Literary Translations and Synthetic Data for Machine Translation of Low-resourced Middle Eastern Languages

Sina Ahmadi[1,*] Razhan Hameed[2] Rico Sennrich[1]

[1]Department of Computational Linguistics, University of Zurich, Switzerland
[2]Vox AI, Netherlands

*sina.ahmadi@uzh.ch

Abstract

Middle Eastern languages represent a linguistically diverse landscape, yet few have received substantial attention in language and speech technology outside those with official status. Machine translation, a cornerstone application in computational linguistics, remains particularly underexplored for these predominantly non-standardized, spoken varieties. This paper proposes data alignment and augmentation techniques that leverage monolingual corpora and large language models to create high-quality parallel corpora for low-resource Middle Eastern languages. Through systematic fine-tuning of a pretrained machine translation model in a multilingual framework, our results demonstrate that corpus quality consistently outperforms quantity as a determinant of translation accuracy. Furthermore, we provide empirical evidence that strategic data selection significantly enhances cross-lingual transfer in multilingual translation systems. These findings offer valuable insights for developing machine translation solutions in linguistically diverse, resource-constrained environments.

DOLMA-NLP/bitext-mining

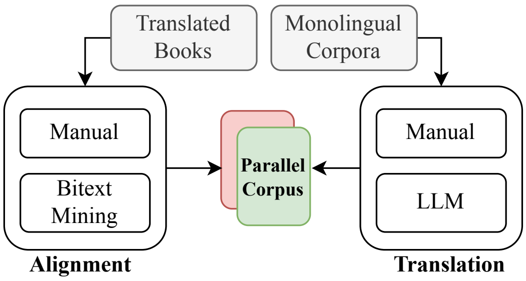

Figure 1: Approaches to create parallel corpora for the selected low-resourced languages in this paper

1 Introduction

Machine translation (MT) represents one of the most transformative applications in natural language processing (NLP), driving numerous breakthrough discoveries in the field. The evolution of MT has progressed from rule-based techniques to sophisticated deep learning approaches and, most recently, to large language models (LLMs) (Zhu et al., 2024b). Despite these paradigm shifts, data availability remains the fundamental constraint, leaving MT far from solved for low-resourced and under-represented languages and varieties. Of particular interest to this paper are such languages in the Middle East–a region with rich linguistic heterogeneity. Many languages in the Middle East lack formal status or standardization, face sociopolitical marginalization, and are systematically disadvantaged in technological development. Consequently, these languages have not benefited equitably from recent advances in MT technology, widening the digital language divide.

In our previous work, PARME–described in detail in (Ahmadi et al., 2025), we explored a participatory research initiative where native speakers contribute to translating sentences into eight Middle Eastern languages: Luri Bakhtiari, Laki Kurdish, Gilaki, Hawrami, Mazandarani, Southern Kurdish, Talysh, and Zazaki. Collecting data in a context where spoken tradition predominates over writing presents significant challenges. This effort resulted in over 36,000 translations, which were used to fine-tune the No Language Left Behind (NLLB) pretrained translation model (Team et al., 2024). Our previous experiments yield BLEU scores ranging from 2.89 to 16.54, indicating substantial room for improvement.

The current paper expands on our previous data collection approach through two complementary approaches illustrated in Figure 1. In the first approach, we leverage literary works by aligning sentences from translated works in the selected languages to the original English texts, using both manual and automated alignment tech-

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 110–118
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

niques. In the second approach, we extract sentences from monolingual corpora and translate them using Gemini-2.0-flash, creating synthetic parallel data. Using these datasets, we then systematically evaluate how these various data sources affect the performance of fine-tuned multilingual translation models. Our findings reveal that incorporating these new datasets improves model performance overall, but with an important caveat: increasing data quantity for one language can sometimes adversely affect performance for others in a multilingual setting. This highlights the complex interplay between data quantity, quality, and distribution in multilingual MT systems for low-resource languages.

