Fine-tuning and evaluation of NMT models for literary texts using RomCro v.2.0

Bojana Mikelenić
Faculty of Humanities
and Social Sciences,
University of Zagreb
bmikelen@ffzg.unizg.hr

Antoni Oliver
Universitat Oberta
de Catalunya
aoliverg@uoc.edu

Sergi Àlvarez Vidal
Universitat Autònoma
de Barcelona
sergi.alvarez@uab.cat

Abstract

This paper explores the fine-tuning and evaluation of neural machine translation (NMT) models for literary texts using RomCro v.2.0, an expanded multilingual and multidirectional parallel corpus. RomCro v.2.0 is based on RomCro v.1.0, but includes additional literary works in five Romance languages (Spanish, French, Italian, Portuguese, Romanian) and Croatian, as well as texts in Catalan, making it a valuable resource for improving MT in underrepresented language pairs. Given the challenges of literary translation, where style, narrative voice, and cultural nuances must be preserved, fine-tuning on high-quality domain-specific data is essential for enhancing MT performance.

We fine-tune existing NMT models with RomCro v.2.0 and evaluate their performance for six different language combinations using automatic metrics and for Spanish-Croatian and French-Catalan using manual evaluation. Results indicate that fine-tuned models outperform general-purpose systems, achieving greater fluency and stylistic coherence. These findings support the effectiveness of corpus-driven fine-tuning for literary translation and highlight the importance of curated high-quality corpora.

1 Introduction

Parallel multilingual corpora play a crucial role in linguistic research and computational applications, serving as foundational resources for a broad range of disciplines. In linguistics, they enable contrastive studies, lexicographic analysis, and phraseology research (Lefer, 2021), offering insights into language structures and translation patterns across multiple languages. In translation studies, they are used to examine translation strategies, detect shifts in meaning, and provide empirical evidence for translation universals. Beyond theoretical applications, parallel corpora are also essential in translation training (López Rodríguez, 2016), providing students and professionals with real-world examples of translated texts that reflect both linguistic variation and different approaches to translation. Additionally, they are widely used in computational linguistics, particularly in training and evaluating machine translation (MT) systems (Koehn et al., 2007; Koehn, 2020), as well as in terminology extraction (Lefever et al., 2009) and multilingual information retrieval.

Since the effectiveness of MT models largely depends on the quality and quantity of their training data, access to well-aligned, diverse, and representative parallel corpora is crucial for improving translation performance. While large-scale datasets exist for widely spoken languages, there remains a significant gap in high-quality parallel data for specific language pairs, especially when literary texts are involved. Unlike technical or legal texts, which are often characterized by terminological consistency and rigid syntactic structures, literary texts pose unique challenges due to their stylistic complexity, cultural nuances, and need for creativity (Guerberof-Arenas and Toral, 2022). Standard MT models trained on general-purpose corpora struggle to capture these intricacies, often producing translations that fail to preserve the author's style, narrative voice, and the overall reading experience.

Given these challenges, the development of high-quality parallel corpora specifically designed for literary translation is essential for advancing MT capabilities in this domain. This paper describes one such use of an updated and improved version of the RomCro corpus, a multilingual and multidirectional parallel corpus of contemporary literary texts in Romance languages and Croatian. This 2.0 version includes more translation units and another

Proceedings of the Second Workshop on Creative-text Translation and Technology (CTT), pages 44–51, June 24, 2025

Category	RomCro v.1.0	RomCro v.2.0	Difference
Languages	6	7	1
Translation units (TU)	142,470	166,742	24,272
Original texts	27	33	6
Total texts	159	213	54
Millions of words (Mw)	15.7	19.4	3.7

Table 1: Comparison of RomCro v.1.0 and v.2.0

language, Catalan. We have used this improved RomCro corpus to fine-tune different MT models, and conducted automatic evaluations on six language combinations, as well as manual evaluations for Spanish-Croatian and French-Catalan language pairs.

2 Previous work

Despite significant advances in NMT, the automatic translation of literary texts remains one of the most challenging areas of MT research. Unlike technical or legal translation, which prioritizes accuracy and consistency, literary translation must preserve elements such as narrative voice, rhythm, metaphorical expressions, and stylistic nuances. Existing studies have shown that standard NMT systems struggle with these aspects, often failing to reproduce the richness and depth of the original text (Toral and Way, 2015).

