 were multiword metaphors (e.g. *'take off', 'broke their word', 'your hands were not that clean'*). Table 6 shows that while there is considerable variation between the post-editors in how many of the machine-translated metaphors they changed overall, ranging from a third to over half of all the metaphors, they consistently changed more translations of multiword metaphors than single metaphors. Four of the PEs changed about one in four of the single metaphors, PE5 a third, and only PE1 more than half. For the multiword metaphors, all six PEs changed more than half of the MT translations, and PE1 even changed 20 out of 28, more than 70%.

This shows that while multiword metaphors may be less frequent in naturally occurring discourse (and perhaps more strongly related to authorial style, i.e. a preference for idioms or phrasal verbs), they are much more likely to be problematic in MT. Interestingly, all of these multiword metaphors are highly conventional, so their mistranslation by MT cannot be due to obscure or infrequent use. The context shows they are also used in their canonical form. However, taken literally, they evoke rather vivid imagery that is relevant to the story unfolding. One of our next steps is therefore also to determine how readers react and respond to these literal translations and whether they consider them nonsensical or creative.

4.2 General patterns in the interview data

As discussed in the previous section, the task revealed differences in post-editing styles. The interview data provides additional insights into these variations. Some participants reflected on their usual PE workflow and reported that they normally adopt a pragmatic approach, preferring to leave acceptable MT output unchanged in order to save time, as they are not compensated sufficiently to deliver perfect quality. PE3 explained this approach as follows:

> *Researcher A: Do I understand correctly that, usually, you save time on [post-editing] texts?*
>
> *PE3: Yes, but then I also think that it's not my best work.*
>
> *Researcher B: Yeah.*
>
> *PE3: Right? So, as far as I'm concerned, they get a passable translation– one that's just okay, and that, I think, most people wouldn't consider bad, because I am actually a pretty decent post-editor. But I do believe that it could use an extra round of translating/revising, and that's what I would prefer to do.*

Similarly, PE6 emphasised that compensation influenced the effort they were willing to invest, focusing on essential corrections only:

> *Researcher A:[...], unless you deliver a lower-quality output.*
>
> *PE6: Yeah, that's definitely what I would do. If I'd get paid less for it [post-editing], I really would... only focus on the vocabulary. And then I'd be like, yeah, figure it out yourself. Yes, ask someone else to do the full editing.*

These quotes are clearly aligned with the lack of authorship and text ownership in post-editing discussed by Taivalkoski-Shilov and Koponen (2023). Similarly, the quantitative results presented in Table 7 align with previous post-editing studies such as those by Moorkens et al. (2018) and Ruffo (2024), showing that the participants frequently encountered inefficiencies and challenges during the post-editing task. The 'Efficiency' categories suggest that, although post-editing was occasionally faster than translating from scratch (e.g. PE3 had a tag in 'Faster'), a substantial portion of the task required more work or was classified as being significantly slower than translating from scratch, particularly for PE4, PE5 and PE6.

Similarly, the 'Experience' categories reveal that the post-editing task often resulted in difficulties such as self-doubt regarding language proficiency, frustration, and reduced creativity, among other issues. The issues of doubting their own language skills and feeling constrained in their creativity were often linked in the interviews: the participants described how they could not "unsee"

Category	PE1	PE2	PE3	PE4	PE5	PE6	Total
Efficiency: Faster	0	0	1	0	0	0	1
Efficiency: More work	4	2	3	4	3	1	17
Efficiency: Slower	4	0	0	2	7	2	15
Experience: Difficult	1	2	1	3	0	0	7
Experience: Doubt language skills	3	4	2	6	7	0	22
Experience: Frustration	0	0	0	3	1	0	4
Experience: Fun	1	1	1	0	1	1	5
Experience: Lack context	2	1	1	1	3	0	8
Experience: Limits creativity	2	1	1	7	2	1	14
Total	17	11	10	26	24	5	93