2 Related Work

2.1 Low-Resourced MT

MT systems typically require millions of parallel sentences for effective training, a requirement met by only a few dozen high and medium-resource languages, primarily European. For low-resource languages, researchers have developed various approaches to address data scarcity. Synthetic data augmentation techniques include leveraging dictionaries and morphological variations (Alam et al., 2024), substituting rare words to create new training sentences (Fadaee et al., 2017), and mapping word embeddings from high-resource to low-resource languages through bilingual lexicon induction (Li et al., 2024). Synthetic data generation via back-translation (Sennrich et al., 2016) or forward-translation (Zhang and Zong, 2016) are common strategies, as well. Other approaches leverage the capacity of multilingual models to enhance related low-resourced languages using transfer learning (Ko et al., 2021), fine-tuning (Moslem et al., 2023) and adapters (Pham et al., 2024).

The emergence of LLMs has opened new possibilities for low-resource MT through prompting (Zhang et al., 2023), few-shot learning (Hendy et al., 2023), and in-context translation (Raunak et al., 2023). However, recent studies emphasize that translation direction (Zhu et al., 2024a) along with parallel data quality during both pre-training and fine-tuning remain crucial for performance (Guo et al., 2024). Furthermore, Iyer et al. (2024) note that "*diversity (in prompts and datasets) tends to cause interference instead of transfer*," highlighting the challenges in leveraging diverse datasets.

2.2 Bitext Mining

To facilitate the creation of parallel corpora from unaligned corpora, bitext mining or bitext retrieval aims to identify potential translation pairs across translated documents or monolingual corpora (Koehn, 2024). This task, of particular interest to low-resourced languages, has been extensively studied previously (Zweigenbaum et al., 2017), including methods for sentence filtering from web-crawled content (Chaudhary et al., 2019). Some approaches to bitext mining rely on automatic translations, as in Bleualign (Sennrich and Volk, 2010), while other approaches leverage semantic representations (Heffernan et al., 2022), with a notable example being Vecalign (Thompson and Koehn, 2019). Recent work by Winata et al. (2024) demonstrates that LLMs can also perform effectively in bitext mining tasks.

Our paper addresses a critical gap in the literature by exploring the intersection of bitext mining, data augmentation using an LLM and, multilingual fine-tuning for low-resource Middle Eastern languages, offering insights into enhancing translation capabilities for these understudied varieties.

3 Methodology

Complementary to PARME (Ahmadi et al., 2025), our previous participatory research where English sentences are translated by experts into one of the selected languages, we explore bitext mining and LLM-based data augmentation to further extract parallel sentences.

3.1 Sentence Alignment

Given a translated content in one of our selected languages, we aim to align the translations to their original sentences. Reaching out to publishers and translators, we could collect 25 translated books and articles for four languages among our eight selected ones: five translated articles for Laki Kurdish, five for Southern Kurdish (two books and three articles), 11 books for Hawrami and four for Gilaki (three articles and one book). All the content were originally translated from English, except in a couple of cases that we excluded as they were originally translated from Persian. The books are all famous novels of George Orwell, Virginia Woolf, Franz Kafka, Ernest Hemingway and Antoine de Saint-Exupéry, except one children book for Southern Kurdish, while the articles discuss specific sociological and medical topics.

To prepare the books for alignment, we first extract the sentences from the original textbooks in English (or their translations in English). Although most of the books are openly available[1], two of them required OCR from the scanned PDFs. Following this step, we preprocess the text in both the original text in English and our translations by normalizing characters, fixing tabulations and excessive newlines and finally, splitting the text into sentences or phrases using KLPT (Ahmadi, 2020b).

Given the set of sentences per work in English along with the translation, we initially aimed to carry out the alignments using an LLM. However, due to the low-resourced status of the selected languages, usage of `Chat GPT-4o` and `Claude 3.7 Sonnet` for our selected languages was far from helpful. As such, we try the following methods.

Manual alignment (M): Providing the sentences in a spreadsheet, we manually align sentences by splitting, merging and editing sentences to create matching translation pairs. Stylistic variation across translators required further attention to the alignment task; for instance, a long passage in the English text might have been translated in one or two short sentences considered to be culturally less relevant to the readers of the translated book. Similarly, specific contexts have required further elaboration by the translator, as in describing "Big Brother" or "Thinkpol" in George Orwell's 1984. Therefore, some alignments require appropriate modifications. The alignment was carried out by expert native speakers.

