Using a multilingual literary parallel corpus to train NMT systems

Bojana Mikelenić
University of Zagreb
bmikelen@ffzg.unizg.hr

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Abstract

This article presents an application of a multilingual and multidirectional parallel corpus composed of literary texts in five Romance languages (Spanish, French, Italian, Portuguese, Romanian) and a Slavic language (Croatian), with a total of 142,000 segments and 15.7 million words. After combining it with very large freely available parallel corpora, this resource is used to train NMT systems tailored to literature. A total of five NMT systems have been trained: Spanish-French, Spanish-Italian, Spanish-Portuguese, Spanish-Romanian and Spanish-Croatian. The trained systems were evaluated using automatic metrics (BLEU, chrF2 and TER) and a comparison with a rule-based MT system (Apertium) and a neural system (Google Translate) is presented. As a main conclusion, we can highlight that the use of this literary corpus has been very productive, as the majority of the trained systems achieve comparable, and in some cases even better, values of the automatic quality metrics than a widely used commercial NMT system.

1 Introduction

Parallel multilingual corpora have a wide use and are known for their application in different kinds of linguistic research (contrastive linguistics, translation studies, phraseology, lexicography, etc.) (Lefer, 2021), translation training (López Rodríguez, 2016) and training of machine translation systems (Koehn et al., 2007; Koehn, 2020), as well as terminology extraction (Lefever et al., 2009).

The parallel corpus RomCro (Bikić-Carić et al., 2023) was created taking into account all these possible applications. This project started in autumn 2019 and it is financed by the Faculty of Humanities and Social Sciences of the University of Zagreb. RomCro is a multilingual and multidirectional parallel corpus, which is aligned and annotated with MSD (Morpho-Syntactic Description) tags. It is composed of original literary texts written in five Romance languages (Spanish, French, Italian, Portuguese, Romanian) as well as Croatian, and their respective published translations into the other five languages. Even though lemmatization and annotation are not relevant for the task at hand, they were completed in order to allow for different uses of the corpus, such as extracting desired structures and their translations for contrastive analysis or translator training.

Most previous studies about machine translation (MT) of literary texts are quite recent (from 2012 onwards). According to Toral and Way (2015), a key challenge in literary translation is preserving not only the meaning, but also the reading experience. This is a key difference to other domains, for example, technical or legal texts. Hansen and Esperanza-Rodier (2022) evaluate a customized MT system tailored for a literary translator specializing in fiction. The study demonstrates that fine-tuning a base model with a smaller subset of custom training data can yield translations closer to human references, despite the raw output still falling short of human quality. Other studies (Oliver, 2023) also suggest the idea of training author-tailored NMT systems for literary texts.

B. Vanroy, M.-A. Lefer, L. Macken, P. Ruffo
Proceedings of the 1st Workshop on Creative-text Translation and Technology, p. 3–11
Sheffield, United Kingdom, June 2024.

Most of the studies remark the idea that translation technologies are still not mature enough for translation of literary works ready to be published and that human translators are needed for this task. Bur NMT systems for literary works can still have several interesting uses, as for example (1) produce draft translation for editorial teams to decide whether to publish a translation of a novel in a given market, promoting the cultural interchange and the visibility of authors writing in smaller languages; (2) produce bilingual electronic books where users can access the translation of difficult sentences or paragraphs, promoting reading in the original language, among others.

We will dedicate the first part of this paper to describe the corpus RomCro, in order to understand its characteristics as well as the process of its creation, and then proceed to explore its application in Neural Machine Translation (NMT).

2 Building RomCro

The corpus contains 27 original titles: seven in Spanish, six in French, four in Italian, four in Romanian, three in Portuguese and three in Croatian, as it is shown in Table 1 (including the author and the year of publication). Adding the translation to all the other languages, that makes 162 texts in total. However, there are three translated texts that are not yet available[1] and two that were acquired and added recently[2] only to the version of the corpus available through Sketch Engine (Kilgarriff et al., 2008). The version of the corpus used in this experiment does not include these five texts, so it contains 157 texts, counting the originals and their translations. The total number of translation units is 142,470, and the total number of words is 15.7 million. The distribution by language in millions of words is as follows: French 2.8, Spanish 2.7, Romanian 2.6, Italian 2.6, Portuguese 2.6, and Croatian 2.4.

