Metaphors in Literary Machine Translation: Close but no cigar?

Alina Karakanta, Mayra Nas, Aletta G. Dorst
Leiden University Centre for Linguistics
[a.karakanta|a.g.dorst]@hum.leidenuniv.nl

Abstract

The translation of metaphorical language presents a challenge in Natural Language Processing as a result of its complexity and variability in terms of linguistic forms, communicative functions, and cultural embeddedness. This paper investigates the performance of different state-of-the-art Machine Translation (MT) systems and Large Language Models (LLMs) in metaphor translation in literary texts (English→Dutch), examining how metaphorical language is handled by the systems and the types of errors identified by human evaluators. While commercial MT systems perform better in terms of translation quality based on automatic metrics, the human evaluation demonstrates that open-source, literary-adapted NMT systems translate metaphors equally accurately. Still, the accuracy of metaphor translation ranges between 64-80%, with lexical and meaning errors being the most prominent. Our findings indicate that metaphors remain a challenge for MT systems and adaptation to the literary domain is crucial for improving metaphor translation in literary texts.

1 Introduction

In 2015, Toral and Way carried out two landmark studies on Literary Machine Translation (LitMT) challenging 'the perceived wisdom [...] that MT is of no use for the translation of literature' (2015a, p. 123) and the claim that literature remains 'the last bastion of human translation' (p. 123). Despite recent improvements in MT quality, they doubted whether MT would be able to tackle what has been called 'perhaps the most creative task a human translator can take on' (Rothwell et al., 2023, p. 10). Yet Toral and Way (2015a; 2015b) convincingly showed that MT has potential in assisting human literary translators, especially in the translation of fiction novels between closely related languages. Their best-performing system equalled professional human quality almost 20% of the time, and a human evaluation with native speakers indicated that over 60% of the translations were considered of the same or even higher quality. A small but steadily growing number of studies has been conducted in LitMT for different genres and languages (Voigt and Jurafsky, 2012; Besacier, 2014; Thai et al., 2022; Toral et al., 2023; Ploeger et al., 2024), showing significant quality gains of literary-adapted NMT systems over general-purpose MT.

Nevertheless, several challenges remain in LitMT. The time is not yet ripe to admit defeat and concede MT's triumphant victory over the human translator. The gap between LitMT and publishable translations is still large, with MT systems lacking in terms of adequacy, style and tone, and the translation of figurative language (Matusov, 2019; Hansen and Esperança-Rodier, 2022). One characteristic of literary texts that continues to pose difficulties is the use of metaphors, which are problematic for NLP (Chakrabarty et al., 2021) and notoriously hard to translate, even for humans, because of their linguistic and cultural embeddedness.

Recently, Large Language Models (LLMs) have demonstrated remarkable performance in several linguistic tasks, including MT (Kocmi et al., 2024). Unlike traditional encoder-decoder models, LLMs have shown potential in translating long documents, in performing style transfer in a zero-shot manner and have even been tested as aids in creative processes (Chakrabarty et al., 2024). These new abilities lead to the questions: Can LLMs address key challenges in LitMT? How well do they perform on metaphor translation, a hallmark of creative expression? To date, only Dorst (2024) and Zajdel (2022) have studied metaphor in LitMT, but each performed a qualitative analysis on a single text and engine. No studies have systematically com-

Proceedings of Machine Translation Summit XX Volume 1, pages 276–286, June 23-27, 2025

pared how different MT systems translate different types of metaphor and whether LLMs offer new possibilities in addressing this persistent problem.

In this paper, we investigate metaphors in LitMT by analysing (i) the performance of state-of-the-art MT systems in translating metaphor and ii) the kinds of errors the different systems make when translating metaphor in literary texts. Our contributions are as follows:

- We compile a new parallel test set of literary texts and their translations (En→NL), annotated with metaphors at the word-level.[1]
- We conduct an evaluation of several commercial and open-source, encoder-decoder and decoder-only, generic and literary-adapted systems in their performance of translating literary texts from English into Dutch using multiple metrics.
- We complement the automatic evaluation with a human evaluation of the accuracy of the systems in translating metaphors, by annotating the errors in metaphor translation and classifying them based on error type.
- Our findings show that metaphor translation is still a challenge in LitMT and that adaptation to the literary domain (regardless of the model architecture) is crucial for addressing metaphor translation in literary texts.

