Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature

Katherine Thai[★♢] **Marzena Karpinska**[★♢] **Kalpesh Krishna**[♢] **William Ray**[♢]
Moira Inghilleri[♠] **John Wieting**[♣] **Mohit Iyyer**[♢]

[♢]Manning College of Information and Computer Sciences, UMass Amherst
[♠]Department of Languages, Literatures, and Cultures; UMass Amherst
[♣]Google Research
{kbthai,mkarpinska,kalpesh,miyyer}@cs.umass.edu
minghilleri@complit.umass.edu, jwieting@google.com

Abstract

Literary translation is a culturally significant task, but it is bottlenecked by the small number of qualified literary translators relative to the many untranslated works published around the world. Machine translation (MT) holds potential to complement the work of human translators by improving both training procedures and their overall efficiency. Literary translation is less constrained than more traditional MT settings since translators must balance meaning equivalence, readability, and critical interpretability in the target language. This property, along with the complex discourse-level context present in literary texts, also makes literary MT more challenging to computationally model and evaluate. To explore this task, we collect a dataset (PAR3) of non-English language novels in the public domain, each aligned at the paragraph level to both human and automatic English translations. Using PAR3, we discover that expert literary translators prefer reference human translations over machine-translated paragraphs at a rate of 84%, while state-of-the-art automatic MT metrics do not correlate with those preferences. The experts note that MT outputs contain not only mistranslations, but also discourse-disrupting errors and stylistic inconsistencies. To address these problems, we train a post-editing model whose output is preferred over normal MT output at a rate of 69% by experts. We publicly release PAR3 to spur future research into literary MT.[1]

1 Introduction

While the quality of machine translation (MT) systems has greatly improved with recent advances in modeling and dataset collection, the application of these new technologies to the task of automatically translating *literary* text (e.g., novels, short stories) has remained limited to small-scale studies (Genzel et al., 2010; Jones and Irvine, 2013; Toral et al., 2018). Translating literary works differs from translating standard MT corpora (e.g., news articles or parliamentary proceedings) in several key ways. For one, it is much more difficult to evaluate. The techniques[2] used by literary translators differ fundamentally from those applied in more standard MT domains (see Table 8 in the Appendix). Literary translators have the freedom (or burden) of both semantic and critical interpretation, as they must solve the problem of *equivalence*, often beyond the word level (Neubert, 1983; Baker, 2018; Baker and Saldanha, 2021). The task of conveying an author's ideas highlights yet another difference between literary and traditional MT: *document-level* context is especially critical for the literary domain due to the presence of complex discourse structure, rendering the typical sentence-level MT pipeline insufficient for this task (Voigt and Jurafsky, 2012; Taivalkoski-Shilov, 2019).

In this work, we seek to understand how both state-of-the-art MT systems and MT evaluation metrics fail in the literary domain, and we also leverage large pretrained language models to improve literary MT. To facilitate our experiments, we introduce PAR3, a large-scale dataset to study *paragraph-level literary translation* into English. PAR3 consists of 121K paragraphs taken from 118 novels originally written in a non-English language, where each paragraph is aligned to multiple human-written English translations of that paragraph as well as a machine-translated paragraph produced

[1] https://github.com/katherinethai/par3/
★Authors contributed equally.

[2] Many terms have been employed by translation scholars to refer to various operations used by translators (Chesterman, 2005). Here, we employ the term "techniques" argued for by Molina and Hurtado Albir (2004) and recently used in the field of NLP (Zhai et al., 2018, 2020).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9882 - 9902
December 7-11, 2022

by Google Translate (see Table 2).

We show that MT evaluation metrics such as BLEU and BLEURT are not effective for literary MT. In fact, we discover that two of our tested metrics (BLEU and the document-level BLONDE) show a preference for Google Translate outputs over reference translations in PAR3. In reality, MT outputs are much worse than reference translations: our human evaluation reveals that professional translators prefer reference translations at a rate of **85%**.

While the translators in our study identified overly literal translations and discourse-level errors (e.g., coreference, pronoun consistency) as the main faults of modern MT systems, a monolingual human evaluation comparing human reference translations and MT outputs reveals additional hurdles in readability and fluency. To tackle these issues, we fine-tune GPT-3 (Brown et al., 2020) on an automatic post-editing task in which the model attempts to transform an MT output into a human reference translation. Human translators prefer the post-edited translations at a rate of 69% and also observe a lower incidence of the above errors.

Overall, we identify critical roadblocks in evaluation towards meaningful progress in literary MT, and we also show through expert human evaluations that pretrained language models can improve the quality of existing MT systems on this domain. We release PAR3 to spur more meaningful future research in literary MT.

2 The PAR3 Dataset: Parallel Paragraph-Level Paraphrases

To study literary MT, we collect a dataset of **par**allel **par**agraph-level **par**aphrases (PAR3) from public domain non-English-language (*source*) novels with their corresponding English translations generated by both humans and Google Translate. PAR3 is a step up in both scale and linguistic diversity compared to prior studies in literary MT, which generally focus on one novel (Toral et al., 2018) or a small set of poems or short stories (Jones and Irvine, 2013). PAR3 contains at least two human translations for every source paragraph (Table 2). In Table 1, we report corpus statistics by the 19 unique source languages[4] represented in PAR3. PAR3 was curated in four stages: selection of source texts, machine translation of source texts, paragraph alignment, and final filtering. This process closely resembles the paraphrase mining methodology described by Barzilay and McKeown (2001); the major distinctions are (1) our collection of literary works that is $\sim$ 20 times the size of the previous work, (2) our inclusion of the aligned source text to enable translation study, and (3) our alignment at the paragraph, not sentence, level. In this section, we describe the data collection process and disclose choices we made during curation of Version 1 of PAR3. See Section A in the Appendix for more details on the different versions of PAR3.

Src lang	#texts	#src paras	sents/para
French (*fr*)	32	50,070	2.7
Russian (*ru*)	27	36,117	3.3
German (*de*)	16	9,170	4.3
Spanish (*es*)	1	3,279	2.0
Czech (*cs*)	4	2,930	3.0
Norwegian (*no*)	2	2,655	3.4
Swedish (*sv*)	3	2,620	3.2
Portuguese (*pt*)	4	2,288	3.7
Italian (*it*)	2	1,931	2.6
Japanese (*ja*)	9	1,857	4.4
Bengali (*bn*)	2	1,499	3.3
Tamil (*ta*)	1	1,489	3.1
Danish (*da*)	1	1,384	3.6
Chinese[3] (zh)	7	1,320	8.8
Dutch (*nl*)	1	963	3.4
Hungarian (*hu*)	1	892	3.7
Polish (*pl*)	1	399	3.9
Sesotho (*st*)	1	374	4.2
Persian (*fa*)	1	148	4.2
All	118	121,385	3.2

Table 1: Corpus statistics for Version 2 of PAR3 by each of the 19 source languages. The average number of sentences per paragraph refers to only the English human and Google translations of the source paragraphs. We did not count tokens or sentences for source paragraphs because of the lack of a reliable tokenizer and sentence segmenter for all source languages.

2.1 Selecting works of literature

For a source text to be included in PAR3, it must be (1) a literary work that has entered the public domain of its country of publication by 2022 with (2) a published electronic version along with (3) multiple versions of human-written, English translations. The first requirement skews our corpus towards older works of fiction. The second requirement ensures the preservation of the source texts' paragraph breaks. The third requirement limits us

[3]The Chinese texts in PAR3 were written in Classical Chinese, an archaic and very different form of the language currently used today.

[4]Languages in PAR3 represent different language families (Romance, Germanic, Slavic, Japonic, Sino-Tibetan, Iranian, Dravidian, Ugric, and Bantu), with different morphological traits (synthetic, fusional, agglutinative), and use different writing systems (Latin alphabet, Cyrillic alphabet, Bengali script, Persian alphabet, Tamil script, Hanzi, and Kanji/Hiragana/Katakana).

SRC (ru): — Извините меня: я, увидевши издали, как вы вошли в лавку, решился вас побеспокоить. Если вам будет после свободно и по дороге мимо моего дома, так сделайте милость, зайдите на малость времени. Мне с вами нужно будет переговорить			
GTr: "Excuse me; seeing from a distance how you entered the shop, I decided to disturb you. If you will be free after and on the way past my house, so do yourself a favour, stop by for a little time. I will need to speak with you.	HUM1: "Pardon me, I saw you from a distance going into the shop and ventured to disturb you. If you will be free in a little while and will be passing by my house, do me the favour to come in for a few minutes. I want to have a talk with you."	HUM2: "I saw you enter the shop," he said, "and therefore followed you, for I have something important for your ear. Could you spare me a minute or two?"	HUM3: 'Excuse me: I saw you from far off going into the shop, and decided to trouble you. If you're free afterwards and my house is not out of your way, kindly stop by for a short while. I must have a talk with you."
SRC (st): Ho bile jwalo ho fela ha Chaka, mora wa Senzangakhona. Mazulu le kajeno a bokajeno ha a hopola kamoo a kileng ya eba batho kateng, mehleng ya Chaka, kamoo ditjhaba di neng di jela kgwebeleng ke ho ba tshoha, leha ba hopola borena ba bona bo weleng, eba ba sekisa mahlong, ba re: "Di a bela, di a hlweba! Madiba ho pjha a maholo!"			
GTr: Such was the end of Chaka, son of Senzangakhona. The Zulus of today when they remember how they once became people, in the days of Chaka, how the nations ate in the sun because of fear of them, even when they remember their fallen kingdom, they wince in their eyes, saying: "They're boiling, they're boiling! The springs are big!"	HUM1: So it came about, the end of Chaka, son of Senzangakhona. Even to this very day the Zulus, when they think how they were once a strong nation in the days of Chaka, and how other nations dreaded them so much that they could hardly swallow their food, and when they remember their kingdom which has fallen, tears well up in their eyes, and they say: "They ferment, they curdle! Even great pools dry away!"	HUM2: And this was the last of Chaka, the son of Senzangakona. Even to-day the Mazulu remember how that they were men once, in the time of Chaka, and how the tribes in fear and trembling came to them for protection. And when they think of their lost empire the tears pour down their cheeks and they say: 'Kingdoms wax and wane. Springs that once were mighty dry away.'	