There were six stages in building the corpus: 1) Selection and collection of texts, 2) Digitization of texts, 3) Preparation for segmentation and sentence alignment, 4) Segmentation, alignment and manual correction, 5) Lemmatization and morphosyntactic annotation (with MSD tags), and 6) Access to the corpus.

One of the main challenges was to find high quality material translated from the original language into the rest of the languages, which is why literary texts were chosen. The uneven divide between the number of originals in each language (Table 1) is due to a higher availability of titles translated from some languages (e.g., Spanish) then other, smaller ones (e.g., Croatian). In order to keep the corpus as synchronic as possible, the texts should have been published relatively recently. This was more difficult for some languages, namely Romanian, where two titles from the first half of the 20th century had to be selected. To maintain homogeneity in the corpus, the inclusion of exclusively European varieties of Spanish, French and Portuguese was preferred. However, since four titles were translated only into Brazilian Portuguese,[3] they were added to the corpus with a possibility of excluding them when consulting it, filtering by notes provided in each segment.

Once the selection of texts was completed, digitization of those not available in digital format was initiated. They were scanned and then an Optical Character Recognition using Abbyy FineReader was performed.

In the next stage, the material was prepared for segmentation and alignment by manually correcting texts in MS Word. Several undergraduate and master level students collaborated on the project, reviewing and correcting the results of this digitization, that is, preparing the texts for automatic alignment.

The segmentation and alignment was performed using LF Aligner,[4] a freely available tool based on Hunalign (Varga et al., 2005). The results were again revised and corrected manually.

The lemmatization and morphosyntactic annotation was done using the annotators available via Sketch Engine, which were FreeLing (Padró, 2011) for Spanish, French, Italian and Portuguese, and MULTEXT-East (Erjavec et al., 2003; Erjavec, 2017) for Romanian and Croatian.

A lemmatized and POS tagged version of the corpus containing 159 texts is available on Sketch Engine. For direct access to the untagged TMX and TSV versions used in this experiment (com-

[1] The book *El asombroso viaje de Pomponio Flato* is not available in Romanian, while *Dora i Minotaur: Moj život s Picassom* is still not translated to Spanish and Portuguese.

[2] These are the Portuguese translation of *Maitreyi* and the Italian translation of *Muzej bezuvjetne predaje*.

[3] The texts are as follows: *A fada carabina* by Daniel Pennac, *A forma da água* by Andrea Camilleri, *Acontecimentos na Irrealidade Imediata* by Max Blecher and *Nostalgia* by Mircea Cărtărescu.

[4] `https://sourceforge.net/projects/aligner/`

n.	Lang.	Titles:
1	ES	La sombra del viento (C.R. Zafón, 2001)
2		La catedral del mar (I. Falcones, 2006)
3		El juego del ángel (C.R. Zafón, 2008)
4		El asombroso viaje de Pomponio Flato (E. Mendoza, 2008)
5		Soldados de Salamina (J. Cercas, 2001)
6		El mapa del tiempo (F. J. Palma, 2008)
7		El tiempo entre costuras (M. Dueñas, 2009)
8	FR	Seras-tu là ? (G. Musso, 2006)
9		HHhH (L. Binet, 2010)
10		Un barrage contre le Pacifique (M. Duras, 1950)
11		La Fée Carabine (D. Pennac, 1987)
12		L'amant (M. Duras, 1984)
13		A l'ombre des jeunes filles en fleur (M. Proust, 1919)
14	IT	Imprimatur (Monaldi & Sorti, 2002)
15		Le otto montagne (P. Cognetti, 2017)
16		La forma dell'acqua (A. Camilleri, 1994)
17		L'amica geniale (E. Ferrante, 2011)
18	RO	Maitreyi (M. Eliade, 1933)
19		Întâmplări în irealitatea imediată (M. Blecher, 1936)
20		Nostalgia (M. Cărtărescu, 1993)
21		Cartea şoaptelor (V. Vosganian, 2009)
22	PT	A viagem do elefante (J. Saramago, 2008)
23		Nenhum olhar (J. L. Peixoto, 2000)
24		As intermitências da morte (J. Saramago, 2005)
25	HR	Muzej bezuvjetne predaje (D. Ugrešić, 1998)
26		Mediteranski brevijar (P. Matvejević, 1987)
27		Dora i Minotaur: Moj život s Picassom (S. Drakulić, 2015)