2 Related work

2.1 Metaphor translation

Since the cognitive turn of the 1980s, metaphors are no longer seen as instances of 'deviant' or 'decorative' language use but recognized as a fundamental cognitive tool in human understanding and communication. Metaphors allow us to think and talk about abstract, complex and unfamiliar concepts (such as time, life or arguments) in terms of more concrete, simple and familiar ones (such as concrete objects, movement or living entities). For example, this is why we said in the Introduction that Toral and Way carried out their 'landmark' studies 'in' 2015 and that their studies 'challenged' the widely believed claim that literature is the 'bastion' of human translation. Lakoff and Johnson's (1999; 2008) groundbreaking work showed that such metaphorical uses of words and phrases form systematic patterns in language because they realize underlying conventional conceptual metaphors in thought. For instance, 'in 2015' realizes TIME IS SPACE and both 'challenged' and 'bastion' are realizations of ARGUMENT IS WAR (where beliefs and theories are locations we defend or attack, gaining or losing ground, challenging our opponents until we win or lose the argument). Since most of the metaphors we use are conventional both in language and thought, we normally use and understand them automatically and effortlessly, without realizing they are metaphors.

Yet even highly conventional linguistic metaphors quickly become problematic once we try to translate them. In fact, metaphors have long been considered a notorious problem in translation as a result of their complexity, variability and linguistic and cultural embeddedness – Newmark even went as far as to consider metaphors "the most important particular problem in translation" (1988, p.104). While a small but consistent stream of studies has focused on detailing procedures for metaphor translation (e.g. Van den Broeck (1981); Newmark (1988); Mandelblit (1995); Dickins (2005); Ali (2006)), most of these focus on metaphor at the linguistic level and finding equivalent forms in the target language (but see Schäffner (2017); Shuttleworth (2017)). Very little attention has been paid to the communicative and rhetorical function of metaphor and the role metaphors play in creating aesthetic effects or stylistic coherence, issues particularly relevant in literary translation where style and content are inseparable (Landers, 2001; Boase-Beier, 2014). As illustrated by Dorst (2019) metaphor translation based on local decisions without considering global textual patterns may disrupt a text's stylistic coherence. A subsequent study by Dorst (2024) on the differences between human and machine translations of literary metaphor found that the human translator frequently opted for deletion and normalization, especially for creative metaphors. When the metaphors were conventional, especially fixed collocations and idiomatic expressions, human translators, both professionals and students, showed more (creative) variation in their solutions and the MT system (Google Translate) made more mistakes (lexical and/or grammatical errors). The current study picks up from this point to investigate more systematically how different MT systems –

[1]While the Dutch translations cannot be released at the time of writing due to pending copyright approval, our annotations of metaphors and code can be found at: https://github.com/fatalinha/MetaphorMT/tree/main.

NMT and LLM – compare in their translation of metaphor and the type of errors they make.

2.2 Literary machine translation

The suitability and feasibility of MT for the literary domain has been a long-standing topic of inquiry in MT research. Techniques that have demonstrated improvements include domain adaptation (Toral and Way, 2015a,b; Toral et al., 2023), author-tailored adaptation (Kuzman et al., 2019; Oliver, 2023), restoration of lexical richness to that of the source text (Ploeger et al., 2024) and automatic post-editing (Thai et al., 2022). Studies assessing MT quality in literary contexts have recognised the importance of conducting human evaluations and error analyses of the generated outputs in addition to computing automatic evaluation metrics. While readers seem to rate a significant percentage of MT sentences as acceptable, error-free or equivalent to human translations, with variations across language pairs (34% for English to Catalan (Toral et al., 2018), ∼20% for English into Russian and German (Matusov, 2019), 44% for English into Dutch (Fonteyne et al., 2020)), a recent multilingual study involving 20 language pairs reported that professional translators preferred human translations in 85% of the cases (Thai et al., 2022). Productivity gains from using MT and post-editing have also been reported as moderate success stories of LitMT (Besacier, 2014; Kuzman et al., 2019). Professional translators, however, still prefer human translation over post-editing for literary texts, mentioning sentence-level fragmentation, wrong level of politeness, vocabulary use, figurative language and cultural items as main limitations of MT (Moorkens et al., 2018).