Table 2: An example of one source paragraph in PAR3, from Nikolai Gogol's *Dead Souls* (upper example) and from Thomas Mofolo's *Chaka* (lower example) with their corresponding Google translation to English and aligned paragraphs from human-written translations.

to texts that had achieved enough mainstream popularity to warrant (re)translations in English. Our most-recently published source text, *The Book of Disquietude*, was published posthumously in 1982, 47 years after the author's death. The oldest source text in our dataset, *Romance of the Three Kingdoms*, was written in the 14th-century. The full list of literary works with source language, author information, and publication year is available in Table 5 in the Appendix.

2.2 Translating works using Google Translate

Before being fed to Google Translate, the data was preprocessed to convert ebooks to lists of plain text paragraphs and to remove tables of contexts, translator notes, and text-specific artifacts.[5] Each paragraph was passed to the default model of the Google Translate API between April 20 and April 27, 2022. The total cost of source text translation was about 900 USD.[6]

2.3 Aligning paragraphs

All English translations, both human and Google Translate-generated, were separated into sentences using spaCy's Sentencizer.[7] The sentences of each human translation were aligned to the sentences of the Google translation of the corresponding source text using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) for global alignment. Since this algorithm requires scores between each pair of human-Google sentences, we compute scores using the embedding-based SIM measure developed by Wieting et al. (2019), which performs well on semantic textual similarity (STS) benchmarks (Agirre et al., 2016). Final paragraph-level alignments were computed using the paragraph segmentations in the original source text.

2.4 Post-processing and filtering

We considered alignments to be "short" if any English paragraph, human or Google generated, contained fewer than 4 tokens or 20 characters. We discarded any alignments that were "short" and contained the word "chapter" or a Roman numeral, as these were overwhelmingly chapter titles. We also discarded any alignments where one English paragraph contained more than 3 times the number of words than another, reasoning that these were actually misalignments. Thus, we also discarded any alignments with a BLEU score of less than 5. Alignments were sampled for the final version of PAR3 such that no more than 50% of the paragraphs for any human translation were included. Finally, alignments for each source text were then shuffled, at the paragraph level, to prevent reconstruction of the human translations, which may not be in the public domain.

[5] From Japanese texts, we removed artifacts of *furigana*, a reading aid placed above difficult Japanese characters in order to help readers unfamiliar with higher-level ideograms.

[6] Google charges 20 USD per 1M characters of translation.

[7] https://spacy.io/usage/linguistic-features#sbd

2.5 Train, test, and validation splits

Instead of randomly creating splits of the 121K paragraphs in PAR3, we define train, test, and validation splits at the document level. Each literary text belongs to one split, and all translations associated with its source paragraphs belong to that split as well. This decision allows us to better test the generalization ability of systems trained on PAR3, and avoid cases where an MT model memorizes entities or stylistic patterns located within a particular book to artificially inflate its evaluation scores. The training split contains around 80% of the total number of source paragraphs (97,611), the test split contains around 10% (11,606), and the validation split contains around 10% (11,606). Appendix 5 shows the texts belonging to each split.

3 How good are existing MT systems for literary translation?

Armed with our PAR3 dataset, we next turn to evaluating the ability of commercial-grade MT systems for literary translation. First, we describe a study in which we hired both professional literary translators and monolingual English experts to compare reference translations to those produced by Google Translate at a paragraph-level. In an A/B test, the translators showed a strong preference (on 84% of examples) for human-written translations, finding MT output to be far too literal and riddled with discourse-level errors (e.g., pronoun consistency or contextual word sense issues). The monolingual raters preferred the human-written translations over the Google Translate outputs 85% of the time, suggesting that discourse-level errors made by MT systems are prevalent and noticeable when the MT outputs are evaluated independently of the source texts. Finally, we address deficiencies in existing *automatic* MT evaluation metrics, including BLEU, BLEURT, and the document-level BLONDE metric. These metrics failed to distinguish human from machine translation, even preferring the MT outputs on average.

3.1 Diagnosing literary MT with judgments from expert translators

As literary MT is understudied (especially at a document level), it is unclear how state-of-the-art MT systems perform on this task and what systematic errors they make. To shed light on this issue, we hire human experts (both monolingual English experts as well as literary translators fluent in both languages) to perform A/B tests on PAR3 which indicates their preference of a Google Translate output paragraph (`GTr`) versus a reference translation written by a human (`HUM`). We additionally solicit detailed free-form comments for each example explaining the raters' justifications. We find that both monolingual raters and literary translators strongly prefer `HUM` over `GTr` paragraphs, noting that overly literal translation and discourse errors are the main error sources with `GTr`.

Experimental setup: We administer A/B tests to two sets of raters: (1) monolingual English experts (e.g., creative writers or copy editors), and (2) professional literary translators. For the latter group, we first provided a source paragraph in German, French, or Russian. Under the source paragraph, we showed two English translations of the source paragraph: one produced by Google Translate and one from a published, human-written translation.[8] We asked each rater to choose the "better" translation and also to give written justification for their choice (2-3 sentences). While all raters knew that the texts were translations, they did **NOT** know that one paragraph was machine-generated. Each translator completed 50 tasks in their language of expertise. For the monolingual task, the set up was similar except for two important distinctions: (1) **NO** source paragraph was provided and (2) each monolingual rater rated all 150 examples (50 from each of 3 language-specific tasks). Tasks were designed and administered via Label Studio,[9] an open-source data-labeling tool, and raters[10] were hired using Upwork, an online platform for freelancers.[11] For the completion of 50 language-specific tasks, translators were paid $200 each. For the set of 150 monolingual tasks, raters were paid $250 each. All raters were given at least 4 days to complete their tasks.

Common MT errors: We roughly categorize the errors highlighted by the professional literary translators into five groups. The most pervasive error (constituting nearly half of all translation errors identified) is the **overly literal** translation of the

[8] Each English paragraph was 130-180 words long.

[9] https://labelstud.io/

[10] For the language-specific task, raters were required to be professional literary translators with experience translating German, French, or Russian to English. We hired one translator for each language. For the monolingual task, we hired three raters with extensive experience in creative writing, copy-editing, or English literature.

[11] https://www.upwork.com/

source text, where a translator adheres too closely to the syntax of the source language, resulting in awkward phrasing or the mistranslation of idioms. The second most prevalent errors are **discourse** errors, such as pronoun inconsistency or coreference issues, which occur when context is ignored–these errors are exacerbated at the paragraph and document levels. We define the rest of the categories and report their the distribution in Table 3.

Monolingual vs translator ratings: Though the source text is essential to the practice of translation, the monolingual setting of our A/B testing allows us to identify attributes other than translation errors that distinguish the MT system outputs from human-written text. Both monolingual and bilingual raters strongly preferred HUM to GTr across all three tested languages[12], as shown in Figure 1, although their preference fell on Russian examples. In a case where all 3 monolingual raters chose HUM while the translator chose GTr, their comments reveal that the monolingual raters prioritized clarity and readability:

> [HUM] "is preferable because it flows better and makes better sense" and "made complete sense and was much easier to read"

while the translator diagnosed HUM with a catastrophic error:

> "[HUM] contains several mistakes, mainly small omissions that change the meaning of the sentence, but also wrong translations ('trained European chef' instead of 'European-educated chef')."

For an example where all 3 monolingual raters chose [GTr] while the translator chose [HUM], the monolingual raters much preferred the contemporary language in [GTr]:

> [GTr] was "much easier for me to grasp because of its structure compared to the similar sentence in [HUM]" and praised for its "use of commonplace vocabulary that is understandable to the reader."

However, the translator, with access to the source text, identified a precision error in GTr, and ultimately declared HUM to be the better translation:

> "*lord* from [HUM] is the exact translation of the Russian бари while *bard* from [GTr] doesn't convey a necessary meaning."[13]

3.2 Can automatic MT metrics evaluate literary translation?

Expert human evaluation, while insightful, is also time-consuming and expensive, which precludes its use in most model development scenarios. The MT community thus relies extensively on *automatic* metrics that score candidate translations against references. In this section, we explore the usage of three metrics (BLEU, BLEURT, and BLONDE) on literary MT evaluation, and we discover that none of them can accurately distinguish GTr text from HUM. Regardless of their performance, we also note that most automatic metrics are designed to work with sentence-level alignments, which are rarely available for literary translations because translators merge and combine sentences. Thus, developing domain-specific evaluation metrics is crucial to make meaningful progress in literary MT.

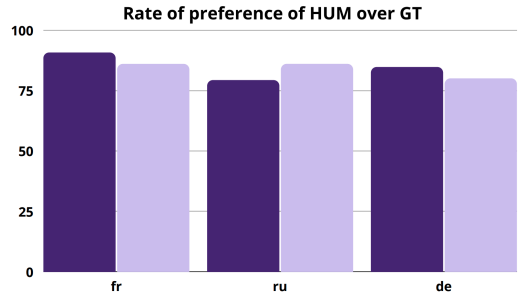

Figure 1: The percentage of cases in which raters preferred the human-written translation to the Google translation by source language. Note that the value for monolingual raters is the average of 3 percentages for 3 monolingual raters.

MT Metrics: To study the ability of MT metrics to distinguish between machine and human translations, we compute three metrics on PAR3:
BLEU (Papineni et al., 2002)[14] is a string-based multi-reference metric originally proposed to evaluate sentence-level translations but also used for document-level MT (Liu et al., 2020).
BLEURT (Sellam et al., 2020) is a pretrained language model fine-tuned on human judgments of translation-reference pairs.[15] BLEURT has been

[12] We report Krippendorff's alpha (Krippendorff, 2011) as the measure of inter-annotator agreement (IAA). The IAA between the monolingual raters was 0.546 (0.437 for Russian, 0.494 for German, and 0.707 for French). The IAA between the aggregated votes of monolingual raters (majority vote) and the translator was 0.524 for Russian, 0.683 for German, and 0.681 for French. These numbers suggest moderate to substantial agreement (Artstein and Poesio, 2008).

[13] To view the SRC, HUM, and GTr texts for these examples, see Tables 13 and 14 in the Appendix.

[14] We compute the default, case-sensitive implementation of BLEU from https://github.com/mjpost/sacrebleu.

[15] We compute BLEURT for PAR3 using the recommended and most recent checkpoint, BLEURT-20. As the maximum input length for BLEURT is 512 sentencepiece tokens, we exclude inputs which exceed this length and would be otherwise truncated. In total, 1.4% of the dataset was excluded.