Automatic Alignment using Vecalign (V): Due to limited workforce for manual alignment, we carried out bitext mining to automatically align remaining translations using a few methods that were far from practical. To do so, we first manually split translations by chapter or long sections to further reduce the range of the possible alignment combinations, also known as *hierarchical mining* (Koehn, 2024). Then, we tried a range of methods: the Microsoft's Bilingual Sentence Aligner (Moore, 2002), Bleualign (Sennrich and Volk, 2010) with translations from PARME's fine-tuned models, embedding-based techniques using LaBSE (Feng et al., 2022), LASER (Artetxe and Schwenk, 2019), SONAR (Duquenne et al., 2023), SBERT (Reimers and Gurevych, 2019) and Vecalign (Thompson and Koehn, 2019). Given that none of the selected languages are included in the pretrained embeddings, we rely on the embeddings of closely-related languages: Persian (PES) for Gilaki and Central Kurdish (CKB) for Laki, Southern Kurdish and Hawrami. To determine the most effective alignment technique, we tested several methods on the manually-aligned corpus of the Little Prince containing 1101 sentence pairs. We measured accuracy as the proportion of sentence pairs that matched between the automatically-aligned and manually-aligned corpora. Table 1 summarizes the accuracy showing that Vecalign with SONAR embeddings produce the highest accuracy. It should be noted that the reported accuracies are limited to a sample in Hawrami without considering the combination of the embeddings and techniques.

Technique		Accuracy (%)
Microsoft Aligner		38.78
Bleualign		32.24
SBERT	LaBSE	2.63
	LASER	2.08
Vecalign	SONAR	**46.5**

Table 1: Accuracy of different bitext mining techniques on a sample of Hawrami translated text. Vecalign with SONAR achieves the highest accuracy (46.5%).

3.2 LLM-based Data Augmentation

Relying on the monolingual corpora available for Southern Kurdish (Ahmadi et al., 2023) and Zazaki, Hawrami (Ahmadi, 2020a) along with Wikipedia dumps[2] for Gilaki, Mazandarani and Zazaki, we implement a few-shot in-context translation approach to optimize in-context example selection using Gemini-2.0-flash, inspired by Agrawal et al. (2023), as follows:

```
Below are examples of {language} to English
translations. Translate the new text
following these patterns:

{language}: {example1}
English: {english_translation1}

[... more examples ...]

Now translate this text to English, only
output the translation:

{language}: {text_to_translate}
English:
```

[1] For English, we relied on the raw text provided by the Project Gutenberg: `https://www.gutenberg.org`

[2] Latest dumps of December 2025

Language	Gemini-2.0-flash		Llama3.3	
	zero	few	zero	few
Luri Bakhtiari	0.06	0.15	0.09	0.09
Gilaki	0.11	0.23	0.09	0.09
Hawrami	0.07	0.21	0.07	0.14
Laki Kurdish	0.10	0.19	0.05	0.11
Mazandarani	0.16	0.36	0.06	0.18
Southern Kurdish	0.18	0.14	0.06	0.13
Talysh	0.07	0.14	0.06	0.11
Zazaki	0.32	0.34	0.13	0.11
Average	0.14	**0.22**	0.08	0.12

Table 2: Zero-shot and few-shot prompting results (BLEU↑ [0, 100]) on Gemini-2.0-flash and Llama3.3. We translate sentences from monolingual corpora using few-shot prompted Gemini.

Our implementation uses BM25 retrieval to find semantically similar examples from a datastore, followed by a custom n-gram based re-ranking method. We calculate n-gram overlap between the test source and retrieved examples using a weighted scoring function that emphasizes coverage of source text terms. Our approach employs a dynamic weighting system where already-covered n-grams receive reduced weight by a lambda factor (set to 0.1) to promote selection of complementary examples.