Table 1: A list of the original titles in the corpus

prising 157 texts) under the CC-BY-NC-4.0 license, please refer to the ELRC (European Language Resource Coordination) platform.[5] In both formats, the order of languages is Spanish (es), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Croatian (hr). All the versions of the corpus contain notes about the original language, writer, and the original title of the text the segment is from. Additionally, segment order was scrambled to protect copyright.

3 Comparison to similar corpora

Many parallel corpora are freely available on the Internet. The main collection can be found in Opus Corpora (Tiedemann, 2012). However, only a minority of these parallel corpora are created from literary texts, and when available, they do not contain many parallel segments (for example Books[6]), or are created from individual or a small number of works (as Salome[7]).

To the best of our knowledge, this is the first multilingual and almost completely multidirectional parallel corpus that aligns literary texts in several Romance languages and one Slavic. In other similar corpora, the Slavic language was the pivot language (Terzić et al., 2020; Grabar et al., 2018; Akimova et al., 2020), which means that all texts were translated from or to the Slavic language, but not necessarily between each other.

Croatian is present in some multilingual literary corpora, such as TransLiTex (Fraisse et al., 2018) or InterCorp (Čermák, 2019), which also includes literary texts. However, TransLiTex contains translations of a single book into 23 languages and InterCorp is made up of 40 languages, including all

[5] https://elrc-share.eu/repository/search/?q=romcro

[6] https://opus.nlpl.eu/Books.php

[7] https://opus.nlpl.eu/Salome-v1.php

those that form part of RomCro, but has Czech as the pivot language.

4 Using RomCro to train NMT systems tailored to literary texts

As an example of use of RomCro, we explore the training of Neural Machine Translation systems tailored to literature. A total of five NMT systems have been trained, combining the Spanish subcorpus (original and translated material) with subcorpora in the other five languages. In other words, the trained NMT systems were Spanish to French, Italian, Portuguese, Romanian and Croatian. In this section all the processes performed to train and evaluate these systems are described.

4.1 Extending the corpus

First of all, the size of RomCro is insufficient for training these NMT systems. We have about 150,000 segments when we would require several million. To obtain the needed segments, we have combined RomCro with a very large parallel corpus, such as CCMatrix[8] (Schwenk et al., 2021) and MultiCCAligned[9] (El-Kishky et al., 2020). Unfortunately, such very large parallel corpora contain errors, as some segments are not in the correct language and some segment pairs are not translation equivalents. To solve this problem, we have rescored the parallel corpora using the MTUOC-PCorpus-rescorer[10] (Oliver and Álvarez, 2023). This tool automatically detects the language of each segment and checks if each segment pair is really a translation equivalent using SBERT[11] (Reimers and Gurevych, 2019), providing a confidence score. We have used a threshold of 0.75 for each check. In Table 2, the size in segments for the raw and rescored versions can be observed. As we can see, for all language pairs except Spanish-Croatian, we have enough segments in the rescored version.

For Spanish-Croatian we have concatenated several parallel corpora,[12] and eliminated repeated parallel segments, obtaining a corpus with 29 million parallel segments, resulting in 12.4 million after rescoring. This is the corpus that we have used combined with RomCro.