Example	Error Type (%)	Translator Comments
From *The Sin of Abbé Mouret*, Emile Zola SRC: L'abbé Mouret dépensa **là** ses économies du séminaire. C'étaient, d'ailleurs, des embellissements dont la naïveté maladroite eût fait sourire. La maçonnerie le rebuta vite. Il se contenta de recrépir le tour de l'église, à hauteur d'homme. La Teuse **gâchait** le plâtre. HUM: Abbé Mouret spent all his seminary savings **on the work**. His embellishments were so clumsy and naive as to raise a smile. The masonry-work soon lost its appeal for him. He contented himself with replastering all round the church to the height of a man's head. La Teuse **mixed** the plaster. GTr: Father Mouret spent his seminary savings **there**. They were, moreover, embellishments whose clumsy simplicity would have made you smile. Masonry soon put him off. He contented himself with replastering around the church, at eye level. La Teuse **ruined** the plaster.	**Discourse** (20.8%) Issues created by lack of context. **Word sense** (7.3%) Incorrect translation chosen where multiple are valid.	The first line in French includes the adverb **"là" which means "there"**. In the translation I selected, **"là" is translated by "on the work"** to mean that that is what the Abbé spent all his money on. It makes it easier to understand and the text flows better. The verb **"gâcher" usually means "to waste" / "to ruin"**. However, when used with "plâtre" (=plaster), it means **"to mix" / "to temper"**– This is a collocation that the author of the second translation missed . . . but that was translated correctly in the passage I selected.
From *We*, Yevgeny Zamyatin SRC: Проснулся: умеренный, синеватый свет; блестит стекло стен, стеклянные кресла, стол. Это успокоило, сердце перестало колотиться. Сок, Будда. . . что за абсурд? Ясно: болен. Раньше я никогда не видел снов. Говорят, у древних это было самое обыкновенное и нормальное – видеть сны. Ну да: ведь и вся жизнь у них была вот такая ужасная карусель: зеленое – оранжевое – Будда – сок. Но мы-то знаем, что сны – это серьезная психическая болезнь. И я знаю: до сих пор мой мозг был хронометрически выверенным, сверкающим, без единой соринки механизмом, а теперь. . . Да, теперь именно так: я чувствую там, в мозгу, какое-то инородное тело – как тончайший ресничный волосок в глазу: всего себя чувствуешь, а вот этот глаз с волоском – нельзя о нем забыть ни на секунду. . . HUM: I woke: soft, bluish light, glimmer of glass walls, glass chairs and table. **This calmed me; my heart stopped hammering. Sap**, Buddha ... what nonsense! Clearly I must be ill. I have never dreamed before. They say that with the ancients dreaming was a perfectly ordinary, normal occurrence. But of course, their whole life was a dreadful whirling carousel—green, orange, Buddhas, **sap**. We, however, know that dreams are a serious psychic disease. And I know that until this moment my brain has been a chronometrically exact gleaming mechanism without a single speck of dust. But now . . . Yes, precisely: I feel some alien body in my brain, like the finest eyelash in the eye. **You do not feel your body, but that eye with the lash in it—you can't forget it for a second.** GTr: Awake: moderate, bluish light; glittering glass walls, glass chairs, table. **It calmed her down and her heart stopped beating. Sok**, Buddha... what an absurdity? Obviously sick. I have never dreamed before. They say that among the ancients it was the most ordinary and normal thing—to dream. Well, yes: after all, their whole life was such a terrible carousel: green - orange - Buddha - **juice**. But we know that dreams are a serious mental illness. And I know: until now, my brain was a chronometrically verified, sparkling, without a single mote mechanism, but now. . . Yes, now it's exactly like this: I feel there, in the brain, some kind of foreign body—like the thinnest ciliary hair in the eye: **everything you feel yourself, but this eye with a hair—you can't forget about it for a second. . .**	**Overly literal** (48.4%) The translation adheres too closely to the syntax of the source language. **Precision** (7.3%) The translation is either too specific or not specific enough. **Catastrophic** (16.1)% Errors that completely invalidate the translation.	The **last sentence** of the passage is pretty tough and requires an understanding of the context. The author of the first translation did a great job and conveyed the meaning of the source sentence properly. The author of the second translation made a mistake by using **word-by-word translation: "everything you feel yourself."** As a result, the phrase makes no sense. The author of the source text mentions "сок" which can be translated as **"sap"** (as in the first translation). The author of the second translation decided to transcribe this word in one sentence as **"Sok"** (which doesn't convey the meaning of the Russian word at all) and then translated it as **"juice"**. According to the source text the narrator is male and he tells a story about himself. There is **a sentence "It calmed her down and her heart stopped beating" in the second translation which makes no sense** if we compare it to the Russian text.

Table 3: Definitions and examples of the five types of translation errors on Google Translate outputs identified by professional literary translators. We report their prevalence as a percentage of all errors identified by the translators and include the translators' explanations.

shown to be effective on document-level tasks such as summarization (Kasai et al., 2021).

BLONDE (Jiang et al., 2022)[16] is a document-level multi-reference evaluation metric that considers discourse coherence by calculating the F1 of four "discourse categories" that each represent a feature of coherence across sentences, such as **tense** or **pronoun** consistency.

Comparing HUM to GTr: Since PAR3 contains a variable number of references for each paragraph, we aggregate metric scores across all references for fair comparison, following the methodology of prior work in crowdsourcing multiple reference translations (Callison-Burch, 2009; Zaidan and Callison-Burch, 2011). Given a specific source paragraph with an aligned translation GTr produced by Google Translate and a set of n human reference paragraphs $\mathsf{HUM}_{1...n}$, we compute aggregate scores (s_{HUM}) of a given metric METRIC for human references (against each other) as:

$$s_{HUM} = \sum_i \frac{\textsc{Metric}(\mathsf{HUM}_i, \mathsf{HUM}_{1...n} - \{\mathsf{HUM}_i\})}{n}$$

We use the same reference sets for each example to compute aggregate scores for Google Translate outputs, which ensures that the numbers are comparable:

$$s_{GTr} = \sum_i \frac{\textsc{Metric}(\mathsf{GTr}, \mathsf{HUM}_{1...n} - \{\mathsf{HUM}_i\})}{n}$$

For BLEURT, which unlike BLEU and BLONDE is not well-defined for multiple references, we compute $\textsc{Bleurt}(\mathsf{GTr}, \mathsf{HUM}_{1...n} - \{\mathsf{HUM}_i\})$ by taking the average over pairwise BLEURT scores between GTr and each reference.

[16] We compute BLONDE.F1, simply referred to as BLONDE in the original paper.

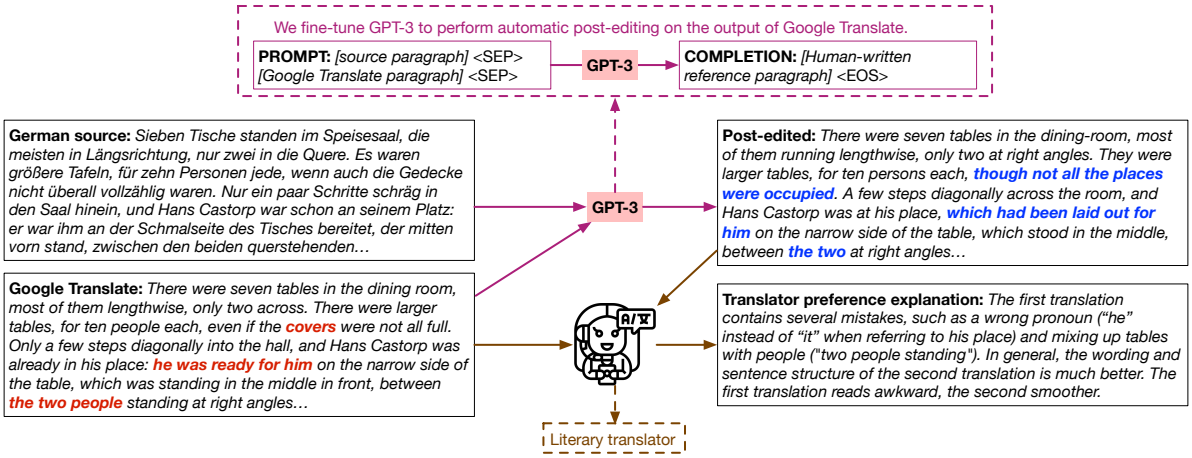

Figure 2: An illustration of our automatic post-editing model on a PAR3 source paragraph from Thomas Mann's *The Magic Mountain*, which fine-tunes GPT-3 to transform a Google Translate paragraph into a human-written reference translation. We hire professional literary translators (in this case, a German translator) to perform a blind A/B test comparing Google Translate against the post-edited output and also to provide free-form explanations as to why they made their choice. In this case, and **69.3%** of the time overall, they prefer the post-editing model's output.

Automatic metrics are not predictive of literary MT quality: We have identified mistakes made by Google Translate in Section 3.1 and the human translations have all been professionally edited and published. Hence, we expect automatic metrics to prefer the human translations. However, we show in Table 4 that two of our three metrics, BLEU and BLONDE,[17] fail to distinguish meaningfully between the human and Google translations, preferring the Google translation to the human one in over 60% of cases. For BLEURT, the choice between Google and human is nearly chance, with human translations preferred 53.6% of the time. A Wilcoxon signed-rank test (Wilcoxon, 1945) reveals that both BLEU (z=-67.344, p<.001, r=.192) and BLONDE (z=-62.862, p<.001, r=.179) prefer Google Translate over human translation. BLEURT, on the other hand, appears to correctly distinguish between the human translation and Google Translate (z=42.462, p<.001, r=.118); however, the effect size is small (r<.30) in all three cases.[18]

4 Can automatic post-editing improve literary MT?

From the experiments in the previous section, we can conclude that human expert evaluation is currently the only way to judge the quality of literary MT. We now turn to improving the quality of Google Translate outputs on PAR3 via *automatic* post-editing (Chatterjee et al., 2018), in which a model corrects the output of a black-box MT system. While Toral et al. (2018) show that manual post-editing on top of MT outputs aids human translator efficiency in the literary domain, no prior work has applied automatic post-editing to literary translations. As shown in Figure 2, we feed both the source paragraph and the Google Translate output to the GPT-3 (Brown et al., 2020) language model, which has been shown to have zero-shot translation capability (although far below state-of-the-art supervised MT systems). We fine-tune GPT-3 to produce a human-written reference translation given these inputs and find that it mitigates issues with overly literal translation and discourse errors.