Table 2 presents preliminary results comparing zero-shot and few-shot prompting on both Gemini-2.0-flash and Llama3.3. While the absolute BLEU scores remain poor, a common challenge when applying general-purpose LLMs to extremely low-resource languages, we observe several important patterns. First, few-shot prompting consistently outperforms zero-shot approaches, with relative improvements for some languages (e.g., Hawrami). Second, Gemini-2.0-flash demonstrates superior performance compared to Llama3.3 across nearly all languages. Through experimentation, we determined that using 16 examples in our prompts produced optimal results, significantly outperforming single-example approaches. Additional examples beyond 16 did not yield further improvements.

Table 3 provides basic statistics of our collected data per language. Luri Bakhtiari (BQI) and Talysh (TLY) are only included in PARME (P), Laki is only included in PARME and manual alignment (PM) while the other languages could benefit from the additional data sources.

Language	P	M	V	L
Luri Bakhtiari (BQI)	999	0	0	0
Gilaki (GLK)	3420	999	1391	22467
Hawrami (HAC)	5796	7050	8367	49987
Laki Kurdish (LKI)	1487	1220	0	0
Mazandarni (MZN)	2345	0	0	49328
Southern Kurdish (SDH)	7806	3681	2495	49992
Talysh (TLY)	1107	0	0	0
Zazaki (ZZA)	2374	0	0	50000
Sum	25,334	12,950	12,253	221,774

Table 3: Basic statistics of the data collected per languages from different data sources: PARME (P), manual (M) and automatic (V) sentence alignment, and LLM (L). Over 272,000 sentence pairs are collected.

4 Experiments

4.1 Experimental Setup

To adapt a multilingual model for our target languages, we leverage NLLB (600M variant) by systematically integrating embeddings from related languages through a structured token-based approach. This integration follows two key steps. First, we expanded the tokenizer's vocabulary by introducing language-specific tokens (e.g., `zza_Latn` for Zazaki) while preserving the existing language tokens. Second, we initialize embeddings for these new tokens by borrowing from phylogenetically related languages: Central Kurdish embeddings for Hawrami, Laki, and Southern Kurdish; Northern Kurdish for Zazaki; and Farsi for Luri Bakhtiari, Gilaki, Mazandarani, and Talysh. For evaluation consistency, we utilize the standardized test sets from PARME, each containing around 1,000 sentences per language in a single orthography. These test sets maintain representativeness across the non-standardized linguistic landscape by incorporating a uniform distribution of dialectal variations.

We conduct X→EN fine-tuning experiments with various data source combinations, e.g., PL for merging PARME and LLM-based datasets. We evaluate the performance using BLEU metric in SacreBLEU (Post, 2018).[3] Our baseline represents the highest BLEU score achieved by NLLB prior to fine-tuning. For fine-tuning, we employ a batch size of 8 with 4-step gradient accumulation, a conservative learning rate of 3e-5, and trained for 20 epochs with 0.1 warmup. Both source and target sequences were truncated to 128 tokens, and

[3] `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2`

Language	Baseline	P	PM	PV	PMV	PL	PMVL	$\text{PML}_{\text{Zazaki}}$
Luri Bakhtiari$^{\text{P}}$	0.75	**4.38**	3.67 ± 0.15	3.55 ± 0.16	3.78 ± 0.29	3.37 ± 0.39	3.26 ± 0.41	3.04 ± 0.19
Gilaki$^{\text{PMVL}}$	1.98	2.73	**4.22** ± 0.15	3.18 ± 0.13	3.92 ± 0.26	3.44 ± 0.17	3.49 ± 0.16	2.94 ± 0.18
Hawrami$^{\text{PMVL}}$	0.9	8.23	**15.46** ± 0.48	11.55 ± 2.78	10.86 ± 0.54	8.11 ± 0.11	8.93 ± 0.70	10.34 ± 2.15
Laki Kurdish$^{\text{PML}}$	1.89	6.33	**9.11** ± 0.67	7.18 ± 2.13	6.81 ± 0.79	4.80 ± 0.37	4.39 ± 0.47	5.43 ± 0.80
Mazandarani$^{\text{PL}}$	1.32	5.23	**5.50** ± 0.30	5.05 ± 0.83	5.32 ± 0.22	4.34 ± 0.28	4.22 ± 0.12	4.62 ± 0.22
Southern Kurdish$^{\text{PMVL}}$	2.77	9.93	**10.64** ± 0.46	8.68 ± 0.27	8.99 ± 0.60	7.61 ± 0.36	7.80 ± 0.48	8.34 ± 0.21
Talysh$^{\text{P}}$	1.03	3.01	**6.70** ± 0.52	5.22 ± 2.28	4.21 ± 1.43	2.36 ± 0.29	2.32 ± 0.56	3.66 ± 1.21
Zazaki$^{\text{PL}}$	2.82	3.45	3.75 ± 0.30	2.55 ± 0.45	3.67 ± 0.35	11.08 ± 0.89	**11.54** ± 0.50	9.99 ± 0.14
Average	1.68	5.41	**7.38** ± 0.19	5.87 ± 0.97	5.94 ± 0.22	5.64 ± 0.27	5.74 ± 0.21	6.04 ± 0.48