Recent research has explored fine-tuning NMT models specifically for literary texts. Hansen and Esperanza-Rodier (2022) investigated the impact of customizing MT systems for fiction translation, demonstrating that fine-tuned models trained on smaller, high-quality datasets perform significantly better than general-purpose NMT systems. However, their study also found that even with domain adaptation, the output still required substantial human post-editing to correct stylistic inconsistencies and ensure readability. Similarly, Oliver (2023) proposed training author-specific NMT models, where an MT system is fine-tuned exclusively on the works of a single writer. This approach has shown promise in maintaining stylistic consistency, although it is highly dependent on the availability of sufficient bilingual training data. Toral et al. (2023) investigate whether it is worthwhile to build a customized MT system trained with a large quantity of in-domain training data (novels) compared to a generic MT system for a fairly well-resourced language pair, English-to-Dutch. A multidimensional evaluation shows that a literary-adapted system performs slightly better. Following the promising results shown in previous research, we use RomCro to fine-tune MT models as well as assess the quality of the resulting translations.

3 RomCro: A Multilingual Parallel Corpus of Literary Texts

RomCro is a multilingual and multidirectional parallel corpus of contemporary literary texts[1] in Romance languages and Croatian. Its first version, RomCro v.1.0 (Bikić-Carić et al., 2023), contains works in six languages: Spanish, French, Italian, Portuguese, Romanian, and Croatian. With 27 original titles and their translations, 142,470 translation units, and 15.7 million words, it stands out for its focus on high-quality literary data and multidirectional alignment, allowing each language to serve as both source and target.

The building of RomCro v.1.0 consisted of several stages, from the selection and the collection of texts, their digitization, segmentation, and alignment, with manual corrections to ensure accuracy. Regarding the selection criteria, novels were prioritized and the availability of translations in all six languages dictated the choices. At the end, out of the possible 162 texts (27 originals and their translations into the remaining five languages), all but three were obtained, making it a well-balanced corpus in terms of language distribution. The corpus is accessible in untagged TMX and TSV documents via the European Language Resource Coordination (ELRC) platform.[2] Annotations contain metadata on original language, author, and title, with segments scrambled to protect copyright.[3]

[1] Out of the 27 originals, 17 were published for the first time in this century. Titles first published in the previous century include one work from the 1910s, two from the 1930s, one from the 1950s, three from the 1980s, and three from the 1990s.

[2] https://elrc-share.eu/repository/search/?q=romcro. This is an earlier version of the corpus containing 157 texts, in other words, missing two additional texts.

[3] For a further discussion about copyright issues and how they were dealt with, see Bikić-Carić et al. (2023).

Recent efforts to expand the corpus have introduced three new titles in the languages with the fewest original works in version 1.0: two in Portuguese and one in Croatian, as well as a new language: Catalan. The latter task was carried out in two phases: the first involved incorporating existing translations into Catalan,[4] while the second added three Catalan novels along with their translations into the six existing languages. Out of now 30 titles (27 from the v.1.0 and three new additions), 17 are available in Catalan, and the three Catalan originals have been obtained in all the remaining languages, resulting in a total of 54 texts added to the corpus.[5] This not only broadens the scope of the corpus but also addresses the scarcity of literary parallel corpora for Catalan. Table 1 shows a comparison of the two versions, namely the augmentation by more than 24,000 translation units and 3.7 million words. The updated version, RomCro v.2.0, is available in Sketch Engine (Kilgarriff et al., 2014), while the untagged TSV and TMX are hosted in the HR-CLARIN repository,[6] ensuring its availability for broader linguistic and computational research.[7]

4 Training and Evaluating NMT Systems Using RomCro

In a prior study (Mikelenić and Oliver, 2024), we explored the use of this corpus in training NMT systems tailored to literature, and we summarize the design and results of this first experiment in the next subsection. Building on this foundation, we proceed to describe the current experiment.