Corpus	Type	Segments	Rescored
spa-fra	CCMatrix	266.5 M	159.7 M
spa-ita	CCMatrix	142.1 M	80.9 M
spa-por	CCMatrix	198.5 M	114.4 M
spa-rom	CCMatrix	53.7 M	25.9 M
spa-cro	MultiCC Al.	2.9 M	88.5 K

Table 2: Size of the corpora before and after rescoring (in millions of segments)

Once we have obtained a curated version of the very large parallel corpora for the different language pairs, or General corpora, we needed to combine them with RomCro, that is, select a subset of the large parallel corpora containing the most similar segment pairs to the segment pairs in RomCro. To combine the corpus, we have used the MTUOC corpus combination algorithm,[13] for all language pairs except Spanish-Romanian and Spanish-Croatian. This program calculates a language model from the Spanish part of the RomCro corpus, and then, for all the segment pairs in the General corpus it calculates the perplexity of the Spanish part using the calculated language model. Then we can select a given number, 20 million in our experiments, of segments with the lowest perplexity. These segments are in a certain way the most similar to those in the RomCro corpus. For Spanish-Romanian and Spanish-Croatian all the available parallel segments after rescoring have been used, so this step was omitted. The training corpus contains some segments from RomCro and some from the General corpus. We assigned a weight of 1 to the segments coming from RomCro and a weight of 0.5 to those coming from the General corpus. These weights were used in the training process, giving greater importance to segments from the literary data. Please note that all the segments for the validation and evaluation corpus come from the RomCro corpus.

4.2 Training NMT systems

We have used Marian[14] (Junczys-Dowmunt et al., 2018) to train general and tailored to literature systems from Spanish to French, Italian, Portuguese, Romanian and Croatian. For the general sys-

[8] https://github.com/facebookresearch/LASER/tree/main/tasks/CCMatrix
[9] https://www.statmt.org/cc-aligned/
[10] https://github.com/mtuoc/MTUOC-PCorpus-rescorer
[11] https://www.sbert.net/
[12] MultiCCAligned, MultiParaCrawl, OpenSubtitles and ELRC-4236.
[13] https://github.com/mtuoc/MTUOC-corpus-combination
[14] https://marian-nmt.github.io/

tem we have used 20 million segments from the rescored corpus, except for Spanish-Romanian, where the whole 25.9M segments after rescoring have been used; and Spanish-Croatian, where we have used the whole concatenated corpus after rescoring consisting of 12.4M segments. The systems tailored to literature have been trained with the corpora described in the section 4.1. These systems were compared with Apertium[15] (Forcada et al., 2011), when available, and Google Translate,[16] as described below.

The training has been performed using marian-nmt with a transformer configuration, using SentencePiece[17] (Kudo and Richardson, 2018) as a subword tokenizer. The weights from the combination step have been used for training.

4.3 Evaluation of the trained NMT systems

In tables 3 to 7 we present the evaluation figures for all the MT systems under study for the language pairs from Spanish to the rest of the currently available languages in RomCro. The evaluation has been performed using Sacrebleu[18] (Post, 2018): BLEU (Papineni et al., 2002), chrF2 (Popović, 2015) and TER (Snover et al., 2006). The Appendix A shows the signatures of the three metrics stating the exact configuration parameters as reported by Sacrebleu. We did not use neural evaluation metrics as COMET (Rei et al., 2020) or BLEURT (Sellam et al., 2020), as these metrics are very dependent of the used model and can give different results for different language pairs, making the results difficult to compare between the studied language pairs. For all language pairs, an evaluation set has been extracted from the RomCro corpus. The segments used in these evaluation sets have been randomly selected and they are not present in the training set nor in the validation set. We have translated these evaluation sets with all the MT systems under study, namely:

- Apertium for those language pairs with an available Apertium system: Spanish-French, Spanish-Italian and Spanish-Portuguese.
- Marian Generic: trained with 20 million segments from the General corpus (except for Spanish-Romanian and Spanish-Croatian, as explained in subsection 4.2).
- Marian RomCro: trained with RomCro and the 20 million segments most similar to RomCro selected from the General corpus (except for Spanish-Romanian and Spanish-Croatian, as explained in subsection 4.2).
- Google Translate through its Python API (Translations with Google Translate were performed between July 21-25, 2023).