Another line of research has focused on identifying common error types in LitMT. One notable issue identified is that MT systems often struggle with maintaining referential cohesion (Voigt and Jurafsky, 2012) and have limited potential in addressing the difficulties of literary translation (Jones and Irvine, 2013), mainly because the typical sentence-level MT pipeline is insufficient for this task, as document-level context is critical for the literary domain. Although NMT demonstrated clear improvements in fluency over statistical MT (Toral et al., 2018), adequacy errors and mistranslations are still primary sources of failure (Hansen and Esperança-Rodier, 2022), with fluency aspects such as coherence and style & register still being present (Fonteyne et al., 2020). Discourse-level errors such as coreference and pronoun consistency were identified by Thai et al. (2022), along with overly literal translations. From English to Arabic, translations were found to lack proper handling of idioms and colloquialisms (Omar and Gomaa, 2020). As stated above, the current study builds on the analysis of Dorst (2024), contributing to the ongoing research on the feasibility of LitMT by introducing a previously unexplored aspect, that of metaphor translation.

3 Methodology

3.1 Data

The data for evaluating the models consists of excerpts of four English fiction texts from the VUAMC corpus (Steen et al., 2010) and their published Dutch translations. Since the VUAMC corpus only contains English texts, the Dutch translations were scanned from the physical books, OCR'd and corrected, and manually aligned to the English excerpts at the sentence level. The resulting test set contains 482 sentences (about 6700 words). Details about the selected excerpts can be found in Appendix A.

3.2 Models

The models used in this study were selected to cover a wide range of architectures and system types, both encoder-decoder and decoder-only language models, open- and closed-source, generic and literary-adapted. The selection was guided by the best-performing systems in the literary domain from WMT 2024 (Kocmi et al., 2024). Specifically, the models tested were the following:

1. Commercial NMT systems: Google Translate[2], DeepL[3] and ModernMT[4];

2. S3Big: a literary-adapted NMT Transformer model trained using Marian[5] on general-domain and back-translated literary monolingual data, and then fine-tuned on real in-domain data (parallel novels En→Nl). This is a sentence-level model (Toral et al., 2023);

3. General purpose LLMs: GPT4, GPT4o, and GEITje 7B Ultra (Vanroy, 2024), a conversational LLM fine-tuned for Dutch, based on

[2]https://translate.google.com/
[3]https://www.deepl.com/
[4]https://www.modernmt.com/
[5]https://marian-nmt.github.io/

Mistral and aligned with AI feedback via Direct Preference Optimisation; and

4. Translation-adapted LLMs: Tower-Instruct-7B-V0.2 and 13B-V0.1 (Alves et al., 2024). These language models have been trained on multilingual data and fine-tuned on translation-specific data, so as to handle several translation-related tasks, e.g. translation, paraphrasing, automatic post-editing.

LLMs received a simple prompt in the form: "Translate the following sentence from English into Dutch (NL)". GPT4 and GPT4o were accessed through the Trados Studio OpenAI API on 23 July 2024, with temperature set to 0.75. To test whether prompting can have a positive effect on translation quality in LitMT, a literary prompt was also tested with GPT4o and Tower13b, mentioned here as GPT4o-Lit and Tower13b-Lit respectively: "You are a professional translator, specializing in the translation of literary texts. Translate the following sentence from an English novel into Dutch (NL), paying special attention to the translation of metaphors". For the TowerInstruct models, the ChatML prompt templates format was used. The models were tested with the default settings and batch size 256.

3.3 Evaluation

The system outputs were evaluated for general translation quality against the human reference using multiple automatic metrics: SacreBLEU (Post, 2018), COMET (Rei et al., 2020), BERTScore (Zhang* et al., 2020) and MetricX (Juraska et al., 2023) [6]. MetricX is a learned regression-based metric based on the mT5-XXL pretrained language model. It achieved among the highest correlation with human judgements in the WMT 2024 Metrics task (Freitag et al., 2024). We use MetricX-24-Hybrid-Large and the corpus-level score is computed by averaging the segment-level scores. BERTScore was computed using the MATEO framework (Vanroy et al., 2023). The selection of string-based, neural and LLM-based metrics aims to compare the rankings assigned to the systems by different evaluation metrics and examine the relation between various types of metrics and the quality of metaphor translation.

In addition, the systems' accuracy in translating different types of metaphors was assessed via human evaluation. One hundred sentences were randomly selected from the test set (4 chunks of approx. 6 sentences from each novel) containing 333 source metaphors in total. The outputs of the five highest-performing systems according to the automatic metrics, as well as the official translations, were annotated in INCEpTION (Klie et al., 2018) by two professional translators, native speakers of Dutch. The outputs were presented to the annotators without any information about which system generated which sentence. For each metaphor in the source (annotated at the word level following VUAMC), the evaluators had to detect the corresponding translation in Dutch and assess whether the translation was "correct" or "incorrect". Subsequently, the errors were analysed and classified in three categories: meaning errors (the Dutch translation of the source metaphor has the wrong meaning), form errors (the Dutch translation of the source metaphor is ungrammatical or unidiomatic) or omissions, when the metaphor is left out in the translation.