4.1 The first experiment: combining RomCro with large parallel corpora

RomCro's potential in training NMT systems tailored to literary texts was demonstrated in Mikelenić and Oliver (2024), where the experiment was completed for five language pairs: from Spanish into French, Italian, Portuguese, Romanian, and Croatian. Five baseline and five tailored to literature systems (using the literary data from RomCro) were trained from scratch in the following manner. RomCro v.1.0 was combined with larger, freely available parallel corpora, such as CCMatrix (sch) and MultiCCAligned (El-Kishky et al., 2020), to create a sufficiently large training dataset. These large corpora were rescored using a confidence-based filtering tool and, for most language pairs, a subset of the rescored data most similar to RomCro was selected to enhance relevance. The training process used Marian (Junczys-Dowmunt et al., 2018) to train general (or baseline) and tailored to literature systems for all five language pairs. Standard metrics, including BLEU (Papineni et al., 2002), chrF2 (Popović, 2015), and TER (Snover et al., 2006), were used to compare the tailored systems with baseline models and Google Translate.

Results showed that the tailored systems outperformed the generic Marian systems and achieved comparable or superior results to Google Translate. However, the Spanish-Croatian system underperformed, probably due to limited training data. Building on these findings, we have used RomCro v.2.0 for the new experiment described in the following subsection.

4.2 Fine-tuning existing models using RomCro

In the current experiment, we were interested in: improving the results for the language pair Spanish-Croatian, making use of RomCro v.2.0 to begin training and evaluating systems for Catalan, and adding human evaluation. We opted for the following language pairs in both directions: Spanish↔Croatian (es↔hr), French↔Croatian (fr↔hr) for control, and French↔Catalan (fr↔ca). The third pair was chosen considering that the results might be more insightful compared to combining Catalan with Spanish, given their similarity and the already demonstrated strong performance of this pair in similar tasks. Since the first experiment where corpora were trained from scratch was not very successful for Spanish-Croatian, and not

[4]This first phase was presented as a poster at the CLARIN Annual Conference 2024, with an extended abstract is available in the Conference Proceedings: https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf .

[5]The new titles in Portuguese and Croatian are: Lídia Jorge – *O vale da paixão* (in 6 languages, with Catalan missing), Afonso Cruz – *Os livros que devoraram o meu pai* (in 6 languages, with Romanian missing) and Miroslav Krleža – *Povratak Filipa Latinovicza* (in 5 languages, with Portuguese and Catalan missing). The originals in Catalan are: Jaume Cabré – *Les veus del Pamano*, Albert Sánchez Piñol – *La pell freda* and Mercè Rodoreda – *La Plaça del Diamant*.

[6]https://repository.clarin.hr/items/fe77001c-0e97-4b58-8031-505bfaa45352

[7]This paper centers on the application of RomCro v.2.0 in MT, and therefore presents only an overview of the corpus, limited to aspects we considered important for the task at hand and the objectives of this study. We will detail the building of RomCro v.2.0 in a separate paper.

enough Croatian data was added to change that, we opted for fine-tuning the existing models instead. To keep the experiment consistent, the same was done for the language pair not including Croatian, French↔Catalan.

4.2.1 Training

The training was completed by fine-tuning the Opus-MT models[8] and the multilingual NLLB200 600M distilled model (No Language Left Behind)[9] with RomCro v.2.0. To perform the fine-tuning, we have used the algorithms published by the MTUOC project[10]. We have split the RomCro corpus for each language pair into validation (5,000 segments), evaluation (1,000 segments) and the remaining segments for training.

4.2.2 Automatic evaluation

In Table 2, we present the evaluation results for all the reference and fine-tuned models using three automatic metrics implemented in Sacrebleu[11] (Post, 2018): BLEU, chrF2, and TER. The appendices provide the metric signatures, detailing the exact configuration parameters as reported by Sacrebleu. The best results are highlighted in bold in the table. If multiple results are highlighted, it indicates that the differences are not statistically significant. The OpusMT model has been used as the baseline for comparison.

The conclusions we can draw from the table are that the fine-tuned (FT) models, OpusMT-FT and NLLB200-distilled-600M-FT, outperform the base models while, compared to each other, they perform very similarly overall. For Spanish-Croatian, OpusMT-FT achieves the best results across all metrics, with statistically significant differences observed for BLEU and TER, but not for chrF2. For the Croatian-Spanish pair, some metrics favor OpusMT-FT while others favor NLLB-FT, though none of these differences are statistically significant. In the case of French-Croatian, OpusMT-FT performs better across all metrics, but the difference in chrF2 is not statistically significant. Similarly, for Croatian-French, OpusMT-FT outperforms NLLB-FT in all metrics, except for chrF2, where they perform the same. Conversely, for French↔Catalan, NLLB-FT outperforms OpusMT-FT, though the difference in TER is not statistically significant.