For each language pair the values of BLEU, chrF2 and TER for all the evaluated systems are presented. Best values for each metric and language pair are marked in bold in the tables. In the same table, significance figures for the comparison of the Marian Generic and Marian RomCro, on one hand, and for Marian RomCro and Google Translate, on the other hand, are presented. These figures have been calculated with paired bootstrap resampling test with 1,000 resampling trials, using the -paired-bs option in Sacrebleu. In this way, one of the systems is pairwise compared to the system considered as the baseline (indicated with a B in the tables). Assuming a significance threshold of 0.05, the null hypothesis can be rejected for p-values < 0.05 (marked with "*" in the tables), indicating that the differences are significant and are not produced by chance.

In Table 3 the evaluation figures for the Spanish-French language pair can be observed. First of all and for all language pairs having Apertium, any neural system achieves better results than this transfer system. For Spanish-French the Marian RomCro achieves slightly better, but statistically significant results for BLEU (an increase of 1.3 points) and TER (a decrease of 1.8 poins) than the Marian Generic. Comparing Marian RomCro and Google Translate, the latter achieves better results in all metrics, but this difference is only significant for chrF2 (with an increase of 1.1 points).

In Table 4 the evaluation figures for Spanish-Italian are shown. For this language pair, training with RomCro is very productive, as this system achieves significantly better results than Marian Generic and Google Translate. For BLEU we get an improvement of 7.5 points with respect to Marian Generic and 1.4 with respect to Google Translate.

In Table 5 the evaluation results for the Spanish-Portuguese language pair are presented. This language pair presents a similar behaviour to Spanish-Italian, with the Marian RomCro system getting

[15] https://www.apertium.org/
[16] https://translate.google.com/
[17] https://github.com/google/sentencepiece
[18] https://github.com/mjpost/sacrebleu

System	BLEU	chrF2	TER
Apertium es-fr	19.4	49.5	72.3
Marian Generic es-fr	31.9	57.2	58.9
Marian RomCro es-fr	33.2	56.9	**57.1**
GoogleT es-fr	**33.5**	**58.0**	57.4

System	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
B: Marian Generic es-fr	31.9 (31.9 ± 1.3)	57.2 (57.2 ± 1.0)	58.9 (58.9 ± 1.5)
Marian RomCro es-fr	**33.2 (33.2 ± 1.4) (p = 0.0020)***	56.9 (56.9 ± 1.0) (p = 0.1179)	**57.1 (57.0 ± 1.4) (p = 0.0010)***
B: Marian RomCro es-fr	33.2 (33.2 ± 1.4)	56.9 (56.9 ± 1.0)	57.1 (57.0 ± 1.4)
GoogleT es-fr	33.5 (33.4 ± 1.4) (p = 0.2118)	**58.0 (58.0 ± 0.9) (p = 0.0030)***	57.4 (57.4 ± 1.5) (p = 0.1918)

Table 3: Evaluation results for Spanish-French

System	BLEU	chrF2	TER
Apertium es-it	20.3	50.6	68.3
Marian Generic es-it	25.5	56.3	67.1
Marian RomCro es-t	**33.0**	**58.7**	**56.3**
GoogleT es-it	31.6	57.6	57.4

System	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
B: Marian Generic es-it	25.5 (25.5 ± 1.4)	56.3 (56.3 ± 1.1)	67.1 (67.1 ± 3.0)
Marian RomCro es-it	**33.0 (32.9 ± 1.4) (p = 0.0010)***	**58.7 (58.6 ± 1.0) (p = 0.0010)***	**56.3 (56.3 ± 1.5) (p = 0.0010)***
B: Marian RomCro es-it	**33.0 (32.9 ± 1.4)**	**58.7 (58.6 ± 1.0)**	**56.3 (56.3 ± 1.5)**
GoogleT es-it	31.6 (31.6 ± 1.3) (p = 0.0030)*	57.6 (57.6 ± 1.0) (p = 0.0010)*	57.4 (57.4 ± 1.5) (p = 0.0170)*

Table 4: Evaluation results for Spanish-Italian

even better results and outperforming the Marian Generic and Google Translate. For this language pair Google Translate is getting worse results than the Marian Generic (with 5.5 less BLEU points) and Marian RomCro (with 7.1 less BLEU points).