Table 4: X→EN BLEU scores for the fine-tuned NLLB model across eight languages using different combinations of data sources. Results are reported as mean ± standard deviation over three runs with different random seeds. Data sources where a language is included appear as superscript.

we implemented beam search with a beam size of 5 during inference. Training was conducted on NVIDIA RTX 3090 GPUs (24GB VRAM) with completion times of 9.4 to 16.1 hours per model.

4.2 Experimental Results

Table 4 presents the results of our experiments. To assess the impact of randomness in fine-tuning, we run the process three times by shuffling the train sets with different seeds.We report the mean values of the three systems per data setup along with standard deviations. Analyzing the results indicates:

A: Data quality matters more than quality Among the data setups, PARME (P) merged with manually aligned sentences (M), i.e. PM, achieves the highest BLEU scores for most languages and on average. Surprisingly, PM also improves the performance of Talysh, Zazaki and Mazandarani even though it does not contain additional data in those languages. Luri Bakhtiari's best performing model remains P, the only dataset covering that language. Although LLM-generated dataset along with PARME, i.e., PL, is the largest dataset, the obtained performances are lower than the PM setup and not much higher than P; so including the LLM-generated data does not improve the average BLEU score substantially.

On the standard deviations, they reveal varying levels of model stability across configurations and languages, with some combinations showing remarkable consistency, e.g., Gilaki with PM at ±0.15, while others demonstrate substantial sensitivity to initialization, e.g., Hawrami with PV at ±2.78 and Talysh with PV at ±2.28, suggesting that optimal data selection should consider both performance and reliability.

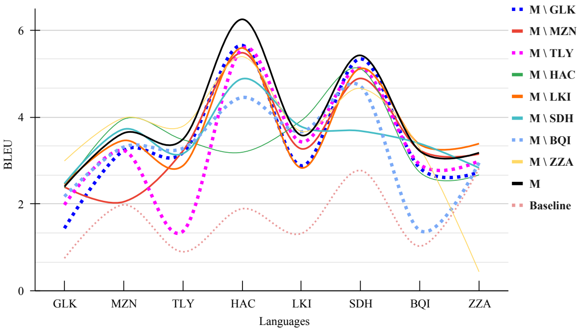

Figure 2: Cross-linguistic dependencies in our multilingual fine-tuning models. Each curve represents performance of a model trained without one language, e.g., `M\GLK`. The solid black line (M) shows the full model.

B: Multilingual data interference While PM is generally the optimal configuration for most languages, Zazaki's performance shows unique sensitivity to dataset composition, particularly when the LLM-generated data (L) is included in the fine-tuning dataset. Within the comprehensive PMVL setup (containing all data sources for all languages), Zazaki achieves its best performance with a BLEU score of 11.54, followed by 11.08 in PL. This observation led us to create a targeted dataset combination–$\text{PML}_{\text{Zazaki}}$–which integrates PM with the Zazaki LLM-generated data only. Although Zazaki still has a comparatively higher BLEU score in this setup (9.99), the average BLEU score is lower than that of PM and other setups where L is included.