4.2.3 Human evaluation

In addition to automatic evaluations, and to obtain a better insight into the results produced by the MT models fine-tuned with the new corpus, we conducted a human evaluation for the Spanish-Croatian and French-Catalan language pairs. For each language pair, we selected 100 randomized segments, with a number of words between 8 and 25. These short segments were extracted from the test set and typically consisted of a single sentence. Although we acknowledge that sentence-level evaluation has limitations, particularly in domains where discourse-level context is important, we considered it appropriate for a preliminary comparison between model outputs. Future work will incorporate context-aware evaluation over longer spans of text.

We selected, for each language pair, the fine-tuned model that achieved the best performance in the automatic evaluation and compared its output against that of the corresponding baseline model. Specifically, we compared the outputs of the OpusMT and fine-tuned OpusMT models for Spanish–Croatian, and the NLLB-DIST-600M and its fine-tuned variant for French–Catalan.

Two experienced linguists, both native speakers of the target language and with professional experience in translation and post-editing, carried out the human evaluation. The evaluators were not informed of which model had produced which output (i.e., the identity of the MT engine was hidden to ensure blind evaluation). For each segment pair, they were asked to assess which of the two outputs represented a better translation, based on adequacy and fluency. When neither translation was clearly better, they were allowed to mark the pair as "equally good."

The evaluation followed a binary preference protocol: each annotator independently selected the better of the two translations (or marked them as equal), without assigning quality scores. Disagreements, which occurred in approximately 10% of the segments, were resolved through discussion in a follow-up meeting. No formal inter-annotator agreement score was calculated, though agreement was high in both language pairs.

As shown in Table 3, for both language pairs, the fine-tuned models produced a higher number of preferred translations. In two cases, both out-

[8] https://github.com/Helsinki-NLP/Opus-MT
[9] https://ai.meta.com/research/no-language-left-behind/es-es/
[10] https://github.com/mtuoc/MTUOC-finetune-OpusMT and https://github.com/mtuoc/MTUOC-finetune-NLLB
[11] https://github.com/mjpost/sacrebleu