Table 7: PE Efficiency and Experience

Metaphor-related code	PE1	PE2	PE3	PE4	PE5	PE6	Total
Aware after probing	0	1	0	0	1	2	4
Mentioned problematic metaphor	16	5	1	8	0	2	32
Mentioned good metaphor	1	1	0	0	0	1	3
Not aware after probing	0	1	1	0	0	1	3
Total	17	8	2	8	1	6	42

Table 8: Metaphor-related quotes

Category	Total quality mentions	Co-occurring with metaphor
Bad overall / other	18	3
Grammar	2	0
Too literal	22	8
Unnatural language	19	0
Vocabulary	7	0

Table 9: Total mentions of negative quality fragments and their co-occurrence with noticing problematic metaphors

the MT solutions; after having seen the MT solution they could no longer think of alternatives. While they had this nagging suspicion that something was wrong with the translation, they struggled to think how they would have said it themselves and whether it was correct Dutch or not. This confirms previous studies on how PE limits translators' creativity (e.g. Guerberof-Arenas and Toral, 2020, 2022).

4.3 Responses to metaphor in MT: Reflections on metaphor during the interview

The interview data provide further insights into whether the participants noticed the metaphors during the PE task, as well as into how the MT translations were perceived. As shown in Table 8, 42 out of in total 280 coded segments across nine themes were metaphor-related (see Appendix A for an overview of all coded themes across participants). Of these, the majority (35) consisted of unprompted mentions, indicating that metaphors were frequently noticed spontaneously by the participants. However, there is considerable variation between the PEs: while PE1 mentioned metaphors 17 times, PE5 mentioned them only once, and only after probing by the researchers. This suggests that metaphor awareness differs across individuals.

As discussed in Section 4.2, PE1 made the highest number of changes to metaphor translations, whereas other PEs, such as PE3 and PE6, changed considerably less. The interview data thus suggest a link between noticing metaphors in the task and mentioning them during the interview: PEs who were more attuned to the metaphorical language were also more likely to revise it.

Several PEs reported issues with unnatural and overly literal language in the texts, as well as errors introduced by the machines (see Table 9). Notably, the researchers purposefully had not yet mentioned metaphors during the part of the interviews when the questions "what did you think of the quality of the machine translations" and "did you notice certain stylistic features" were asked, as illustrated in the quote below.

> *Researcher: Did you notice any particular stylistic features?*
>
> *PE3: Let me think- short sentences, repetition, yes, those are things I took into account, of which I thought: this is ugly, but I'll leave it as it is.*
>
> *Researcher: And why did you find it ugly?*
>
> *PE3: Well, because... in some cases it just sounded too strange in Dutch; that you'd think, this is an American expressing things in a certain way and I get that, but it doesn't work like that. So I merged one sentence because of that, and otherwise tried to adapt it... There were simply sentences where I thought: Yes, but this is just not how we would say it.*

However, references to metaphors frequently co-occurred with comments about overall poor quality and the literalness of the MT output. In eight of the 32 instances in which metaphors were mentioned, the PEs also mentioned that the metaphor was translated too literally. This pattern suggests that by mentioning metaphors in conjunction with excessive literalness or broader issues (as reflected by the dual tagging), the PEs consciously linked the MT output as being too literal and not good enough in terms of metaphor translation. PE4, for example, explicitly connected incorrect metaphor translations to literalness:

> *Researcher A: And the overly literal things? Could you give an [example]?*
>
> *PE4: Yes, so 'breaking ciphers' was translated as 'ciphers breken'.*

Here too, the PE connects this problem to the struggle to think of alternatives:

> *PE4: And I also often had the problem that there was something idiomatic in the source text, which, as far as I could tell, had not been correctly translated in... by the machine, but I, because I had already been primed, I found it difficult to think of a solution that would be correct in Dutch.*

The same problem is raised by PE4:

> *Researcher: What we were secretly looking for in these texts was what you did with the literally translated metaphors. So, for example, 'under your skin'. Did those stand out or not necessarily more than other things?*
>
> *PE4: They definitely stood out, but because I got stuck in that track, it was often hard to find what the correct solution actually was.*

5 Conclusion

This study set out to examine how post-editors react and respond to metaphor in literary MT. The findings from the post-editing task and subsequent interviews demonstrate that metaphor translation continues to pose a challenge for both NMT and LLM-based systems. While the general editing patterns demonstrated greater variation between both fragments and PEs, all six participants changed more than a third of the MT metaphors across engines and fragments, indicating that one in three metaphors in MT may be considered problematic.