Following VUAMC, the metaphor translations were also annotated at the word level. However, this word-based approach is not without problems, since words in the source may be expressed by a multi-word expression in the target and vice versa. For example, the English verb 'glare' is correctly translated into Dutch as 'boos kijken' [lit. 'angry look'] while 'wiped out' translates as 'vernietigd' ['destroyed']. In addition, metaphors frequently form multi-word expressions (MWE) (e.g. collocations, idiomatic expressions) in which the translation of the metaphorical word may be considered correct in isolation but not in the multi-word expression. For example, in the phrase 'made good time' the verb 'made' is annotated as a source metaphor and in isolation the translation 'maakte' is technically correct, but the combination 'maakte goede tijd' is incorrect because it is ungrammatical and unidiomatic. An additional problem is that in some cases it is clear that there is an error, because the MWE as a whole is incorrect in the Dutch translation, but it is hard to pinpoint which individual word(s) to annotate. Despite this, the word-based approach is necessary to obtain a measure of accuracy in metaphor translation.

After collecting the error annotations, the per-

[6] SacreBLEU:nrefs:1|bs:1000|s:12345|c:mixed|eff:no|tok:13a|smooth:exp|v:2.4.3
BERTScore: nrefs:1|bs:1000|s:12345|l:other|v:0.3.12|mateo:1.1.3
COMET: nrefs:1|bs:1000|s:12345|c:Unbabel/wmt22-comet-da|v:2.2.2

centage of correctly translated metaphors is reported per system. We compute inter-annotator agreement using Cohen's κ (Cohen, 1960), based on whether annotators agree on their judgement of a source metaphor being translated correctly or not. In addition, we report inter-rater reliability (IRR) as the percentage of matches between the two raters.

4 Results

4.1 Automatic evaluation

Table 1 presents the automatic quality scores for the various systems. In response to our first research question "Which is the highest-performing system for LitMT from English into Dutch?", notably, **different types of metrics assign higher scores to different systems, not allowing to pinpoint a clear winner**. The best-performing system based on the neural metrics is the commercial system Google Translate (GT) with a COMET score of 84.02 and a BERTScore of 83.96, while BLEU favours DeepL with a score of 31.31 and 3 points difference from GT, the second-scoring system. However, MetricX scores the output of GPT4o-Lit as the best with a score of 2.0461 (the lower the score the better). Similar to the neural metrics, MetricX scores Google Translate higher than DeepL with a score of 2.1075 and 2.1545 respectively. It appears that MetricX favours the outputs of GPT models, but not those of the Tower models, even though all of them are LLMs. These differences in ranking highlight variations in how each metric evaluates translation quality and the difficulty of relying solely on automatic evaluation in LitMT.

Another hypothesis put forward in the Introduction is that LLMs may outperform NMT systems in LitMT. However, based on the automatic scores, **there is no clear indication that LLMs can surpass NMT systems yet**. LLMs perform similarly with commercial systems and the literary-adapted system S3Big. Bootstrap resampling on COMET and BERTScore scores shows a second-place tie among DeepL, the GPT4 models, and S3Big (light gray). A similar second tie is observed for BLEU scores. This is notable, given that GPT models have not been explicitly trained for translation or on literary data. On the contrary, translation-specific LLMs (Tower7b and 13b) unexpectedly scored significantly lower according to all metrics, forming a third tie together with ModernMT. Lastly, GEITje has the lowest score, despite being fine-tuned for Dutch. This is expected since it is not fine-tuned to the task of translation, which often leads to hallucinations and the inability to follow instructions. Therefore, vanilla LLMs do not seem to be bringing a transformative breakthrough in LitMT yet.

LLM adaptation to translation tasks or the target language did not demonstrate promising results, but does adaptation to the literary domain make a difference? S3Big performs on par with commercial systems, **demonstrating that domain adaptation can still yield high-performing non-commercial NMT systems**. Similar results were reported by Toral et al. (2023) where S3Big showed only a 2% reduction in COMET compared to DeepL. Even though MetricX assigned a low score to S3Big, the best score was assigned to GPT4o-Lit, the system adapted with the literary prompt. For the neural and string-based metrics, the literary prompt (GPT4o-Lit) also led to minor improvements in scores. However, this is not the case with Tower13b, where the literary prompt hurt performance. Both these observations indicate that further adaptation and careful fine-tuning of LLMs to the literary domain could lead to improvements in LitMT, a direction to be explored in future work.