System	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
Spanish-Croatian			
Baseline: OpusMT	16.6 (16.6 ± 1.1)	43.5 (43.5 ± 1.3)	71.1 (71.1 ± 1.4)
OpusMT-FT	**22.0 (21.9 ± 1.3) (p = 0.0010)***	**49.1 (49.1 ± 1.4) (p = 0.0010)***	**65.4 (65.4 ± 1.6) (p = 0.0010)***
NLLB-dist-600M	14.6 (14.6 ± 1.3) (p = 0.0020)*	42.9 (42.9 ± 1.1) (p = 0.0689)	76.5 (76.4 ± 4.7) (p = 0.0190)*
NLLB-dist-600M-FT	20.4 (20.3 ± 1.2) (p = 0.0010)*	48.2 (48.2 ± 1.2) (p = 0.0010)*	67.0 (67.0 ± 1.4) (p = 0.0010)*
eval.es-GoogleT.hr	20.6 (20.6 ± 1.1) (p = 0.0010)*	48.9 (48.9 ± 1.0) (p = 0.0010)*	67.2 (67.2 ± 1.3) (p = 0.0010)*
Croatian-Spanish			
Baseline: OpusMT	22.9 (22.9 ± 1.2)	48.3 (48.3 ± 1.3)	65.5 (65.5 ± 1.5)
OpusMT-FT	29.6 (29.6 ± 1.4) (p = 0.0010)*	53.8 (53.8 ± 1.4) (p = 0.0010)*	**58.7 (58.7 ± 1.6) (p = 0.0010)***
NLLB-dist-600M	22.1 (22.1 ± 1.5) (p = 0.0480)*	48.4 (48.4 ± 1.3) (p = 0.2547)	68.9 (68.8 ± 4.2) (p = 0.0559)
NLLB-dist-600M-FT	**30.3 (30.3 ± 1.4) (p = 0.0010)***	**54.5 (54.5 ± 1.2) (p = 0.0010)***	**58.8 (58.8 ± 1.7) (p = 0.0010)***
GoogleT	27.1 (27.0 ± 1.1) (p = 0.0010)*	52.9 (52.9 ± 1.1) (p = 0.0010)*	61.0 (61.0 ± 1.4) (p = 0.0010)*
French-Croatian			
Baseline: OpusMT	14.2 (14.2 ± 1.0)	40.9 (40.9 ± 1.2)	74.6 (74.6 ± 1.3)
OpusMT-FT	**19.1 (19.1 ± 1.2) (p = 0.0010)***	**46.0 (46.1 ± 1.4) (p = 0.0010)***	**68.9 (68.9 ± 1.5) (p = 0.0010)***
NLLB-dist-600M	12.5 (12.5 ± 1.1) (p = 0.0010)*	40.5 (40.5 ± 1.0) (p = 0.1479)	79.8 (79.8 ± 3.1) (p = 0.0010)*
NLLB-dist-600M-FT	**18.8 (18.8 ± 1.1) (p = 0.0010)***	**46.3 (46.3 ± 1.2) (p = 0.0010)***	**69.4 (69.4 ± 1.4) (p = 0.0010)***
GoogleT	**18.7 (18.6 ± 1.2) (p = 0.0010)***	**47.0 (47.0 ± 1.1) (p = 0.0010)***	**69.8 (69.8 ± 1.3) (p = 0.0010)***
Croatian-French			
Baseline: OpusMT	20.3 (20.3 ± 1.1)	46.7 (46.7 ± 1.3)	71.4 (71.4 ± 1.5)
OpusMT-FT	**25.6 (25.6 ± 1.3) (p = 0.0010)***	**51.7 (51.7 ± 1.3) (p = 0.0010)***	**65.0 (65.0 ± 1.7) (p = 0.0010)***
NLLB-dist-600M	19.3 (19.2 ± 0.9) (p = 0.0120)*	47.2 (47.2 ± 0.8) (p = 0.1479)	74.1 (74.2 ± 1.3) (p = 0.0010)*
NLLB-dist-600M-FT	**26.0 (25.9 ± 1.3) (p = 0.0010)***	**51.7 (51.7 ± 1.2) (p = 0.0010)***	**65.6 (65.6 ± 1.7) (p = 0.0010)***
GoogleT	24.2 (24.2 ± 1.4) (p = 0.0010)*	**51.5 (51.4 ± 1.2) (p = 0.0010)***	67.7 (67.7 ± 1.6) (p = 0.0010)*
French-Catalan			
Baseline: OpusMT	24.1 (24.1 ± 1.2)	48.7 (48.7 ± 1.3)	64.7 (64.6 ± 1.6)
OpusMT-FT	34.0 (34.1 ± 1.6) (p = 0.0010)*	56.8 (56.9 ± 1.7) (p = 0.0010)*	54.4 (54.3 ± 2.1) (p = 0.0010)*
NLLB-dist-600M	25.3 (25.3 ± 2.0) (p = 0.0440)*	51.7 (51.7 ± 1.2) (p = 0.0010)*	66.8 (66.6 ± 6.2) (p = 0.1648)
NLLB-dist-600M-FT	**37.9 (37.9 ± 1.5) (p = 0.0010)***	**60.2 (60.3 ± 1.4) (p = 0.0010)***	**52.1 (52.0 ± 2.2) (p = 0.0010)***
GoogleT	33.3 (33.3 ± 1.6) (p = 0.0010)*	57.6 (57.6 ± 1.3) (p = 0.0010)*	55.0 (55.1 ± 1.6) (p = 0.0010)*
Catalan-French			
Baseline: OpusMT	24.3 (24.3 ± 1.1)	50.1 (50.2 ± 1.4)	66.1 (66.0 ± 1.7)
OpusMT-FT	33.3 (33.3 ± 1.5) (p = 0.0010)*	57.3 (57.4 ± 1.6) (p = 0.0010)*	56.7 (56.6 ± 2.1) (p = 0.0010)*
NLLB-dist-600M	28.1 (28.2 ± 1.4) (p = 0.0010)*	55.3 (55.3 ± 1.1) (p = 0.0010)*	64.7 (64.6 ± 4.2) (p = 0.1678)
NLLB-dist-600M-FT	**37.4 (37.4 ± 1.4) (p = 0.0010)***	**60.8 (60.9 ± 1.4) (p = 0.0010)***	**53.2 (53.1 ± 2.0) (p = 0.0010)***
GoogleT	33.4 (33.4 ± 2.1) (p = 0.0010)*	58.9 (58.9 ± 1.4) (p = 0.0010)*	56.4 (56.4 ± 2.1) (p = 0.0010)*