One key finding is that this is particularly true for multiword metaphors, which were changed over half of the time by all PEs across fragments and engines. The observed lack of differences between the three engines suggests that there is at present no indication that LLM-based systems are better at translating metaphors than NMT, at least for literary texts. All three systems struggled with idiomatic expressions, phrasal verbs, and collocations, even when they were highly conventional, frequent, and occurring in their standard form.

The results from the interviews align with previous studies on literary MT and PE, confirming that post-editing literary MT output is generally perceived as effortful, and as less efficient than translating from scratch. The participants reported frustration, reduced creativity, and a feeling of doubting their own language skills, supporting previous claims in the literature. As far as their responses to metaphor were concerned, the data suggest that when PEs were more aware of the metaphors during the task, they were also more likely to change them and mention them during the interview (i.e. they also remembered them). References to metaphors co-occurred with references to overall poor quality as well as to the output being too literal. Interestingly, a number of PEs indicated that they felt their mind was being "tricked" by the output into accepting translations that were not in fact idiomatic in Dutch, but once they "got stuck in the MT's track", they could no longer think of the right way to say it.

The current paper is of course limited in generalisability, analysing data from only six post-editors for three short excerpts, for one language pair, and comparing only three commercial, black-box systems. Moreover, further research is needed to determine how many of the changes that were made by the PEs were optional rather than required, and how often they retained metaphor translations that could be classified as errors. This will provide more insights into when and how often PEs are "lured into" accepting incorrect MT output. As to why they retain MT metaphors, more research still needs to be done on whether incorrect metaphor translations were in fact not noticed by the post-editors or whether they left the output unchanged

because they could not think of an alternative, because they did not care enough to change it given the poor overall quality and poor remuneration, or because they suspected the reader might actually like it and find it creative.

6 Acknowledgments

This publication is part of the project Metaphor in Machine Translation: Reactions, Responses, Repercussions with file number VI.Vidi.231C.014 of the research program Vidi SGW which is (partly) financed by the Dutch Research Council (NWO) under the grant https://doi.org/10.61686/ASYVT63546.

The data for this publication were collected with financial support from the European Association for Machine Translation (EAMT) through its 2024 sponsorship of activities programme, proposal entitled "Metaphor in Literary Machine Translation and Post-Editing". We would like to thank the post-editors for their participation.

References

Bentivogli, L., A. Bisazza, M. Cettolo, and M. Federico (2016). Neural versus Phrase-Based Machine Translation Quality: A Case Study. In J. Su, K. Duh, and X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 257–267. https://aclanthology.org/D16-1025/.

Besacier, L. (2014). Traduction Automatisée d'une Oeuvre Littéraire: Une Étude Pilote. *Traitement Automatique du Langage Naturel (TALN)*. TALN Conference, Marseille.

Braun, V. and V. Clarke (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa.

Castilho, S., J. Moorkens, F. Gaspari, R. Sennrich, V. Sosoni, P. Georgakopoulou, P. Lohar, A. Way, A. V. Miceli-Barone, and M. Gialama (2017). A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators. In S. Kurohashi and P. Fung (Eds.), *Proceedings of Machine Translation Summit XVI: Research Track: September 18-22, Nagoya*, pp. 116–131. Nagoya, Japan. https://aclanthology.org/2017.mtsummit-papers.10/.

Daems, J. (2021). Wat denken literaire vertalers echt over technologie? WEBFILTER. https://www.tijdschrift-filter.nl/webfilter/dossier/literair-vertalen-en-technologie/2021-1/wat-denken-literaire-vertalers-echt-over-technologie/.