To further analyze the implications on other languages in the multilingual setup, we fine-tune models on 1000 randomly-selected sentences in PARME data by excluding data of a language per model; for instance, `M\GLK` is a model fine-tuned on all but Gilaki data. Figure 2 illustrates the evaluation of these models. As expected, re-

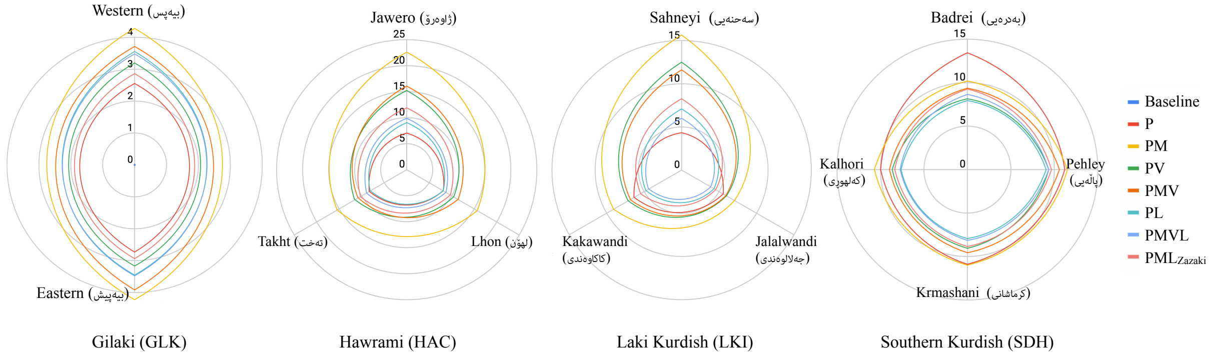

Figure 3: Performance across dialects and model configurations. Each radar chart displays mean BLEU scores from three randomly initialized models for different dialects. Greater extension of curves toward a dialect's axis indicates higher translation performance for that specific dialect.

moving one language's data deteriorates performance for that language, visible in the performance drops along the curves. However, several notable cross-language dependencies emerge. Removing Talysh (TLY) data negatively impacts Gilaki (GLK) and Mazandarani (MZN) performance, while removing Luri Bakhtiari (BQI) data hurts Hawrami (HAC) and Southern Kurdish (SDH). The dependencies manifest asymmetrically, with Zazaki (ZZA) exhibiting both high vulnerability to the removal of its own data and relative resilience to the removal of others, corroborating our earlier observations of its unique behavior.

C: Performance varies depending on the variety Gilaki, Hawrami, Laki and Southern Kurdish include sentences of different varieties/dialects in the test set making cross-dialectal evaluation possible. Figure 3 provides our analysis results for these languages revealing considerable performance disparities within each language. While in Hawrami, the Jawero dialect achieves substantially higher BLEU scores than Takht and Lhon, particularly with PM and PMV configurations, the performance of the models for Eastern and Western varieties of Gilaki is more consistent. Similarly, for Laki Kurdish, the Sahneyi variety benefits more from our fine-tuning approaches than Kakawandi and Jalalwandi varieties. Southern Kurdish shows more balanced performance across its dialects, though Badrei and Krmashani tend to receive slightly higher scores. Nevertheless, we caution against concluding that certain varieties are inherently more difficult to translate, as train and validation sets do not equally represent all varieties, and the test set does not contain the same sentences translated across different varieties. These observed differences may instead reflect varying degrees of representation in training data or linguistic proximity to the source material rather than intrinsic translation difficulty.

5 Conclusion and Discussion

This paper sheds light on eight low-resourced Middle Eastern languages by fine-tuning a pretrained MT model using different sources of data, from manually translated and aligned sentences to automatically aligned and automatically-translated ones. Our experiments demonstrate three key findings. First, data quality consistently outperforms quantity as a determinant of translation accuracy, with the manually aligned (M) data providing the most substantial improvements despite its relatively smaller size. Second, we observed complex cross-linguistic transfer effects where adding data for one language sometimes adversely affects performance for others, highlighting the importance of strategic dataset selection in multilingual systems. Third, we found significant performance variations across dialectal varieties within the same language. While our models perform well on all languages in comparison to the baseline, achieving 15.46 BLEU score for Hawrami at the highest, there remains substantial room for improvement.