For the Spanish-Romanian language pair (see Table 6), the Marian RomCro again outperforms the Marian Generic systems, and achieves very similar scores to Google Translate. In fact, Google Translate only gets significantly better results for the chrF2 measure (an increment of 0.7 points). For this language pair, all the parallel segments available after rescoring have been used, meaning no corpus combination was performed. This suggests that the results could potentially improve if segments more similar to RomCro could have been selected.

For the Spanish-Croatian language pair (Table 7) our training systems are getting bad results for all the metrics, very far from the values obtained for Google Translate (a decrement of 8.2 BLEU points). This should be due to the small size of the training parallel corpus and the missing corpus combination step. Even so, the use of RomCro improves significantly the results of the Generic MT engine (with an increment of 2 BLEU points).

As a general conclusion from the evaluation, we can confirm that the use of RomCro to create neural machine translation tailored to literature is promising. But there is still a lot of work to be done. Further training experiments should be performed, using some known techniques to further

System	BLEU	chrF2	TER
Apertium es-pt	31.7	58.9	53.9
Marian Generic es-pt	36.4	61.1	51.4
Marian RomCro es-pt	**38.0**	**61.9**	**49.2**
GoogleT es-pt	30.9	57.4	55.8

System	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
B: Marian Generic es-pt	36.4 (36.3 ± 1.5)	61.1 (61.1 ± 1.0)	51.4 (51.4 ± 1.6)
Marian RomCro es-pt	**38.0 (38.0 ± 1.5) (p = 0.0010)***	**61.9 (61.9 ± 1.1) (p = 0.0010)***	**49.2 (49.2 ± 1.5) (p = 0.0010)***
B: Marian RomCro es-pt	**38.0 (38.0 ± 1.5)**	**61.9 (61.9 ± 1.1)**	**49.2 (49.2 ± 1.5)**
GoogleT es-pt	30.9 (30.9 ± 1.3) (p = 0.0010)*	57.4 (57.4 ± 1.0) (p = 0.0010)*	55.8 (55.8 ± 1.4) (p = 0.0010)*

Table 5: Evaluation results for Spanish-Portuguese

System	BLEU	chrF2	TER
Marian Generic es-ro	18.4	45.4	73.4
Marian RomCro es-ro	**21.4**	48.2	69.5
GoogleT es-ro	20.7	**48.9**	**69.1**

System	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
B: Marian Generic es-ro	18.4 (18.4 ± 1.0)	45.4 (45.4 ± 0.8)	73.4 (73.4 ± 1.3)
Marian RomCro es-ro	**21.4 (21.4 ± 1.0) (p = 0.0010)***	**48.2 (48.2 ± 0.9) (p = 0.0010)***	**69.5 (69.5 ± 1.4) (p = 0.0010)***
B: Marian RomCro es-ro	21.4 (21.4 ± 1.0)	48.2 (48.2 ± 0.9)	69.5 (69.5 ± 1.4)
GoogleT es-ro	20.7 (20.7 ± 1.0) (p = 0.0639)	**48.9 (48.9 ± 0.9) (p = 0.0170)***	69.1 (69.1 ± 1.4) (p = 0.1708)

Table 6: Evaluation results for Spanish-Romanian

improve the quality. We plan to experiment with backtranslation, compiling a monolingual literary corpus for the target language, and machine translate these corpora into the source language to create the backtranslated data. So far the only source language in the experiments is Spanish, and we plan to perform further experiments with the other RomCro languages as source languages.

5 Conclusions and future work

We presented a possible use of RomCro, a multilingual and multidirectional parallel corpus of literary texts in six languages. Our study has illustrated the viability of using the RomCro corpus for training neural machine translation systems specifically designed for literary texts. Notably, our findings indicate that these specialized systems outperform generic models and achieve comparable, if not superior, performance compared to Google Translate.