Daems, J. (2022a). Dutch Literary Translators' Use and Perceived Usefulness of Technology. In J.L. Hadley, K. Taivalkoski-Shilov, C.S.C. Teixeira, and A. Toral (Eds.), *Using Technologies for Creative-Text Translation*, pp. 40–65. Routledge. https://doi.org/10.4324/9781003094159.

Daems, J. (2022b). Literaire automatische vertaling? Enkele inzichten uit het onderzoeksveld. ELV-dossier Vertaling en technologie. https://literairvertalen.org/kennisbank/literaire-automatische-vertaling-enkele-inzichten-uit-het-onderzoeksveld.

Do Carmo, F. (2022). Debunking a Few Machine Translation Myths: From Zero-Shot Translation to Human Parity and No Language Left Behind. University of Surrey, Convergence Lecture Series. https://www.surrey.ac.uk/news/convergence-lecture-series-debunking-few-machine-translation-myths-zero-shot-translation-human.

Dorst, A. G. (2023). Metaphor in Literary Machine Translation: Style, Creativity and Literariness. In A. Rothwell, A. Way, and R. Youdale (Eds.), *Computer-Assisted Literary Translation*, pp. 173–186. Routledge. doi:10.4324/9781003357391-12.

Fonteyne, M., A. Tezcan, and L. Macken (2020). Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 3790–3798. https://aclanthology.org/2020.lrec-1.468/.

Genzel, D., J. Uszkoreit, and F. Och (2010). "Poetic" Statistical Machine Translation: Rhyme and Meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, USA, pp. 158–166. https://aclanthology.org/D10-1016/.

Greene, E., T. Bodrumlu, and K. Knight (2010). Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, USA, pp. 524–533. https://aclanthology.org/D10-1051/.

Guerberof-Arenas, A. and A. Toral (2020). The Impact of Post-Editing and Machine Translation on Creativity and Reading Experience. *Translation Spaces 9*(2), 255–282. https://doi.org/10.1075/ts.20035.gue.

Guerberof-Arenas, A. and A. Toral (2022). Creativity in Translation: Machine Translation as a Constraint for Literary Texts. *Translation Spaces 11*(2), 184–212. https://doi.org/10.1075/ts.21025.gue.

Guerberof-Arenas, A. and A. Toral (2024). To Be or Not to Be: A Translation Reception Study of a Literary Text Translated into Dutch and Catalan Using Machine Translation. *Target 36*(2), 215–244. https://doi.org/10.1075/target.22134.gue.

Hansen, D. and E. Esperanca-Rodier (2022). Human-Adapted MT for Literary Texts: Reality or Fantasy? In *Proceedings of the New Trends in Translation and Technology Conference*, Rhodes, Greece, pp. 178–190. https://hal.science/hal-04038025.

Hassan, H., A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou (2018). Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint*. arXiv:1803.05567.

Jones, R. and A. Irvine (2013). The (Un)Faithful Machine Translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Sofia, Bulgaria, pp. 96–101. https://aclanthology.org/W13-2713/.

Karakanta, A., M. O. Nas, and A. G. Dorst (2025). Metaphors in Literary Machine Translation: Close but No Cigar? In P. Bouillon, J. Gerlach, and S. Girletti (Eds.), *Proceedings of Machine Translation Summit XX*. European Association for Machine Translation. https://aclanthology.org/2025.mtsummit-1.21/.

Kenny, D. and M. Winters (2020). Machine Translation, Ethics and the Literary Translator's Voice. *Translation Spaces 9*(1), 123–149. https://doi.org/10.1075/ts.00024.ken.

Kolb, W. (2023a). Brazilian short prose in German: A study of literary post-editing. *Revista Tradumàtica. Tecnologies de la Traducció 21*, 63–70. https://doi.org/10.5565/rev/tradumatica.347.