To sum up, the automatic evaluation suggests that commercial NMT systems are the strongest for LitMT, followed by closed-source LLMs. However, the open-source, literature-adapted NMT system S3Big remains competitive despite being trained on significantly less data, demonstrating the effectiveness of domain adaptation. In contrast, open-source LLMs still lag behind, even when specifically trained for translation tasks. However, another question remains: How accurate are these systems in translating metaphors? To answer this question, the top performing systems from each architecture group are selected to conduct a human evaluation of their accuracy in metaphor translation. The selected systems include DeepL, Google Translate (GT), GPT4, Tower13b and S3Big.

4.2 Human evaluation

Table 2 shows the human evaluation scores in metaphor translation for the selected systems, as well as for the human translation (Ref). In general, the scores for the accuracy of translating metaphors range between 64-80%, showing that **metaphors are still a challenge for MT systems**. The literary-adapted NMT system S3Big has the highest accuracy in translating metaphors with 75% of the metaphors on average annotated as correctly trans-

system	MetricX24↓	COMET↑	BERTScore↑	BLEU↑
DeepL	2.1545	83.55	83.46	**31.31**
Google Translate	2.1075	**84.02**	**83.96**	28.47
ModernMT	2.4435	82.89	82.83	26.30
GPT4	2.1464	83.45	83.42	27.59
GPT4o	2.1256	83.18	83.15	26.35
GPT4o-Lit	**2.0461**	83.25	83.20	26.68
GEITje	3.7430	77.64	77.62	14.89
TOWER7b	2.3195	82.73	82.72	23.66
TOWER13b	2.2156	82.81	82.83	24.53
TOWER13b-Lit	2.3778	82.04	82.13	24.94
S3Big	2.3593	83.31	83.30	28.72

Table 1: MetricX24, COMET, BERTScore and BLEU scores of different systems on the En→Nl literary test set. Best score in **bold**. Different colours (light blue , medium blue and dark blue) indicate statistically significant differences between systems. Systems sharing the same colour are not statistically different from each other.

	Ref	DeepL	Google Tr.	GPT4	Tower13b	S3Big
An1	87%	78%	76%	75%	75%	**80%**
An2	87%	**70%**	68%	65%	64%	**70%**
Avg	87%	74%	72%	70%	69.5%	**75%**
κ	0.57	0.47	0.52	0.47	0.46	0.42
IRR	90%	79%	80%	77%	77%	78%

Table 2: Accuracy in the translation of metaphors by the two annotators (An1 and An2) and on average (Avg), Cohen's κ and inter-rater reliability (IRR). Different colours (light blue , medium blue and dark blue) indicate statistically significant differences between systems ($p < .05$) based on pairwise comparisons. Systems sharing the same letter are not significantly different from each other.

lated. The second-best system was found to be DeepL (74%). The highest-performing system based on the neural metrics, GT, comes third (72%). However, a logistic mixed-effects model did not reveal statistically significant differences in accuracy between S3Big, DeepL and GT. LLMs, despite promises to address issues in LitMT, have the lowest scores in metaphor translation with 70% for GPT4 (β = -0.339, p = .019 compared to S3Big) and 69.5% for Tower13b (β = -0.398, p = .006), suggesting that adaptation techniques may be required for these systems to address literary aspects more accurately.

Interestingly, metaphors in the human translation were also sometimes annotated as incorrect (84% accuracy). Most of the identified errors in the human translation were meaning errors or omissions (see Table 3 for a classification of errors). These appeared to occur when the source metaphors were rather hard to interpret or their meaning was ambiguous (for example, 'knotted' in 'once free of the knotted tentacles of the suburbs') or when the translation may have sounded forced or awkward rather than literary and creative. In such cases, the human translator may have decided to go for the "safe" option of omitting the metaphor. After all, as pointed out by Guerberof-Arenas and Toral (2022), creativity involves both novelty and acceptability. This is a particularly interesting area for future investigations: while omission is generally considered an error in MT, it is often a deliberate risk-avoiding strategy in human translation. Future studies could explore in more detail whether metaphor omissions in MT occur in the same contexts and under the same conditions as in HT.