Table 2: Evaluation metrics for all language pairs

	Spanish-Croatian	**French-Catalan**
Base model	32	27
Fine-tuned model	**66**	**71**
Both	2	2
Total	100	100

Table 3: Results of the human evaluation

puts were judged to be of equal quality. These results align with the automatic evaluation metrics and support the conclusion that fine-tuning on the RomCro corpus improves MT output quality.

5 Conclusions and future work

This study has demonstrated the effectiveness of fine-tuning existing NMT models using RomCro v.2.0 for literary translation tasks. Our experiments confirm that the fine-tuned models consistently outperform baseline models and, in most cases, achieve results comparable to or better than Google Translate. These findings are supported by both automatic evaluation metrics and human assessments, reinforcing the value of domain-specific corpora for improving translation quality in literary contexts. The fine-tuning approach has shown particular promise for underrepresented language pairs, such as Spanish-Croatian, where the improvements in translation quality were substantial.

However, despite these advances, some challenges remain. While fine-tuning led to improved performance across all evaluated language pairs, certain discrepancies were noted, particularly in the translation of stylistically complex literary segments. This highlights the need for further refinements, including more targeted preprocessing and domain adaptation techniques. Additionally, the human evaluation process, though valuable, was limited in scope and should be expanded to include a larger number of segments, additional language pairs, and a more detailed qualitative analysis of translation errors.

For future work, we plan to expand human evaluation efforts by increasing the number of annotated segments and incorporating more translation directions. We will also conduct a more in-depth analysis of translation errors to identify specific linguistic phenomena where NMT models struggle. We also plan to use these fine-tuned models to produce bilingual e-books and pedagogical materials for language learners, leveraging the improved literary translations produced by the fine-tuned models.

6 Limitations of this study

One limitation of our study concerns the choice of evaluation metrics. While we primarily relied on automatic metrics such as BLEU, chrF2, and TER, we initially chose not to include COMET in our main evaluation pipeline. This decision was motivated by the fact that COMET's results depend heavily on the underlying model used and on the possibility of training custom models, both of which, in our view, warrant further investigation.

However, for the sake of completeness, we later computed COMET scores using the default model `wmt22-comet-da` to compare the systems evaluated in this study. The results, presented in the tables 4 to 9, confirm two main trends. First, Google Translate consistently achieves the highest COMET scores across all evaluated language pairs (except for the mean in the Catalan-French scenario, see table 9). Second, and importantly, our fine-tuned models show clear improvements over their respective baselines even under COMET evaluation, which reinforces the effectiveness of fine-tuning with RomCro v.2.0, regardless of the evaluation metric applied. These results provide further support for our conclusions and highlight the robustness of our approach.

We did not include the use of large language models (LLMs) for translation in our study for several reasons. First, the wide variety of available models and the strong dependency of translation quality on prompt design would require a detailed and systematic analysis, which we believe deserves a dedicated study and a separate publication. Second, fine-tuning large language models demands significantly more powerful computational resources than were available to us at the time of conducting this research.

Acknowledgments

This work was supported by the Croatian Science Foundation under the project number MOBODL-2023-08-9511, funded by the European Union – NextGenerationEU.

References

Gorana Bikić-Carić, Bojana Mikelenić, and Metka Bezlaj. 2023. Construcción del RomCro, un corpus paralelo multilingüe. *Procesamiento del lenguaje natural*, 70:99–110.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation. *Translation Spaces*, 11(2):184–212.

Damien Hansen and Emmanuelle Esperança-Rodier. 2022. Human-Adapted MT for Literary Texts: Reality or Fantasy? In *NeTTT 2022*, pages 178–190.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Philipp Koehn. 2020. *Neural machine translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. Association for Computational Linguistics.

Marie-Aude Lefer. 2021. Parallel corpora. In *A practical handbook of corpus linguistics*, pages 257–282. Springer.

Els Lefever, Lieve Macken, and Veronique Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 496–504.