Kolb, W. (2023b). 'I Am a Bit Surprised': Literary Translation and Post-Editing Processes Compared. In A. Rothwell, A. Way, and R. Youdale (Eds.), *Computer-Assisted Literary Translation*, pp. 53–68. Routledge.

Kuzman, T., S. Vintar, and M. Arčan (2019). Neural Machine Translation of Literary Texts from English to Slovene. In J.L. Hadley, M. Popović, H. Afli, and A. Way (Eds.), *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, pp. 1–9. European Association for Machine Translation. https://aclanthology.org/W19-7301/.

Läubli, S., R. Sennrich, and M. Volk (2018). Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, pp. 4791–4796. Association for Computational Linguistics. https://aclanthology.org/D18-1512/.

Lavie, A. and A. Agarwal (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, pp. 228–231. Association for Computational Linguistics. https://aclanthology.org/W07-0734/.

Macken, L., P. Ruffo, and J. Daems (2025). The Role of Translation Workflows in Overcoming Translation Difficulties: A Comparative Analysis of Human and Machine Translation (Post-Editing) Approaches. In *Proceedings of the Second Workshop on Creative-Text Translation and Technology (CTT)*, Geneva, pp. 1–13. European Association for Machine Translation. https://aclanthology.org/2025.ctt-1.1/.

Matusov, E. (2019). The Challenges of Using Neural Machine Translation for Literature. In J.L. Hadley, M. Popović, H. Afli, and A. Way (Eds.), *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, pp. 10–19. European Association for Machine Translation. https://aclanthology.org/W19-7302/.

Moorkens, J., A. Toral, S. Castilho, and A. Way (2018). Translators' Perceptions of Literary Post-Editing Using Statistical and Neural Machine Translation. *Translation Spaces 7*(2), 240–262. https://doi.org/10.1075/ts.18014.moo.

Oliver, A. (2023). Author-Tailored Neural Machine Translation Systems for Literary Works. In A. Rothwell, A. Way, and R. Youdale (Eds.), *Computer-Assisted Literary Translation*, pp. 126–141. Routledge.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, and D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 311–318. https://aclanthology.org/P02-1040/.

Ploeger, E., H. Lai, R. Van Noord, and A. Toral (2024). Towards Tailored Recovery of Lexical Diversity in Literary Machine Translation. In C. Scarton, C. Prescott, C. Bayliss, C. Oakley, J. Wright, S. Wrigley, X. Song, E. Gow-Smith, R. Bawden, V. M. Sánchez-Cartagena, P. Cadwell, E. Lapshinova-Koltunski, V. Cabarrão, K. Chatzitheodorou, M. Nurminen, D. Kanojia, and H. Moniz (Eds.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, Sheffield, UK, pp. 286–299. European Association for Machine Translation. https://aclanthology.org/2024.eamt-1.24/.

Reedz (2023). Swedish AI Startup Reedz Launches Reedz Books. https://sl1nk.com/367tzqd.

Rei, R., C. Stewart, A. C. Farinha, and A. Lavie (2020). COMET: A Neural Framework for MT Evaluation. In B. Webber, T. Cohn, Y. He, and Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 2685–2702. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.213/.

Ruffo, P. (2018). Human-Computer Interaction in Translation: Literary Translators on Technology and Their Roles. In *Proceedings of the 40th Conference Translating and the Computer*, pp. 127–131. AsLing.

Ruffo, P. (2022). Collecting Literary Translators' Narratives: Towards a New Paradigm for Technological Innovation in Literary Translation. In J.L. Hadley, K. Taivalkoski-Shilov, C.S.C. Teixeira, and A. Toral (Eds.), *Using Technologies for Creative-Text Translation*, pp. 18–39. Routledge.

Ruffo, P. (2024). Literary translators in-between: An exploration of their self-imaging discourse and relationship to technology. *Translation in Society 3*(1), 87–103. https://doi.org/10.1075/tris.23015.ruf.