4.1 Literary post-editing with GPT-3

Our analysis experiments reveal that discourse-level errors that span multiple sentences (e.g., coreference, stylistic consistency, contextual word sense selection) are a huge problem for Google Translate when applied to literary MT. Motivated to address these issues, we select the 175B parameter GPT-3 *davinci* model as our base post-editing system, as it can operate over paragraph-length inputs (max sequence length of 2048 tokens), encode text in multiple languages, and exhibits impressive ability to learn complex tasks with limited training data. To form fine-tuning examples for GPT-3, we concatenate a source paragraph `SRC`, an aligned Google Translate paragraph `GTr`, and a human reference translation `HUM` using special separator and end-of-

[17] Jiang et al. (2022) show that BLONDE has very high correlation to BLEU.

[18] We also perform bootstrapping which yields comparable results.

Source lang	BLEU		BLEURT		BLONDE	
	HUM	GTr	HUM	GTr	HUM	GTr
fr	26.8	29.4	0.630	0.630	25.6	27.5
ru	28.8	29.6	0.642	0.622	25.2	26.0
de	23.1	24.6	0.598	0.597	22.0	23.6
no	29.0	26.5	0.628	0.595	28.3	29.6
es	24.8	22.4	0.623	0.547	27.4	24.2
cs	15.4	20.4	0.560	0.566	14.6	20.2
sv	36.7	36.4	0.680	0.669	39.5	41.0
pt	31.8	27.9	0.646	0.598	29.2	27.3
it	21.8	24.6	0.646	0.628	23.3	24.8
ja	14.8	12.5	0.568	0.512	15.0	12.5
bn	10.4	12.1	0.596	0.572	9.9	11.1
ta	15.5	14.6	0.581	0.561	11.2	10.4
da	26.7	25.5	0.614	0.566	19.1	16.8
zh	11.8	11.7	0.482	0.434	8.7	8.8
nl	26.0	23.9	0.640	0.625	23.1	22.3
hu	26.3	19.5	0.640	0.602	26.4	18.7
pl	34.89	18.5	0.667	0.563	28.2	14.8
st	16.9	15.78	0.559	0.499	16.4	14.7
fa	15.2	16.2	0.540	0.503	8.9	11.0
All	26.4	27.6	0.536	0.613	24.5	25.6
Win %	38.2%	61.8%	53.6%	46.4%	37.4%	62.6%

Table 4: Average BLEU, BLEURT, and BLONDE scores for PAR3 by source language, computed using the same reference set on human and Google translations. See Section 3.2 for details on the computation of the average metric score. The Win % in the final row is the percentage of cases, out of 121,385 unique source paragraphs, in which the metric prefers the human or the Google translation.

sequence tokens:[19]

seq = SRC <SEP> GTr <SEP> HUM <EOS>

where SRC <SEP> GTr is considered the *prompt* and HUM <EOS> is the *completion*.

Data filtering: Before fine-tuning our model, we filtered the PAR3 training set to remove examples where the aggregated BLEU scores between GTr and HUM were either in the 10^{th} or 90^{th} percentiles, which ignores both noisy alignments and near-perfect GTr outputs that do not need any edits. For each example, we also only use the HUM paragraph with the maximum BLEU against the GTr output for that source paragraph, since we could not use all references during fine-tuning.[20] Finally, we randomly sample 30K of the filtered training examples because of the prohibitive cost of fine-tuning and using the GPT-3 *davinci* model.[21] See Appendix C for our fine-tuning configuration.

4.2 Human evaluation of post-edited outputs

Having established that human evaluation is critical for literary MT, we had the same 3 professional translators perform A/B testing on GTr and the outputs of our post-editing model GPT-3.[22] The translators prefer GPT-3 over GTr at a rate of 69% (p<.001, 95% CI [0.613, 0.770]). The comments show that the model often improved on the "overly literal" nature of many GTr paragraphs:

> "The phrase с знакомыми, очень знакомыми улыбкой и взглядом from the source text should be translated as 'with a familiar, very familiar smile and gaze' as in [GPT-3]. The author of [GTr] makes mistakes in choosing the words and suggests "with acquaintances, very familiar smile and look."[23]

Finally, we also had the 3 professional translators perform A/B testing on the post-edited GPT-3 outputs and HUM. While the translators still preferred HUM, their preference rate decreased from 84% (vs. GTr) to 63%. Their comments reveal an interesting caveat to their judgments: overall, raters are much more confident when selecting GPT-3 than when selecting GTr when choosing between the two machine translations. When they did choose GTr, they were often unsure because both translations were equivalently good or bad. When comparing HUM to GPT-3, our annotators were unsure around half of the time, regardless of whether they selected HUM or GPT-3 (they were slightly more confident when choosing HUM), suggesting that the task of discerning the better translator was particularly challenging. We present the results of a small-scale quantitative analysis of the 150 comments across the 3 raters in Figure 3.

Characterizing the behavior of GPT-3 post-editing: We performed a fine-grained analysis of the comments provided by professional translators regarding the behavior of the GPT-3 post-editing model. Overall, the translators observe several positives, including correcting pronoun errors and mistranslations in addition to better capturing the sense of the original work compared to GTr. For example, the professional Russian translator noted an

[19] For the separator token between the source and the Google translation paragraphs, we arbitrarily chose "##" and for the separator token between the prompt and the completion, we used "\n\n###\n\n" as recommended by OpenAI guidelines. The EOS token (stop sequence) was "DNE".

[20] We excluded any examples in which the total number of tokens in the source, GTr, and human paragraphs was greater than 2,000, as GPT-3 training examples (prompt *and* completion) must be fewer than 2,048 tokens.

[21] *davinci* costs 0.03 USD per 1k tokens to fine-tune and 0.12 USD per 1k tokens to use a fine-tuned model.

[22] We report the scores of the 3 automatic MT metrics on the outputs of GPT-3 in Table 9 in the Appendix.

[23] See Table 15 in the Appendix for the texts.

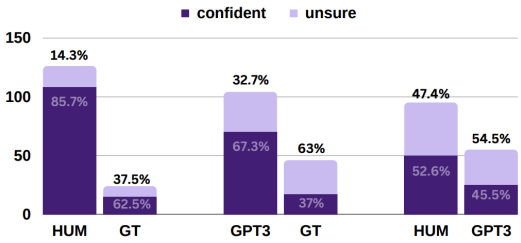

Figure 3: The number of votes for HUM vs GTr, GPT-3 vs GTr, and HUM vs GPT-3 along with their corresponding raters' confidence.

instance where GPT-3 resolved GTr's inconsistent use of character names:

> The narrator mentions one character whose name is "Ippolit." The author of the [Gtr] translation uses this spelling, but then changes it into "Hippolyte" for no reason.

On the other hand, the GPT-3 text occasionally omits some details or contains stylistic choices with which the translators disagree:

> The only reason I selected [GTr] is because I like how it kept the full French nickname "La Teuse" and did not translate the determiner into "the Teuse" [as GPT-3 did].

We show more examples of the post-edit model correcting errors in GTr in Table 7. In 31 out of 150 cases, the translators felt that GPT-3 did not notably improve upon GTr, often mentioning that neither translation was preferred.

Did GPT-3 see the translations in pretraining? One potential criticism of the GPT-3 post-editing model is that it may have seen the reference translations in its pretraining data, and thus any improvements could just be a result of memorization rather than actual understanding of the source text. We qualitatively measure this by creating a small dataset of translated paragraphs that could not have been seen by GPT-3. These translations were either published after GPT-3 was pretrained (2022), or manually translated by an author of this paper from previously untranslated works of literature. Even on this previously unseen data, our model can correct mistranslations, grammatical errors, and stylistic inconsistencies:

> SRC: 朱丽默默走下楼去，都没坐电梯，一路回想惊鸿一瞥的明成的脸。
>
> GTr: Zhu Li walked downstairs silently, without taking the elevator, all the way back to the face of Ming Cheng who had caught a glimpse.
>
> GPT-3: Zhu Li walked downstairs in silence, without taking the elevator, and all the way back she kept recalling the face of Ming Cheng, which she had seen for a moment.

5 Related Work

Our work builds on previous work in literary machine translation. Some early work focused on *poetry* translation (Genzel et al., 2010; Jones and Irvine, 2013), which has recently been tackled with neural approaches (Chakrabarty et al., 2021). Other works have targeted novels, like those in PAR3, with focuses on manual post-editing (Toral et al., 2018, 2020) and comparisons of neural MT to statistical MT systems (Moorkens et al., 2018; Toral and Way, 2018, 2015). Most of these works experiment with datasets far smaller than PAR3 (Arenas and Toral, 2022; Fonteyne et al., 2020). More recent work has involved studying the linguistic characteristics of post-edited literary machine-translated text (Castilho and Resende, 2022; Macken et al., 2022).

Work towards document-level NMT has built on sentence-level MT (Tiedemann and Scherrer, 2017; Jean et al., 2017; Bawden et al., 2018; Miculicich et al., 2018; Agrawal et al., 2018). The a critical lack of parallel document-level corpora has inspired the creative use of parallel sentence-level data (Zhang et al., 2018) and techniques for creating parallel document-level data (Junczys-Dowmunt, 2019). Our work also builds on efforts to tackle discourse-level errors specific to document-level MT and is very similar to that of Voita et al. (2019), but we specifically focus on the literary domain.

6 Conclusion

We study document-level literary machine translation by collecting a dataset (PAR3) of 121K parallel paragraphs from 104 novels. Our experiments show that existing automatic metrics of translation quality are not meaningful in the literary domain. A human evaluation experiment with professional literary translators reveals that commercial-grade MT systems are too literal in their translations and also suffer from discourse-level errors. We mitigate these problems to a certain extent by developing an automatic post-editing model using GPT-3. Overall, our work uncovers new challenges to progress in literary MT, and we hope that the public release of PAR3 will encourage researchers to tackle them.

Limitations

While PAR3 covers a diverse array of genres and languages, there are potential confounding factors in the translation data to be aware of when performing analysis or modeling on top of it. First, multiple human translations of the same source text may not have been written independently: a later translator might have used an earlier translation as a reference, or a new translation may be commissioned because of dissatisfaction with older translations. Additionally, translators in our dataset differ in aspects such as years of experience, familiarity with the author of the source text (some were the exclusive translator for a single author), and bilinguality. The circumstances of each translation are also unique geographically and temporally. It is unclear whether (or how) to model such differences computationally, but it is an intriguing direction for future work.

We also acknowledge that our dataset has a single target-language; the curation of data in other target languages and the improvement of literary MT for other target languages is an essential step towards an equitable and more culturally-conscious field of NLP.