As for future work, other than experimenting with backtranslation and changing the source language, we plan to enlarge the corpus by adding more literary works and other Romance languages. The main difficulty is the lack of works translated to all the languages in the corpus, and this will be even more difficult if we add more languages.

Appendix A - Metric signatures

- BLEU: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.3.1
- chrF2: nrefs:1 | bs:1000 | seed:12345 |

System	BLEU	chrF2	TER
eval.es-MarianGeneric.hr	13.4	39.0	75.6
eval1K.es-Marian.hr	15.4	41.3	74.6
eval1K.es-GoogleT.hr	**23.6**	**51.2**	**63.7**

System	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
B: Marian Generic es-hr	13.4 (13.4 ± 1.2)	39.0 (39.1 ± 2.2)	75.6 (75.6 ± 2.0)
Marian RomCro es-hr	**15.4 (15.4 ± 1.3) (p = 0.0010)***	**41.3 (41.3 ± 2.3) (p = 0.0010)***	**74.6 (74.6 ± 2.7) (p = 0.1019)**
B: Marian RomCro es-hr	15.4 (15.4 ± 1.3)	41.3 (41.3 ± 2.3)	74.6 (74.6 ± 2.7)
GoogleT es-hr	**23.6 (23.5 ± 1.1) (p = 0.0010)***	**51.2 (51.2 ± 1.1) (p = 0.0010)***	**63.7 (63.7 ± 1.3) (p = 0.0010)***

Table 7: Evaluation results for Spanish-Croatian

case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.3.1

- TER: nrefs:1 | bs:1000 | seed:12345 | case:lc | tok:tercom | norm:no | punct:yes | asian:no | version:2.3.1

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

Akimova, Marina, Anastasia Belousova, Igor Pilshchikov, and Vera Polilova. 2020. Cpcl: A multilingual parallel corpus of poetic texts and new perspectives for comparative literary studies. In *DHN2020: The workshop "Parallel Corpora as Digital Resources and Their Applications"(Riga, 2020): Abstracts*.

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

Čermák, Petr. 2019. A parallel corpus of 40 languages. *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*, 90:93.

El-Kishky, Ahmed, 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.

Erjavec, Tomaž, Cvetana Krstev, Vladimir Petkevic, Kiril Simov, Marko Tadić, and Duško Vitas. 2003. The multext-east morphosyntactic specification for slavic languages. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 25–32.

Erjavec, Tomaž. 2017. Multext-east. *Handbook of linguistic annotation*, pages 441–462.

Forcada, Mikel L, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Fraisse, Amel, Quoc-Tan Tran, Ronald Jenn, Patrick Paroubek, and Shelley Fisher Fishkin. 2018. Translitex: A parallel corpus of translated literary texts. In *Eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Grabar, Natalia, Olga Kanishcheva, and Thierry Hamon. 2018. Multilingual aligned corpus with ukrainian as the target language. In *SLAVICORP*.

Hansen, Damien and Emmanuelle Esperança-Rodier. 2022. Human-adapted mt for literary texts: Reality or fantasy? In *NeTTT 2022*, pages 178–190.

Junczys-Dowmunt, Marcin, 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.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2008. The sketch engine. *Practical Lexicography: a reader*, pages 297–306.

Koehn, Philipp, 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.

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

Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

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

Lefever, Els, 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.

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

Oliver, Antoni and Sergi Álvarez. 2023. Filtering and rescoring the ccmatrix corpus for neural machine translation training. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 39–45.

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

Padró, Lluís. 2011. Analizadores multilingües en freeling. *Linguamática*, 3(1):13–20.

Papineni, Kishore, 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.

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

Post, Matt. 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, October. Association for Computational Linguistics.

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

Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sellam, Thibault, 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.

Snover, Matthew, 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.

Terzić, Dušica, Saša Marjanović, Dejan Stosic, and Aleksandra Miletic. 2020. Diversification of serbian-french-english-spanish parallel corpus parcolab with spoken language data. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 61–70. Springer.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

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

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING*, page 590.