Schuyt, K. (2019). *Scientific Integrity: The Rules of Academic Research*. Leiden University Press. https://hdl.handle.net/1887/79152.

Skjøtt Linneberg, M. and S. Korsgaard (2019). Coding Qualitative Data: A Synthesis Guiding the Novice. *Qualitative Research Journal 19*(3), 259–270. https://doi.org/10.1108/QRJ-12-2018-0012.

Steen, G., E. Biernacka, A. G. Dorst, A. A. Kaal, C. I. López Rodríguez, and T. Pasma (2010). Pragglejaz in Practice: Finding Metaphorically Used Words in Natural Discourse. In G. Low, Z. Todd, A. Deignan, and L. Cameron (Eds.), *Researching and Applying Metaphor in the Real World*, pp. 165–184. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.26.11ste.

Steen, G., A. G. Dorst, J. B. Herrmann, A. A. Kaal, and T. Krennmayr (2010a). Metaphor in Usage. *Cognitive Linguistics 21*(4), 765–796. https://doi.org/10.1515/cogl.2010.024.

Steen, G., A. G. Dorst, J. B. Herrmann, A. A. Kaal, and T. Krennmayr (2010b). VU Amsterdam Metaphor Corpus. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2541.

Steen, G., A. G. Dorst, J. B. Herrmann, A. A. Kaal, T. Krennmayr, and T. Pasma (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam: John Benjamins.

Taivalkoski-Shilov, K. and M. Koponen (2023). Literary Post-Editing and the Question of Copy-

right. *HERMES - Journal of Language and Communication in Business* (63), 195–207. https://doi.org/10.7146/hjlcb.vi63.137012.

Tezcan, A., J. Daems, and L. Macken (2019). When a 'Sport' is a Person and Other Issues for NMT of Novels. In J.L. Hadley, M. Popović, H. Afli, and A. Way (Eds.), *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, pp. 40–49. European Association for Machine Translation. https://aclanthology.org/W19-7306/.

Thai, K., M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, J. Wieting, and M. Iyyer (2022). Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, pp. 9882–9902. Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.672/.

Toral, A., A. van Cranenburgh, and T. Nutters (2023). Literary-Adapted Machine Translation in a Well-Resourced Language Pair: Explorations with more Data and Wider Contexts. In A. Rothwell, A. Way, and R. Youdale (Eds.), *Computer-Assisted Literary Translation*, pp. 27–52. New York: Routledge.

Toral, A. and A. Way (2015a). Machine-Assisted Translation of Literary Text: A Case Study. *Translation Spaces 4*(2), 240–267. https://doi.org/10.1075/ts.4.2.04tor.

Toral, A. and A. Way (2015b). Translating Literary Text between Related Languages using SMT. In A. Feldman, A. Kazantseva, S. Szpakowicz, and C. Koolen (Eds.), *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, Colorado, USA, pp. 123–132. Association for Computational Linguistics. https://aclanthology.org/W15-0714/.

Toral, A. and A. Way (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text? In J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice*, pp. 263–287. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7$_1$2.

Toral, A., M. Wieling, and A. Way (2018). Post-Editing Effort of a Novel with Statistical and Neural Machine Translation. *Frontiers in Digital Humanities 5*(9), 1–11. https://doi.org/10.3389/fdigh.2018.00009.

Van Egdom, G.-W., O. Kosters, and C. Declercq (2023). The Riddle of (Literary) Machine Translation Quality: Assessing Automated Quality Evaluation Metrics in a Literary Context. *Revista Tradumàtica. Tecnologies de la Traducció 21*, 129–159. https://doi.org/10.5565/rev/tradumatica.345.

Voigt, R. and D. Jurafsky (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. In D. Elson, A. Kazantseva, R. Mihalcea, and S. Szpakowicz (Eds.), *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, Montréal, Canada, pp. 18–25. Association for Computational Linguistics. https://aclanthology.org/W12-2503/.