Ethical Considerations

We acknowledge that the vast majority of the authors of our source texts are male. Because literary translation requires training, time, and money, our source texts skew towards older texts that achieved international popularity. We hope that our efforts towards better literary MT can aid literary translators in sharing more minority voices. The experiments involving humans were IRB-approved, and each hired rater was fairly compensated, with wages adjusted as we determined the average amount of time each task took.

Acknowledgements

We would like to thank the translators and English language professionals hired on Upwork for the efforts they put in the evaluation and their insightful comments. We would also like to show our appreciate to Tu Vu for sharing his knowledge about MT evaluation metrics and to Sergiusz Rzepkowski for his help in cleaning the data, as well as multiple translators whom we consulted when exploring our dataset. Finally, we would like to thank the UMass NLP community for their insights and discussions during this project. This project was partially supported by awards IIS-1955567 and IIS-2046248 from the National Science Foundation (NSF) as well as an award from Open Philanthropy.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.

Ana Guerberof Arenas and Antonio Toral. 2022. Creamt: Creativity and narrative engagement of literary texts translated by translators and nmt. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 355–356.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Mona Baker. 2018. *In other words*, 3 edition. Routledge, London, England.

Mona Baker and Gabriela Saldanha, editors. 2021. *Routledge encyclopedia of translation studies*, 3 edition. Taylor & Francis, London, England.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language*

Processing, pages 286–295, Singapore. Association for Computational Linguistics.

Sheila Castilho and NatÃ¡lia Resende. 2022. Post-editese in literary translations. *Information*, 13(2).

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Rubino Raphael, and Marco Turchi. 2018. Findings of the wmt 2018 shared task on automatic post-editing. In *Third Conference on Machine Translation (WMT)*, pages 723–738. Association for Computational Linguistics (ACL).

Andrew Chesterman. 2005. Problems with strategies. In K. Károly and A. Fóris, editors, *Trends in Translation Studies. In honour of Kinga Klaudy.*, pages 17–28. Akademiai Kiado, Budapest.

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 12th Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *ArXiv*, abs/1704.05135.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

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.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. 2021. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv preprint arXiv:2112.04139*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lieve Macken, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process : machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110. European Association for Machine Translation.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Lucía Molina and Amparo Hurtado Albir. 2004. Translation techniques revisited: A dynamic and functionalist approach. *Meta*, 47(4):498–512.

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(2):240–262.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Albrecht Neubert. 1983. Discourse analysis of translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kristiina Taivalkoski-Shilov. 2019. Free indirect discourse: an insurmountable challenge for literary mt systems? In *Proceedings of the Qualities of Literary Machine Translation*, pages 35–39.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Antonio Toral, Antoni Oliver, and Pau Ribas Ballestín. 2020. Machine translation of novels in the age of transformer. *arXiv preprint arXiv:2011.14979*.

Antonio Toral and Andy Way. 2015. Translating literary text between related languages using smt. In *CLfL@NAACL-HLT*.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, page 9.

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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

Yuming Zhai, Lufei Liu, Xinyi Zhong, Gbariel Illouz, and Anne Vilnat. 2020. Building an English-Chinese parallel corpus annotated with sub-sentential translation techniques. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4024–4033, Marseille, France. European Language Resources Association.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Dataset Versions

The first version of PAR3 was created in April 2022 as described in Section 2. The post-edit model and all human evaluations were conducted on this version of the dataset, which can still be found at `https://github.com/katherinethai/par3/`. In October 2022, we expanded PAR3 to include three additional languages: Bengali, Sesotho, and Danish, along with new books in Russian and German. Those texts were translated using the Google Translate API in September 2022. The remaining data processing steps were the same.

B Post-editing Details

Automatic evaluation of post-edited texts: We compute BLEU, BLEURT, and BLONDE on the outputs of the post-editing model and present the results by source language in Table 9. All 3 metrics show a clear preference for the human translations or the post-edited outputs of GPT-3.

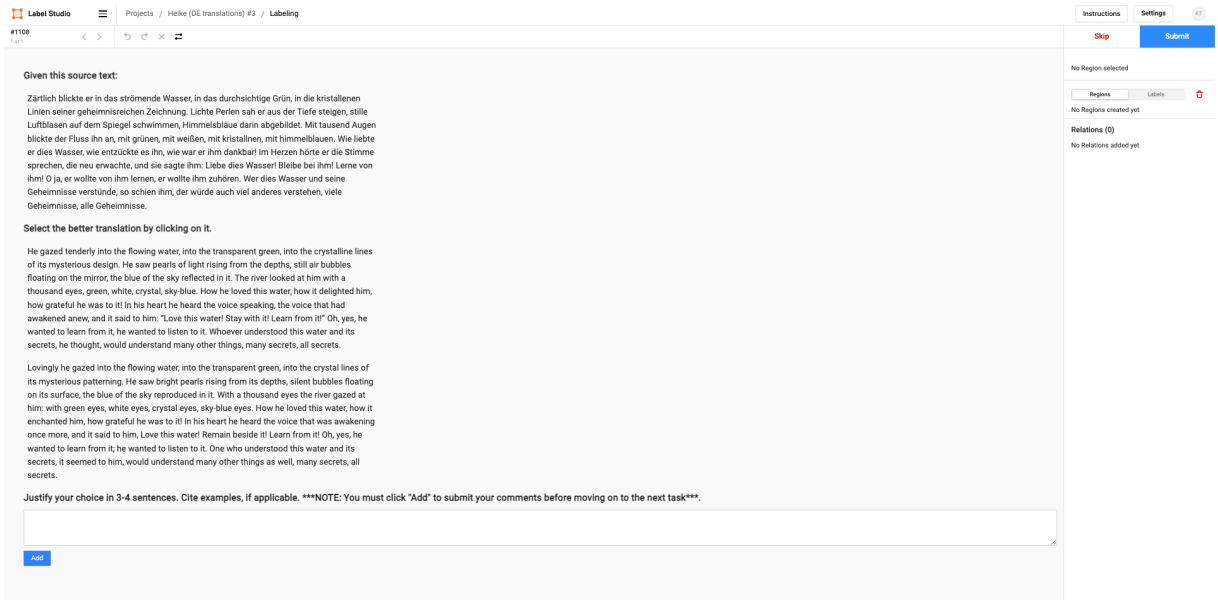

Figure 4: Example of the labeling interface.

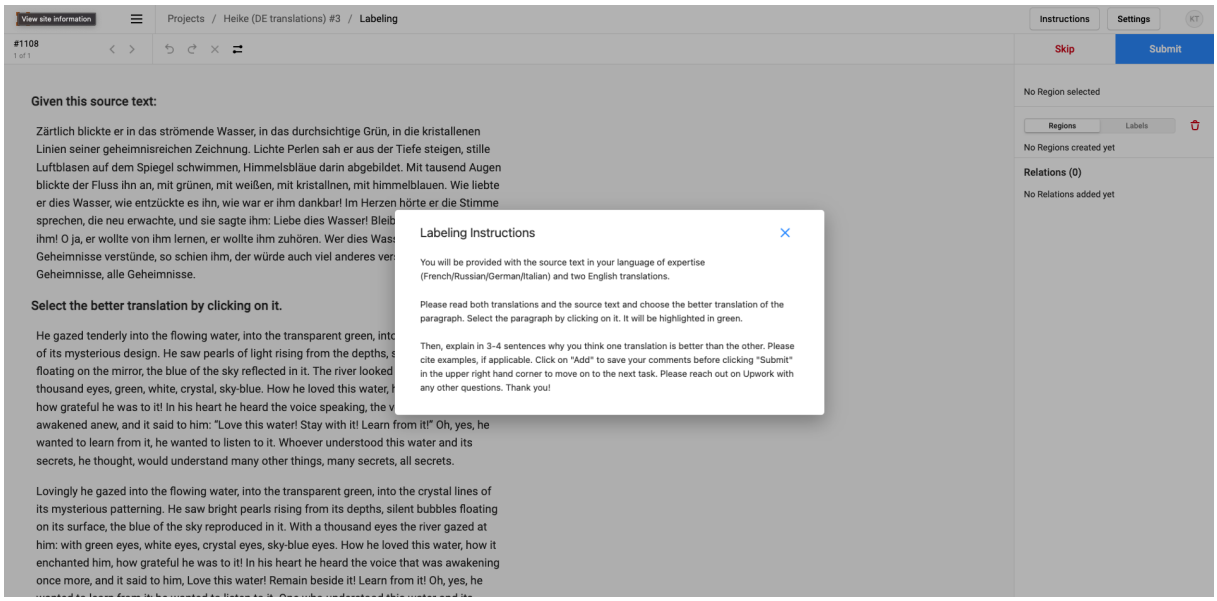

Figure 5: Example of the labeling instructions.

C GPT-3 fine-tuning configuration for post-editing:

The model was fine-tuned on OpenAI's servers for 2 epochs, with a batch size of 32, a learning rate multiplier of 0.2, and a weight of 0.1 for loss on the prompt tokens. The finetuning took 3 hours total and cost $565. Decoding on 9,648 test set examples[24] was performed using nucleus sampling (Holtzman et al., 2019) with $p = 0.2$.[25]

[24] Some test set examples exceeded *davinci*'s input limits.

[25] We performed a small-scale qualitative validation experiment on different values of p to determine this hyperparameter.

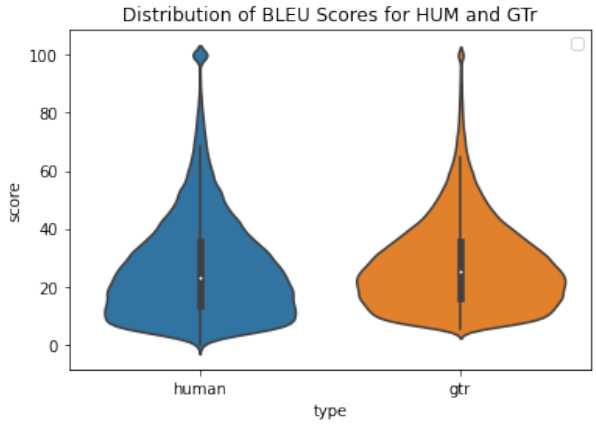

Figure 6: Distribution of BLEU scores for the HUMAN and GTr translations.