On the total number of annotations, a moderate agreement was found between the annotators with Cohen's κ at 0.49 (Landis and Koch, 1977) and a total IRR of 80%. The annotators agreed more on their assessments of the human translation (κ=0.57, IRR=90%) and less on the metaphors translated by S3Big (κ=0.42, IRR=78%). When comparing the scores of the two annotators, we observe that Annotator 2 was more strict than Annotator 1 when

assessing the machine-translated metaphors, by 9% on average, even though the annotators agree on the percentage of correct human-translated metaphors. The moderate inter-annotator agreement shows that the task of assessing metaphor translations is difficult and even trained professional translators may disagree on whether a particular metaphor translation counts as an error. As discussed above, this may be due to the inherent difficulty of pinpointing whether and where errors occur in metaphor translation. Similar agreement scores have been reported in other studies involving error annotation in literary translation (Fonteyne et al., 2020), showing a potential subjectiveness of error assessment in the literary domain. More importantly, what professional translators or linguists consider errors may at times be considered acceptable or creative by the average reader, especially in literary texts.

5 How do MT systems translate metaphors?

The automatic evaluation (Table 1) indicated that the commercial NMT systems obtained the highest quality overall for literary MT. The human evaluation (Table 2) showed that the literary-adapted NMT system S3Big had the highest accuracy for metaphor translation, together with DeepLand GT, with LLMs falling short. To determine whether the NMT and LLM systems make the same or different types of errors and compare the types of errors with the human translation, Annotator 1 labelled the identified errors for their error type, i.e. Form, Meaning or Omission. Table 3 shows the error types by system. A total of 228 errors were identified in all outputs and the human translation by Annotator 1, divided in 128 meaning errors, 85 form errors and 15 omissions. In general, meaning errors are the most prevalent in all MT systems, however, there are differences: form errors are more common for GT while for LLMs (GPT and Tower) as well as for the literary-adapted system S3Big the difference between the number of meaning errors and form errors is much clearer.

Overall, the observations made during the error annotation support previous findings that lexical errors (mistranslations) are the most frequent type of error. Conversely, this raises the question whether lexical errors are often the most frequent type of error in MT output because of the pervasiveness of metaphors in everyday discourse. A closer look at the translations shows that, as suggested by Dorst (2019, 2024), most of these meaning errors concern highly conventional linguistic metaphors. For example, in 'the dark *mouth* of a concrete pillbox' (S10, C8T), 'mouth' was incorrectly translated by DeepL and GPT as 'mond' [mouth] and by Google Translate as 'monding' [mouth, estuary]; Tower uses the correct 'opening' [opening] and S3Big system the correct 'ingang' [entrance]. Similarly, in '*Wiped out* twenty million Russians' (S132, G0L), Google Translate has the incorrect 'weggevaagd' [erased, swept away], GPT the incorrect 'verwijderd' [removed] and S3Big the incorrect 'uitgebuit' [exploited], while DeepL and Tower use the correct 'uitgeroeid' [exterminated]. Table 3 shows that the human translator made the fewest meaning errors, and the two commercial systems slightly fewer than the LLMs and S3Big.

For the form errors, the situation is slightly different: DeepL and S3Big make fewer form errors than the other systems. Here, the contrast is quite striking between Google Translate (with 22 form errors) and S3Big (with only 9 form errors). This suggests that DeepL and S3Big may be better at correctly translating multi-word metaphors such as idiomatic expressions. For example, both systems correctly translated '*keep* your voice *down*' as 'praat niet zo hard' [lit. talk not so loud] rather than the incorrect (unidiomatic) direct translation 'houd je stem laag' produced by Google Translate and GPT. The translation produced by Tower - 'houd je stem maar eens in' - is particularly puzzling because it sounds idiomatic but is in fact meaningless, since 'inhouden' is something you can do with your breath (e.g. hold your breath) but not with your voice. The combination 'je stem inhouden' simply does not exist in Dutch (but 'je adem inhouden' does). Something slightly different happened with the expression 'get under his skin' (S89, G0L), where DeepL, Google Translate and GPT all have the incorrect direct translation 'onder zijn huid kruipen' (which does not exist as a conventional metaphorical expression in Dutch), while Tower outputs a correct and idiomatic alternative 'van streek te maken' [to upset] and S3Big outputs the idiomatic but incorrect (wrong meaning) 'overdonderen' [= overwhelm].