Limitations Despite these advances, our work has several limitations. First, we explored only a limited set of open-weight LLMs for data augmentation; future work could investigate a broader range of models, such as MADLAD-400 (Kudugunta et al., 2023) and Mistral (Jiang et al., 2023), and in-context learning strategies. Second, our automatic alignment approach relies on embeddings from closely-related languages, which could be improved by training or fine-tuning embeddings on monolingual data of our selected languages. Third, our data augmentation techniques could be expanded to include synthetic data generation using bilingual lexicon induction, morphological variations, and back-translation methods. Finally, unlike the test sets that are uniform in orthography, our collected data for training and validation are composed of more than one orthography, as in Hawrami, Zazaki and Gilaki. Given that normalization and transliteration of these orthographies are not trivial, future work can also study the effect of orthographical variation on MT.

Ethics Statement Our data collection process adhered to rigorous ethical standards with careful attention to fairness and representation. While we maintained comprehensive inclusion criteria appropriate for low-resource language documentation, we acknowledge that the literary nature of our corpus means some character dialogue may contain language that reflects historical or cultural contexts that modern readers might find objectionable. All materials were obtained through formal agreements with publishers and translators, with appropriate intellectual property permissions secured. Contributors received fair compensation for their work, and their contributions are explicitly acknowledged. Our research prioritizes expanding NLP for underrepresented languages while maintaining responsible data stewardship practices.

Acknowledgments

This work was supported by the Swiss National Science Foundation through the MUTAMUR project (no. 213976).

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8857–8873. Association for Computational Linguistics.

Sina Ahmadi. 2020a. Building a corpus for the Zaza-Gorani language family. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2020, Barcelona, Spain (Online), December 13, 2020*, pages 70–78. International Committee on Computational Linguistics (ICCL).

Sina Ahmadi. 2020b. KLPT–Kurdish Language Processing Toolkit. In *Proceedings of the second Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics.

Sina Ahmadi, Zahra Azin, Sara Belelli, and Antonios Anastasopoulos. 2023. Approaches to corpus creation for low-resource language technology: the case of Southern Kurdish and Laki. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 52–63, Dubrovnik, Croatia. Association for Computational Linguistics.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin Jalal Hamaamin, Saya Kamal Hama, Ardeshir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi Sheikhalikelayeh, Mohammadreza Yadegari, Kewyar Yadegari, and Sedighe Zamani Roodsari. 2025. PARME: Parallel corpora for low-resourced Middle Eastern languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages. *CoRR*, abs/2402.01939.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 261–266. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *CoRR*, abs/2308.11466.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 567–573. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15685–15697. ELRA and ICCL.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 1393–1409. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona T. Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 802–812. Association for Computational Linguistics.

Philipp Koehn. 2024. Neural methods for aligning large-scale parallel corpora from the Web for South and East Asian languages. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 1454–1466. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset.

Fuxue Li, Beibei Liu, Hong Yan, Mingzhi Shao, Peijun Xie, Jiarui Li, and Chuncheng Chi. 2024. A bilingual templates data augmentation method for low-resource neural machine translation. In *Advanced Intelligent Computing Technology and Applications - 20th International Conference, ICIC 2024, Tianjin, China, August 5-8, 2024, Proceedings, Part III*, volume 14877 of *Lecture Notes in Computer Science*, pages 40–51. Springer.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 6-12, 2002, Proceedings*, volume 2499 of *Lecture Notes in Computer Science*, pages 135–144. Springer.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Fine-tuning large language models for adaptive machine translation. *CoRR*, abs/2312.12740.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3168–3184. Association for Computational Linguistics.

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.

Vikas Raunak, Hany Hassan Awadalla, and Arul Menezes. 2023. Dissecting in-context learning of translations in GPTs. *CoRR*, abs/2310.15987.

Nils Reimers 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*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based Sentence Alignment for OCR-generated Parallel Texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers, AMTA 2010, Denver, Colorado, USA, October 31 - November 4, 2010*. Association for Machine Translation in the Americas.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. 2024. MINERS: multilingual language models as semantic retrievers. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 2742–2766. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 388–409. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora, BUCC@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 60–67. Association for Computational Linguistics.