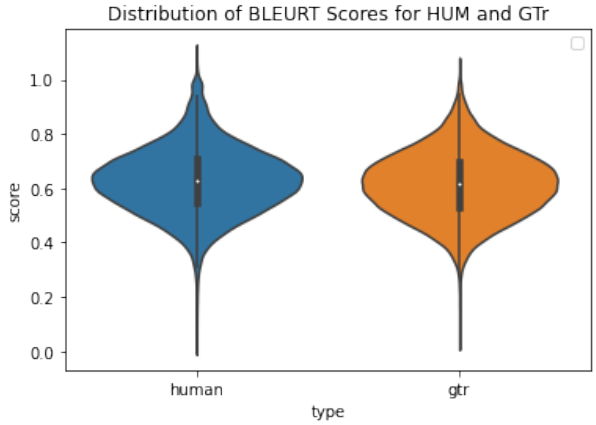

Figure 7: Distribution of BLEURT scores for the HUMAN and GTr translations.

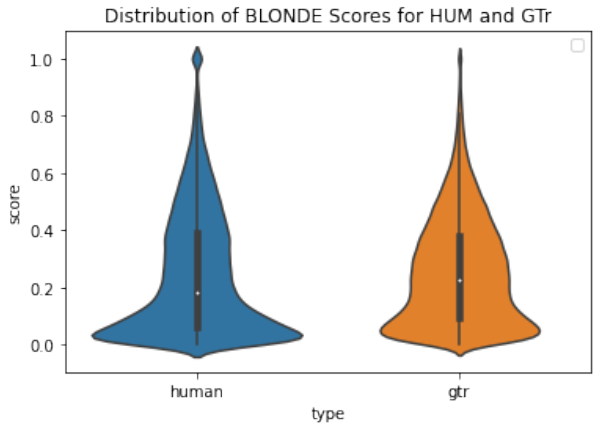

Figure 8: Distribution of BLONDE scores for the HUMAN and GTr translations.

Title	Author	Gender	Split	Source Lang	Pub Year	# Trans
A Confession	Leo Tolstoy	M	test	*ru*	1882	2
Botchan	Natsume Soseki	M	test	*ja*	1906	2
Doctor Glass	Hjalmar Soderberg	M	test	*sv*	1905	2
Dom Casmurro	Machado De Assis	M	test	*pt*	1899	2
The Castle	Franz Kafka	M	test	*de*	1924	4
Chaka	Thomas Mofolo	M	test	*st*	1948	2
Envy	Yury Olesha	M	test	*ru*	1927	2
Fairytales Part 1	Dahans Christian Andersen	M	test	*da*	1875	2-3
Gora	Rabindranath Tagore	M	test	*bn*	1941	2
Journey By Moonlight	Antal Szerb	M	test	*hu*	1937	2
Kokoro	Natsume Soseki	M	test	*ja*	1914	2
Romance Of The Three Kingdoms 1	Luo Guanzhong	M	test	*zh*	1399	2
Romance Of The Three Kingdoms 2	Luo Guanzhong	M	test	*zh*	1399	2
The Adventures Of Captain Hatteras	Jules Verne	M	test	*fr*	1866	2
The Gentleman From San Francisco	Ivan Bunin	M	test	*ru*	1915	3
The Little Prince	Antoine De Saint-Exupery	M	test	*fr*	1943	2
The Magic Mountain	Thomas Mann	M	test	*de*	1924	2
The Trial	Franz Kafka	M	test	*de*	1925	4
War With The Newts	Karel Capek	M	test	*cs*	1936	2
We	Yevgeny Zamyatin	M	test	*ru*	1920	5
The Sorrows of Young Werther	Johann Wolfgang Von Goethe	M	test	*de*	1774	2
A Hero Of Our Time	Mikhail Lermontov	M	train	*ru*	1840	2
A Raw Youth	Fyodor Dostoevsky	M	train	*ru*	1875	2
Against The Grain	Joris Karl Huysmans	M	train	*fr*	1884	2
Amerika	Franz Kafka	M	train	*de*	1927	2
Anna Karenina	Leo Tolstoy	M	train	*ru*	1878	2
Around The World In Eighty Days	Jules Verne	M	train	*fr*	1873	2
Beware Of Pity	Stefan Zweig	M	train	*de*	1939	2
Brothers Karamazov	Fyodor Dostoevsky	M	train	*ru*	1879	3
Buddenbrooks	Thomas Mann	M	train	*de*	1901	2
Call To Arms	Lu Xun	M	train	*zh*	1923	2
Crime And Punishment	Fyodor Dostoevsky	M	train	*ru*	1866	3
Dead Souls	Nikolai Gogol	M	train	*ru*	1842	4
Death In Venice	Thomas Mann	M	train	*de*	1912	3
Demons	Fyodor Dostoevsky	M	train	*ru*	1871	2
Don Quixote	Miguel De Cervantes	M	train	*es*	1605	2
Elective Affinities	Johann Wolfgang Von Goethe	M	train	*de*	1809	2
Fairytales Part 2	Dahans Christian Andersen	M	train	*da*	1875	2-3
Fathers And Sons	Ivan Turgenev	M	train	*ru*	1862	3
Gargantua And Pantagruel	François Rabelais	M	train	*fr*	1532	2
Germinal	Emile Zola	M	train	*fr*	1885	2
Heidi	Johanna Spyri	F	train	*de*	1881	3
Hesitation	Lu Xun	M	train	*zh*	1926	2
Home Of The Gentry	Ivan Turgenev	M	train	*ru*	1859	2
In A Grove	Ryunosuke Akutagawa	M	train	*ja*	1922	2
In The Shadow Of Young Girls In Flower	Marcel Proust	M	train	*fr*	1918	2
Jacques The Fatalist	Denis Diderot	M	train	*fr*	1796	2
Kallocain	Karin Boye	F	train	*sv*	1940	2
Kappa	Ryunosuke Akutagawa	M	train	*ja*	1927	2
Kristin Lavransdatter 1 The Wreath	Sigrid Undset	F	train	*nb*	1920	2
Kristin Lavransdatter 2 The Wife	Sigrid Undset	F	train	*nb*	1920	2
Lassommoir	Emile Zola	M	train	*fr*	1877	2
Les Miserables	Victor Hugo	M	train	*fr*	1862	3
Manon Lescaut	Antoine François Prevost	M	train	*fr*	1731	2
Nana	Emile Zola	M	train	*fr*	1880	3
No Longer Human	Osamu Dazai	M	train	*ja*	1948	2
Notes From Underground	Fyodor Dostoevsky	M	train	*ru*	1864	3

Title	Author	Gender	Split	Source Lang	Pub Year	# Trans
Oblomov	Ivan Goncharov	M	train	*ru*	1859	3
Petersburg	Andrei Bely	M	train	*ru*	1913	3
Pinocchio	Carlo Collodi	M	train	*it*	1883	2
Poor Folk	Fyodor Dostoevsky	M	train	*ru*	1846	3
Rashomon	Ryunosuke Akutagawa	M	train	*ja*	1915	3
Song Of The Little Road	Bibhutibhushan Bandyopadhyay	M	train	*bn*	1950	2
Steppenwolf	Hermann Hesse	M	train	*de*	1927	2
Strange Tales From A Chinese Studio	Pu Songling	M	train	*zh*	1740	2
Swanns Way	Marcel Proust	M	train	*fr*	1913	2
The Blind Owl	Sadegh Hedayat	M	train	*fa*	1937	2
The Book Of Disquietude	Fernando Pessoa	M	train	*pt*	1982	2
The Count Of Monte Cristo	Alexandre Dumas	M	train	*fr*	1844	2
The Dancing Girl Of Izu	Yasunari Kawabata	M	train	*ja*	1926	2
The Death Of Ivan Ilyich	Leo Tolstoy	M	train	*ru*	1886	3
The Debacle	Emile Zola	M	train	*fr*	1892	2
The Diary Of A Young Girl	Anne Frank	F	train	*nl*	1947	2
The Fortune Of The Rougons	Emile Zola	M	train	*fr*	1871	2
The Good Soldier Schweik 1 Behind The Lines	Jaroslav Hasek	M	train	*cs*	1921	2
The Good Soldier Schweik 2 At The Front	Jaroslav Hasek	M	train	*cs*	1922	2
The Good Soldier Schweik 3 The Glorious Licking	Jaroslav Hasek	M	train	*cs*	1922	2
The Hunchback Of Notre Dame	Victor Hugo	M	train	*fr*	1833	2
The Journey To The West	Wu Cheng-En	M	train	*zh*	1592	4
The Kill	Emile Zola	M	train	*fr*	1871	2
The Kreutzer Sonata	Leo Tolstoy	M	train	*ru*	1889	2
The Manuscript Found In Saragossa	Jan Potocki	M	train	*pl*	1805	2
The Master And Margarita	Mikhail Bulgakov	M	train	*ru*	1966	2
The Mate Mattia Pascal	Luigi Pirandello	M	train	*it*	1904	2
The Metamorphosis	Franz Kafka	M	train	*de*	1915	3
The Notebooks Of Malte Laurids Brigge	Rainer Maria Rilke	M	train	*de*	1910	3
The Nun	Denis Diderot	M	train	*fr*	1780	2
The Phantom Of The Opera	Gaston Leroux	M	train	*fr*	1909	3
The Posthumous Memoirs Of Bras Cubas	Joaquim Maria Machado De Assis	M	train	*pt*	1881	2
The Queen Of Spades	Alexander Pushkin	M	train	*ru*	1834	2
The Red And The Black	Stendhal	M	train	*fr*	1830	2
The Story Of Gosta Berling	Selma Agerlof	F	train	*sv*	1891	2
The Three Musketeers	Alexandre Dumas	M	train	*fr*	1844	2
Twenty Thousand Leagues Under The Sea	Jules Verne	M	train	*fr*	1869	3
Venus In Furs	Leopold Von Sacher-Masoch	M	train	*de*	1870	2
War And Peace	Leo Tolstoy	M	train	*ru*	1865	2
Wild Geese	Mori Ogai	M	train	*ja*	1911	2
Voyage Around My Room	Xavier De Maistre	M	valid	*fr*	1794	2
Bel Ami	Guy De Maupassant	M	valid	*fr*	1885	2
Candide	Voltaire	M	valid	*fr*	1759	2
Dream Of The Red Chamber	Cao Xueqin	M	valid	*zh*	1791	2
Dream Story	Arthur Schnitzler	M	valid	*de*	1926	2
Kusamakura	Natsume Soseki	M	valid	*ja*	1906	2
Madame Bovary	Gustave Flaubert	M	valid	*fr*	1856	2
Ponniyin Selvan 1 The First Floods	Kalki Krishnamurthy	M	valid	*ta*	1950	3
Siddhartha An Indian Tale	Hermann Hesse	M	valid	*de*	1922	2
The Alienist	Machado De Assis	M	valid	*pt*	1881	2
The Captains Daughter	Aleksandr Pushkin	M	valid	*ru*	1836	2
The Idiot	Fyodor Dostoevsky	M	valid	*ru*	1868	4
The Sin Of Abbe Mouret	Emile Zola	M	valid	*fr*	1875	2
The Twelve Chairs	Ilya Ilf And Yevgeny Petrov	M	valid	*ru*	1947	2

Table 5: A full list of the literary texts from which the source paragraphs in PAR3 are sampled with author name, author gender, publication year, source language, and test/train/val split designations.