More examples need to be collected and analysed to determine whether these patterns are consistent across larger datasets but this first exploration may suggest that as NMT and LLMs become better at avoiding incorrect direct translations of multi-word expressions they may start making more meaning errors (which will be harder

System	Meaning	Form	Omission	Total
Ref	10 (50%)	3 (15%)	7 (35%)	20 (100%)
DeepL	18 (49%)	14 (38%)	5 (13%)	37 (100%)
GT	20 (48%)	22 (52%)	0 (0%)	42 (100%)
GPT4	25 (57%)	19 (43%)	0 (0%)	44 (100%)
Tower13b	30 (61%)	18 (37%)	1 (2%)	49 (100%)
S3Big	25 (70%)	9 (25%)	2 (5%)	36 (100%)
Total	128 (100%)	85 (100%)	15 (100%)	228 (100%)

Table 3: Errors in meaning form and omissions for each system and reference translation.

to spot, especially for readers without access to the source text). LitMT is advancing to a stage in which the number of obviously incorrect and unidiomatic translations is decreasing, and some of the metaphor translations are indistinguishable from human translation. However, the big question is whether the remaining errors - both form and meaning - affect the readers' understanding of the metaphors and their role in the narrative. If the cost for obtaining idiomatic metaphor translations is a shift in meaning is that a price we are willing to pay?

6 Conclusion and future work

In this paper, we addressed the translation of metaphor in literary MT from English into Dutch by comparing different transformer-based architectures. We investigated how the different systems translate metaphors and determined what type of errors they tend to make, asking whether LLMs provide new opportunities in tackling this long-standing challenge in NLP. Regarding the performance of state-of-the-art MT systems in metaphor translation, our conclusion is that they come close, but no cigar. The automatic evaluation showed that different types of metrics favoured different systems and no single system consistently outperformed the others. No clear evidence was found indicating that LLMs in their current setting outperform NMT systems in LitMT and metaphor translation, at least for this language pair and type of literary content. Commercial NMT systems produced the overall highest quality output, followed by closed-source LLMs. Notably, the open-source, literature-adapted NMT system S3Big remained competitive despite having been trained on significantly less data, demonstrating the effectiveness of domain adaptation. Additionally, the human evaluation revealed that the accuracy of the systems for metaphor translation specifically ranged between 64-80%, highlighting that metaphors remain a challenge for MT systems. A closer look at the errors identified in human evaluation revealed that most were meaning errors (i.e. lexical) rather than form errors (i.e. grammatical) and most of the errors concerned highly conventional metaphorical expressions.

Our current findings suggest that further research is needed to assess whether the errors made by MT - both in form and meaning - affect readers' understanding of the metaphors and the narrative. In addition, the structural similarities between English and Dutch may result in more "false friend" metaphor translations, which may appear to be fluent and correct while the metaphor technically does not exist in Dutch. The next phase of this project is therefore to also extend it to more languages. Another continuing objective of the current project is to develop a clearer taxonomy to identify and label different types of errors and shifts in metaphor translation, especially given the difficulties in deciding where (which word) the error occurs, what type of error it is, and whether it should be counted as an error or as a creative solution. We are therefore also conducting a reader-response study that investigates how readers respond to the different MT metaphor translations that were classified as errors in our human evaluation.

Sustainability statement

The experiments presented in this paper involving running model inference and computing neural automatic metrics ran for 7h and 30min on 1 GPU NVIDIA A100 40GB PCIe, while larger models ran for 1h on 1 GPU NVIDIA A100 80GB PCIe. In total, our experiments drew 1,389 kgCO2e. Based in [country removed for anonymity], this had a carbon footprint of 3.70 kWh, which is equiva-

lent to 1.52 tree-months. (calculated using green-algorithms.org v3.0 (Lannelongue et al., 2021)).

Limitations

In this paper, we addressed the translation of metaphors in literary MT. However, this does not encompass the translation of metaphors in other domains and genres. We evaluated different systems in a high-resource language pair that consists of relatively similar languages, in one translation direction. Given the sparsity of metaphor-annotated data and difficulty of obtaining literary translations we found this difficult to avoid. Experiments with more languages and literary texts may retrieve richer results. In addition, even though our data is transparent, in the sense that we have reported the exact excerpts and sentence numbers from the VUAMC corpus, for the time being we cannot distribute the Dutch translations ourselves, due to copyright restrictions. Lastly, we acknowledge that the list of models, prompts and settings tested is not exhaustive, even though we found it to be representative of the range of models currently available.

References

Abdul Sahib Mehdi Ali. 2006. On the translation of metaphor: Notions and pedagogical implications. *IJAES*, 7:121–136.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Laurent Besacier. 2014. Machine translation for litterature: a pilot study (traduction automatisée d'une oeuvre littéraire: une étude pilote) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 389–394, Marseille, France. Association pour le Traitement Automatique des Langues.