Clara Inés López Rodríguez. 2016. Using corpora in scientific and technical translation training: resources to identify conventionality and promote creativity. *Cadernos de traduçao*, 36:88–120.

Bojana Mikelenić and Antoni Oliver. 2024. Using a multilingual literary parallel corpus to train NMT systems. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 1–9.

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

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 of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

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 *Computer-Assisted Literary Translation*. Routledge. Num Pages: 26.

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

Appendices:

Metric signatures

- BLEU: nrefs:1 | bs:1000 | seed:12345 |case:mixed | eff:no | tok:13a | smooth:exp| version:2.4.3
- chrF2: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.4.3
- TER: nrefs:1 | bs:1000 | seed:12345 | case:lc | tok:tercom | norm:no | punct:yes | asian:no | version:2.4.3
- COMET: Unbabel/wmt22-comet-da

System	COMET Mean	vs. OpusMT	vs. OpusMT-FT	vs. NLLB200	vs. NLLB200-FT	vs. GoogleT
OpusMT	0.7863	-	F	T	F	F
OpusMT-FT	0.8180	T	-	T	T	F
NLLB200	0.7761	F	F	-	F	F
NLLB200-FT	0.8150	T	F	T	-	F
GoogleT	0.8343	T	T	T	T	-

Table 4: COMET mean scores and pairwise system comparisons for Spanish-Croatian. T indicates the system in the row outperforms the system in the column, F otherwise.

System	COMET Mean	vs. OpusMT	vs. OpusMT-FT	vs. NLLB200	vs. NLLB200-FT	vs. GoogleT
OpusMT	0.7697	-	F	F	F	F
OpusMT-FT	0.8011	T	-	T	F	F
NLLB200	0.7703	T	F	-	F	F
NLLB200-FT	0.8092	T	T	T	-	F
GoogleT	0.8124	T	T	T	T	-

Table 5: COMET mean scores and pairwise system comparisons for Croatian-Spanish. T indicates the system in the row outperforms the system in the column, F otherwise.

System	COMET Mean	vs. OpusMT	vs. OpusMT-FT	vs. NLLB200	vs. NLLB200-FT	vs. GoogleT
OpusMT	0.7738	-	F	T	F	F
OpusMT-FT	0.8060	T	-	T	F	F
NLLB200	0.7631	F	F	-	F	F
NLLB200-FT	0.8083	T	T	T	-	F
GoogleT	0.8238	T	T	T	T	-

Table 6: COMET mean scores and pairwise system comparisons for French-Croatian. T indicates the system in the row outperforms the system in the column, F otherwise.

System	COMET Mean	vs. OpusMT	vs. OpusMT-FT	vs. NLLB200	vs. NLLB200-FT	vs. GoogleT
OpusMT	0.7378	-	F	T	F	F
OpusMT-FT	0.7477	T	-	T	F	F
NLLB200	0.7319	F	F	-	F	F
NLLB200-FT	0.7725	T	T	T	-	F
GoogleT	0.7796	T	T	T	T	-

Table 7: COMET mean scores and pairwise system comparisons for Croatian-French. T indicates the system in the row outperforms the system in the column, F otherwise.

System	COMET Mean	vs. OpusMT	vs. OpusMT-FT	vs. NLLB200	vs. NLLB200-FT	vs. GoogleT
OpusMT	0.7104	-	F	F	F	F
OpusMT-FT	0.7698	T	-	T	F	F
NLLB200	0.7568	T	F	-	F	F
NLLB200-FT	0.7945	T	T	T	-	F
GoogleT	0.7973	T	T	T	T	-

Table 8: COMET mean scores and pairwise system comparisons for French-Catalan. T indicates the system in the row outperforms the system in the column, F otherwise.

System	COMET Mean	vs. OpusMT	vs. OpusMT-FT	vs. NLLB200	vs. NLLB200-FT	vs. GoogleT
OpusMT	0.6876	-	F	F	F	F
OpusMT-FT	0.7507	T	-	T	F	F
NLLB200	0.7377	T	F	-	F	F
NLLB200-FT	0.7950	T	T	T	-	F
GoogleT	0.7898	T	T	T	T	-

Table 9: COMET mean scores and pairwise system comparisons for Catalan-French. T indicates the system in the row outperforms the system in the column, F otherwise.