	HUM > GTr	GPT-3 > GTr	HUM > GPT-3
French	86.0%*	66.0%*	64.0%*
Russian	80.0%*	68.0%*	64.0%*
German	86.0%*	78.0%***	62.0%*

Table 6: The preference of translators for all three evaluations: (1) HUM vs GTr, (2) GPT-3 vs GTr, (3) HUM vs GPT-3, divided by the language of translator. The translators unanimously preferred (1) HUM, (2) GPT-3, and (3) HUM. The significance is reported for the following levels *p<.05, **p<.01, ***p<.001 (binomial test).

Edit-Type	Example
Corrected Mistranslation	GTr: First put me in a coffin and bury me in the ground, **then give me my daughter**. GPT-3: Put me in the grave first and bury me in the ground, **then give my daughter away**. GTr: I have only one thing left: to tell you everything, my unknown readers (now you are **so for me the same roads**, and close, and inaccessible - as he was at that moment). GPT-3: I have only one thing left: to tell you everything, my unknown readers (now you are **as dear to me**, as close and as unattainable as he was at that moment).
Corrected Pronouns	GTr: I tasted in **his** arms the delights of paradise, which produced those torments of hell with which you see me devoured; (...). GPT-3: I tasted the delights of paradise in **her** arms, which produced the torments of hell that you see me devoured by. GTr: **It** calmed **her** down and her heart stopped beating. GPT-3: This calmed **me** down, **my** heart stopped pounding.
Corrected Named Entities	GTr: (...) by Lebedev's nephew, known to readers, and secondly, by **Ippolit. Hippolyte** was a very young man, about seventeen, maybe eighteen, with an intelligent, but constantly irritated expression on his face, on which illness left terrible marks. GPT-3: (...) by Lebedev's nephew, a young man known to the reader, and secondly, by **Ippolit. Ippolit** was a very young man, about seventeen, or even eighteen, with an intelligent but constantly irritated expression on his face, on which illness had left terrible traces.
Less Literal / Better Wording	GTr: **To amuse himself, he employed himself at home as a laborer**, and he even tried to paint the attic with **a residue of color** that the painters had left behind. GPT-3: **To distract himself, he did odd jobs around the house**, and even tried to paint the attic with **some paint** that the painters had left over.

Table 7: Examples of the GPT-3-post-edit (GPT-3) model correcting translations by Google Translate (GTr)

Table 8: Examples of translators' techniques from PAR3. Catastrophic GTr mistakes were highlighted in red.

Technique	Description	Example
Borrowing	Words or phrases left untranslated to introduce the *flavor* of source language culture.	**SRC**:– Пиво есть? – сиплым голосом осведомился Бездомный.(*ru*) **HUM**: "Got any beer?" inquired **Bezdomny** in a hoarse voice. **GTr**: - Do you have beer? **Homeless** inquired in a hoarse voice. (from *Master and Margarita*)
Established Equivalence	An equivalent of the source language using different stylistic and structural methods. This technique is applied frequently to idioms, clichés, simile, and proverbs.	**SRC**: 中学と師範とはとこの県下でも犬と猿のように仲がわるいそうだ。 (*ja*) **HUM**: The middle school and the normal, I understood, are as much **friendly as dogs and monkeys**. **GTr**: It seems that junior high school and **instructors get along with each other like dogs and monkeys** in any prefecture. (from *Botchan*)
Transposition	A change in grammatical category, such like word class, number, tense, etc.	SRC: Et il reprit son carnet, biffant **avec le plus grand soin** les sommes qu'il venait de payer. (*fr*) HUM: And he took up his notebook, **carefully** crossing out the amounts he had just paid. GTr: And he went back to his notebook, crossing out **with the greatest care** the sums he had just paid. (from *The Count of Monte Cristo*)
Modulation	A shift in point of view, focus, cognitive category.	**SRC**: Bei der Schnelligkeit ihres Wesens war ihr **nicht leicht** zu widersprechen. (*de*) **HUM**: Being so quick in her manner she was **hard** to contradict. **GTr**: Given the quickness of her nature, it was **not easy** to contradict her. (from *Elective Affinities*)
Addition	An addition of a new piece of information, which is not easily inferable from the source language.	**SRC**: 清が物をくれる時には必ずおやじも兄も居ない時に限る。 (*ja*) **HUM**: When Kiyo gave me these presents she would always **be careful to choose** times when the old man and my brother were not around. **GTr**: When **Qing** gives me something, **I** always do it only when my father and brother are not there. (from *Botchan*)
Omission	An omission of information present in the source language to the extent that it is not even easily inferable in the target language.	**SRC**: 健全なる男女の河童よ (*ja*) **HUM**: IF YOU ARE HEALTHY ___ KAPPAS **GTr**: Healthy **male and female** kappa (from *Kappa*)
Generalization	A word or phrase translated into a more general one (hypernym).	**SRC**: 妹子是被大哥吃了，母知道没有，我可不得而知。 (*zh*) **HUM**: My sister was eaten by my **brother**, but I don't know whether Mother realized it or not. **GTr**: The sister was eaten by **the elder brother**, and whether the mother knew it or not, I don't know. (from *Call to Arms*)
Particularization	A word or phrase is translated into a more precise or concrete term (hyponym).	**SRC**: (...) Andrea saisit la main du comte, la serra, **sauta** dans son phaéton et disparut. (*fr*) **HUM**: (...) Andrea seized his hand, pressed it, **leapt** into his phaeton and rode off. **GTr**: (...) Andrea seized the count's hand, squeezed it, **jumped** into his phaeton and disappeared. (from *The Count of Monte Cristo*)
Adaptation	Content is adapted to the target culture. It may include adapting the portrayed situation so that it is appropriate for the target culture (cultural substitution).	**SRC**: 父はこの前の冬に帰って来た時ほど将棋を差したがらなくなった。 (*ja*) **HUM**: My father did not show as much interest in **chess** as he had done the previous winter. **GTr**: My dad was less reluctant to play **shogi** than when he came back last winter. (from *Kokoro*)
Description	A term or expression from the source language is described in text in the translation.	**SRC**: 或時先生が例の通りさっさと海から上がって来て、いつもの場所に脱ぎ棄ててた浴衣を着ようとすると、どうした訳か、その浴衣に砂がいっぱい着いていた。 (*ja*) **HUM**: One day, however, after his usual swim, Sensei was about to put on his **summer dress** which he had left on the bench, when he noticed that the dress, for some reason, was covered with sand. **GTr**: At one point, as usual, the teacher came up from the sea and tried to put on the **yukata** that had been taken off and thrown away at the usual place, but for some reason, the yukata was full of sand. (from *Kokoro*)
Sentence Diffusion	The source sentence is being translated into two or more sentences in the translation.	**SRC**: Prodal jsem tě, kamaráde, hanebně prodal. (*cs*) **HUM**: I've sold you, buddy. Shamefully sold you. **GTr**: I sold you, my friend, I shamefully sold you. (from *The Good Soldier Schweik*)
Sentence Merging	Two or more sentences from the source language are combined together into one sentence in the translation.	**SRC**: 三年前の夏のことです。僕は人並みにリュック・サックを背負い、あの上高地の温泉宿から穂高山へ登ろうとしました。 (*ja*) **HUM**: One summer morning three years ago, I left an inn at Kamikōchi hot spring to climb Mt. Hodaka, with a rucksack on my back. **GTr**: It was the summer three years ago. I carried a rucksack on my back like a person and tried to climb Mt. Hotaka from that hot spring inn in Kamikochi. (from *Kappa*)
Reordering	Information is moved from one place in the paragraph to another for better coherence in the target language.	SRC: 私が先生と知り合いになったのは鎌倉である。 (*ja*) HUM: It was at Kamakura, **during the summer holidays**, that I first met Sensei. GTr: It was Kamakura that I got to know the teacher. (from *Kokoro*)

Source lang	BLEU			BLEURT			BLONDE		
	Hum	GPT-3	GTr	Hum	GPT-3	GTr	Hum	GPT-3	GTr
fr	20.0	27.2	26.1	0.641	0.681	0.658	24.7	27.7	29.3
ru	46.0	38.2	36.8	0.636	0.631	0.612	30.1	24.5	24.3
de	19.8	22.2	19.0	0.525	0.552	0.530	18.0	21.1	18.6
ja	11.4	9.5	6.9	0.545	0.514	0.457	12.7	11.1	8.5
zh	2.4	4.6	3.6	0.324	0.351	0.310	3.2	4.3	3.7
cs	19.4	22.7	19.1	0.625	0.621	0.590	18.3	22.0	19.7
pt	28.9	32.4	25.3	0.643	0.636	0.590	28.9	30.8	25.8
sv	28.1	33.8	29.2	0.649	0.673	0.538	27.2	33.5	31.7
hu	22.3	25.1	16.9	0.613	0.628	0.581	22.2	22.3	16.0
All	21.2	23.3	20.6	0.564	0.580	0.549	20.0	21.0	19.6
Win %*	28.5%	49.5%	22.0%	30.9%	52.1%	17.0%	30.5%	40.5%	29.0%

Table 9: The percentage of cases in which the automatic MT metric ranks the human, GPT-3, or Google translations above the other two. *Note: There are 9,648 unique source paragraphs that were input to the post-editing model, but we exclude ties in the calculation of Win %. The total number of ties was 340, 94, and 33, for BLEU, BLEURT, and BLONDE respectively.