Way, A., A. Rothwell, and R. Youdale (2023). Why More Literary Translators Should Embrace Translation Technology. *Revista Tradumàtica. Tecnologies de la Traducció 21*, 87–102. https://doi.org/10.5565/rev/tradumatica.344.

Winters, M. and D. Kenny (2023). Mark My Keywords: A Translator-Specific Exploration of Style in Literary Machine Translation. In A. Rothwell, A. Way, and R. Youdale (Eds.), *Computer-assisted Literary Translation*, pp. 69–87. New York, NY: Routledge. https://doi.org/10.4324/9781003357391.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint*. http://arxiv.org/abs/1609.08144.

Zajdel, A. (2022). Catching the Meaning of Words: Can Google Translate Convey Metaphor? In J.L. Hadley, K. Taivalkoski-Shilov, C.S.C. Teixeira, and A. Toral (Eds.), *Using Technologies for Creative-Text Translation*, pp. 116–138. New York: Routledge.

A Distribution of coded themes across participants

Category	PE1	PE2	PE3	PE4	PE5	PE6	Total
Efficiency / Time-saving	8	2	4	6	9	3	32
Faster	0	0	1	0	0	0	1
More work	4	2	3	4	3	1	17
Slower	4	0	0	2	7	2	15
Experience	8	8	5	17	13	2	53
Difficult	1	2	1	3	0	0	7
Doubts about language skills	3	4	2	6	7	0	22
Frustration	0	0	0	3	1	0	4
Fun	1	1	1	0	1	1	5
Lack of context	2	1	1	1	3	0	8
Limits creativity	2	1	1	7	2	1	14
Fragments	1	1	1	1	1	1	6
Fragment 2 problematic	1	1	1	1	1	0	5
No difference	0	0	0	0	0	1	1
Metaphors	17	8	2	8	1	6	42
Noticed issues	16	5	1	8	0	2	32
Noticed positive use	1	1	0	0	0	1	3
Aware after probing	0	1	0	0	1	2	4
Not aware after probing	0	1	1	0	0	1	3
Quality (negative)	10	17	4	15	12	9	67
Overall issues	3	3	3	4	5	0	18
Grammar	0	2	0	0	0	0	2
Too literal	3	6	0	6	5	2	22
Unnatural language	2	5	2	4	2	4	19
Vocabulary	2	1	0	1	0	3	7
Quality (positive)	9	4	0	0	0	4	17
MT sufficient for PE	9	4	0	0	0	4	17
Post-editing process	3	2	5	11	1	2	24
Time/pay affects quality	2	2	4	3	1	2	14
Full rewriting	1	0	0	2	0	0	3
Poor source text quality	0	0	1	7	0	0	8
Style	0	1	2	5	4	7	19
Fiction problematic for MT	0	0	0	3	3	5	11
Fiction suitable for MT	0	0	1	0	0	2	3
Stylistic issues	0	1	1	3	1	0	6
Training	4	4	3	3	3	3	20
MT training	1	1	1	1	0	1	5
MT experience	1	1	1	0	1	0	4
No MT training	0	0	0	0	1	0	1
No MT experience	0	0	0	1	0	1	2
No PE training	1	0	1	0	1	0	3
No PE experience	1	0	0	1	0	1	3
PE training	0	1	0	1	0	1	3
PE experience	0	1	1	0	1	0	3
Total	121	95	55	138	91	75	575

Note. Counts represent coded segments identified in participant responses

B Literal English translation of all "to get under his skin" Dutch renderings

Sentence	Source	Translation (word-by-word English)
"It's important you understand the background," said Sorge, "**not allowing the American to get under his skin.**"	ChatGPT	Without the American under his skin let crawl
	PE1	Who the American no chance gave to get hold of him
	PE2	Who the American the chance not gave to him to chase
	DeepL	Not allowing that the American under his skin
	PE3	Who the American him not on rod let chase
	PE4	Who himself not let chase by the American
	GT	While he the American not under his skin let crawl
	PE5	Insensitive to the nagging of the American
	PE6	While he the American not the chance gave to him on the nerves work