Jean Boase-Beier. 2014. *Stylistic approaches to translation*. Routledge.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, page 132–155, New York, NY, USA. Association for Computing Machinery.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:1–37.

James Dickins. 2005. Two models for metaphor translation. *Target*, 17(2).

Aletta G Dorst. 2019. Translating metaphorical mind style: Machinery and ice metaphors in ken kesey's one flew over the cuckoo's nest. *Perspectives*, 27(6):875–889.

Aletta G. Dorst. 2024. Metaphor in literary machine translation: style, creativity and literariness. In *Computer-Assisted Literary Translation*, pages 173–186. New York: Routlegde.

Margot Fonteyne, 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. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.

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

Ruth Jones and Ann Irvine. 2013. The (un)faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101, Sofia, Bulgaria. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings*

of the Eighth Conference on Machine Translation, pages 756–767, Singapore. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland. European Association for Machine Translation.

George Lakoff and Mark Johnson. 1999. Philosophy in the flesh—the embodied mindand its challenge to western thought. *NY: Basic Books*.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Clifford E Landers. 2001. Literary translation: A practical guide. *Multilingual Matters*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Adv. Sci.*, 1.

Nili Mandelblit. 1995. The cognitive view of metaphor and its implications for translation theory. *Translation and meaning*, 3(1):483–495.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7:240–262.

Peter Newmark. 1988. *A textbook of translation*. Prentice Hall International.

Antoni Oliver. 2023. Author-tailored neural machine translation systems for literary works. In *Computer-Assisted Literary Translation*, pages 126–141. Routledge.

A. Omar and Y. Gomaa. 2020. The machine translation of literature: Implications for translation pedagogy. *International Journal of Emerging Technologies in Learning (iJET)*, 15:228–235.

Esther Ploeger, Huiyuan Lai, Rik Van Noord, and Antonio Toral. 2024. Towards tailored recovery of lexical diversity in literary machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 286–299, Sheffield, UK. European Association for Machine Translation (EAMT).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Andrew Rothwell, Andy Way, and Roy Youdale. 2023. *Computer-Assisted Literary Translation (1st ed.)*. Routledge.

Cristina Schäffner. 2017. Metaphor in translation. In E. Semino and Z. Demjen, editors, *The Routledge Handbook of Metaphor and Language*, pages 247–262. Abingdon: Routledge.

Mark Shuttleworth. 2017. *Studying scientific metaphor in translation*. Routledge.

G.J. Steen, 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.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Toral, Andreas Van Cranenburgh, and Tia Nutters. 2023. Literary-adapted machine translation in a well-resourced language pair. explorations with more data and wider contexts. In Andrew Rothwell, Andy Way, and Roy Youdale, editors, *Computer-Assisted Literary Translation*. Routledge: New York.

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

Antonio Toral and Andy Way. 2015b. Translating literary text between related languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 123–132, Denver, Colorado, USA. Association for Computational Linguistics.

Antonio Toral, Martijn Wieling, Sheila Castilho, Joss Moorkens, and Andy Way. 2018. Project PiPeNovel: Pilot on post-editing novels. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 385, Alicante, Spain.

Raymond Van den Broeck. 1981. The limits of translatability exemplified by metaphor translation. *Poetics today*, 2(4):73–87.

Bram Vanroy. 2024. Geitje 7b ultra: A conversational model for dutch. *Preprint*, arXiv:2412.04092.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500. European Association for Machine Translation (EAMT).

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Alicja Zajdel. 2022. Catching the meaning of words: Can google translate convey metaphor? In *Using Technologies for Creative-Text Translation*, pages 116–138. Routledge.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Dataset

Metaphor in Fiction sample from VUAMC. The following excerpts have been selected as the test set for this study:

- C8T: *Devices and desires*. James, P D (1989). (sentences 2-14, 27-49, 114-131)
- CDB: *A fatal inversion*. Vine, Barbara (1987). (fragment 02: sentences 380-400, 422-465, fragment 04: 855-881)
- FPB: *Crimson*. Conran, Shirley (1992). (1060-1102, 1249-1290, 1312-1373)
- G0L: *The Lucy ghosts*. Shah, Eddy (1993). (sentences 19-69, 75-152, 162-222)

And their Dutch translations:

- PD James – Melodie des doods
- Ruth Rendell – Ongewenst weerzien
- Shirley Conran - Karmozijn
- Eddy Shah – Het Lucy komplot