Source Lang	PRISM		PRISM-QE		MOVERSCORE	
	Hum	GTr	Hum	GTr	Hum	GTr
fr	-2.3329	-2.1711	-2.1812	-1.0883	0.5976	0.5985
ru	-2.2142	-2.1532	-2.1472	-1.2995	0.6109	0.5997
de	-2.5624	-2.3874	-2.4816	-1.5152	0.5912	0.5922
ja	-3.0987	-3.2028	-3.2498	-2.0923	0.5468	0.5281
zh	-4.3927	-4.2472	-4.3900	-3.3711	0.5191	0.5211
cs	-3.0720	-2.5455	-2.5142	-1.3088	0.5515	0.5704
pt	-2.8693	-2.4973	-2.4732	-1.0264	0.5805	0.5827
no	-2.3435	-2.2936	-2.3826	-1.1298	0.5938	0.5897
sv	-1.7067	-1.5924	-1.6552	-1.0648	0.6443	0.6408
it	-2.1974	-2.1698	-2.1216	-1.0742	0.5869	0.5894
es	-2.1496	-2.2906	-2.2182	-1.1592	0.6170	0.5875
fa	-2.9812	-2.9559	-4.3144	-4.0303	0.5735	0.5596
hu	-2.3005	-2.3425	-2.3417	-1.3059	0.6008	0.5701
nl	-2.3712	-2.1936	-2.3491	-1.0664	0.6010	0.6074
pl	-2.0920	-2.5984	-2.6809	-1.3299	0.6219	0.5685
ta	-3.6783	-3.6200	-4.5426	-4.3912	0.5341	0.5361
All	-2.4207	-2.2985	-2.3290	-1.3275	0.5966	0.5928
Win %*	34.61%	65.39%	3.41%	96.59%	45.44%	54.56%

Table 10: Results of PRISM, PRISM-QE and MOVERSCORE on PAR3 . Higher score is better for all metrics. Scores were calculated on the entirety of version one of the PAR3 dataset across its 107,467 unique source paragraphs. Again, we exclude ties from the calculation of Human Win %. The total number of ties was 80, 82, and 100 for PRISM, PRISM-QE (Thompson and Post, 2020), and MOVERSCORE (Zhao et al., 2019), respectively.

Metrics	**Kendall Tau**
BLEU	0.209***
BLONDE	0.120***
BLEURT	0.262***

Table 11: Metrics correlation with human evaluation. Significant correlation at ***p<.001

	Type	Wilcoxon-Pratt Signed-Rank Test	Effect Size*
BLEU	HUM vs GTr	$z = 4.093, p < .001$	0.236
	GPT-3 vs GTr	$z = -7.256, p < .001$	0.419
	HUM vs GPT-3	$z = -1.888, p = .059$	0.109
BLONDE	HUM vs GTr	$z = 1.423, p = .155$	0.082
	GPT-3 vs GTr	$z = -5.127, p < .001$	0.296
	HUM vs GPT-3	$z = -3.027, p = .003$	0.175
BLEURT	HUM vs GTr	$z = 7.0612, p < .001$	0.408
	GPT-3 vs GTr	$z = -7.553, p < .001$	0.436
	HUM vs GPT-3	$z = 1.827, p = .068$	0.105

Table 12: Results of the performance of automatic metrics on the 150 paragraphs used in human evaluation. (*The common interpretation of the effect size is the following: 0.10-<0.30 (small), 0.30-<0.50 (moderate), >=0.50 (large))

SRC: Joachim ging, und es kam die »Mittagssuppe«: ein einfältig symbolischer Name für das, was kam! Denn Hans Castorp war nicht auf Krankenkost gesetzt, – warum auch hätte man ihn darauf setzen sollen? Krankenkost, schmale Kost war auf keine Art indiziert bei seinem Zustande. Er lag hier und zahlte den vollen Preis, und was man ihm bringt in der stehenden Ewigkeit dieser Stunde, das ist keine »Mittagssuppe«, es ist das sechsgängige Berghof-Diner ohne Abzug und in aller Ausführlichkeit, – am Alltage üppig, am Sonntage ein Gala-, Lust- und Parademahl, von einem europäisch erzogenen Chef in der Luxushotelküche der Anstalt bereitet. Die Saaltochter, deren Amt es war, die Bettlägrigen zu versorgen, brachte es ihm unter vernickelten Hohldeckeln und in leckeren Tiegeln; sie schob den Krankentisch, der sich eingefunden, dies einbeinige Wunder von Gleichgewichtskonstruktion, quer über sein Bett vor ihn hin, und Hans Castorp tafelte daran wie der Sohn des Schneiders am Tischlein deck dich.

GTr: Joachim went, and "Lunchtime Soup" came: a simple symbolic name for what was coming! Because Hans Castorp was not put on sick food - why should he have been put on it? Sick diet, small fare, was in no way indicated in his condition. He lay here and paid the full price, and what is brought to him in the standing eternity of this hour is not a "lunchtime soup," it is the six-course Berghof dinner without deduction and in great detail - sumptuous in everyday life, closed on Sundays Gala, pleasure and parade meal, prepared by a European-educated chef in the luxury hotel kitchen of the institution. The maid, whose job it was to look after the bedridden, brought it to him under nickel-plated hollow lids and in delicious jars; She pushed the patient's table that appeared, this one-legged marvel of balanced construction, across his bed in front of him, and Hans Castorp ate at it like the tailor's son at the little table, cover yourself.

HUM: Joachim would leave, and the “midday soup” would arrive—soup was the simplified, symbolic name for what came. Because Hans Castorp was not on a restricted diet—why should he have been? A restricted diet, short commons, would hardly have been appropriate to his condition. There he lay, paying full price, and what they brought him at this hour of fixed eternity was “midday soup,” the six-course Berghof dinner in all its splendor, with nothing missing—a hearty meal six days a week, a sumptuous showpiece, a gala banquet, prepared by a trained European chef in the sanatorium’s deluxe hotel kitchen. The dining attendant whose job it was to care for bedridden patients would bring it to him, a series of tasty dishes arranged under domed nickel covers. She would shove over the bed table, which was now part of the furniture, a marvel of one-legged equilibrium, adjust it across his bed in front of him, and Hans Castorp would dine from it like the tailor’s son who dined from a magic table.

Table 13: An example SRC from Thomas Mann’s *The Magic Mountain* that was administered as an A/B test with its corresponding GTr and HUM. Though all monolingual raters chose HUM, the translator chose GTr.

SRC: Еще вначале, как только князь вошел в гостиную, он сел как можно дальше от китайской вазы, которою так напугала его Аглая. Можно ли поверить, что после вчерашних слов Аглаи в него вселилось какое-то неизгладимое убеждение, какое-то удивительное и невозможное предчувствие, что он непременно и завтра же разобьет эту вазу, как бы ни сторонился от нее, как бы ни избегал беды? Но это было так. В продолжение вечера другие сильные, но светлые впечатления стали наплывать в его душу; мы уже говорили об этом. Он забыл свое предчувствие. Когда он услышал о Павлищеве и Иван Федорович подвел и показал его снова Ивану Петровичу, он пересел ближе к столу и прямо попал на кресло подле огромной, прекрасной китайской вазы, стоявшей на пьедестале, почти рядом с его локтем, чуть-чуть позади.

GTr: Even at the beginning, as soon as the prince entered the drawing room, he sat down as far as possible from the Chinese vase, with which Aglaya had so frightened him. Is it possible to believe that after Aglaya's words yesterday, some indelible conviction came into him, some amazing and impossible premonition that he would certainly break this vase tomorrow, no matter how he avoided it, no matter how he avoided trouble? But it was. In the course of the evening other strong but bright impressions began to flood into his soul; we already talked about this. He forgot his premonition. When he heard about Pavlishchev and Ivan Fyodorovich let him down and showed him again to Ivan Petrovich, he moved closer to the table and fell straight into an armchair beside a huge, beautiful Chinese vase, which stood on a pedestal, almost next to his elbow, a little behind.

HUM: From the very beginning, as soon as the prince entered the drawing room, he sat down as far as possible from the Chinese vase, with which Aglaya had frightened him so. Can one possibly believe that, after Aglaya's words the day before, some sort of indelible conviction settled in him, some sort of astonishing and impossible premonition that the next day he would unfailingly break that vase, however far away he kept from it, however much he avoided the disaster? But it was so. In the course of the evening other strong but bright impressions began to flow into his soul; we have already spoken of that. He forgot his premonition. When he heard about Pavlishchev, and Ivan Fyodorovich brought him and introduced him again to Ivan Petrovich, he moved closer to the table and ended up right in the armchair next to the enormous, beautiful Chinese vase, which stood on a pedestal almost at his elbow, slightly behind him.

Table 14: An example SRC from Fyodor Dostoevsky’s *The Idiot* that was administered as an A/B test with its corresponding GTr and HUM. Though all monolingual raters chose HUM, the translator chose GTr.

SRC: Князь, однако же, слышал, как его назвали идиотом, и вздрогнул, но не оттого, что его назвали идиотом. «Идиота» он тотчас забыл. Но в толпе, недалеко от того места, где он сидел, откуда-то сбоку — он бы никак не указал, в каком именно месте и в какой точке, — мелькнуло одно лицо, бледное лицо, с курчавыми темными волосами, с знакомыми, очень знакомыми улыбкой и взглядом, — мелькнуло и исчезло. Очень могло быть, что это только вообразилось ему; от всего видения остались у него в впечатлении кривая улыбка, глаза и светло-зеленый франтовской шейный галстук, бывший на промелькнувшем господине. Исчез ли этот господин в толпе или прошмыгнул в вокзал, князь тоже не мог бы определить.

GTr: The prince, however, heard how he was called an idiot, and shuddered, but not because he was called an idiot. "Idiot" he immediately forgot. But in the crowd, not far from the place where he was sitting, from somewhere on the side - he would not have indicated exactly in what place and at what point - one face flashed, a pale face, with curly dark hair, with acquaintances, very familiar smile and look, flashed and disappeared. It could very well be that it was only his imagination; from the whole vision, he was impressed by the crooked smile, eyes and light green dandy neck tie that the gentleman flashed by. Whether this gentleman disappeared into the crowd or slipped into the station, the prince could not determine either.

GPT-3: The prince, however, heard how he was called an idiot, and he shuddered, but not because he was called an idiot. “Idiot” he immediately forgot. But in the crowd, not far from the place where he was sitting, from somewhere on the side—he would not have been able to indicate exactly where and in what place—a face flashed, a pale face, with curly dark hair, with a familiar, very familiar smile and gaze, flashed and disappeared. It could very well have been that it was only his imagination; from the whole vision he retained an impression of a crooked smile, eyes, and a light green necktie of the dandy who had flashed by. Whether this dandy disappeared into the crowd or slipped into the station, the prince would also not have been able to say.

Table 15: An example SRC from Fyodor Dostoevsky’s *The Idiot* that was administered as an A/B test with its corresponding GTr and GPT-3. The translator preferred GPT-3.