PARME: Parallel Corpora for Low-Resourced Middle Eastern Languages

Sina Ahmadi[α,*] **Rico Sennrich**[α]
Erfan Karami[ð] **Ako Marani**[ə] **Parviz Fekrazad**[ə] **Gholamreza Akbarzadeh Baghban**[ə]
Hanah Hadi[ə] **Semko Heidari**[ə] **Mahîr Dogan**[ʏ] **Pedram Asadi**[ə] **Dashne Bashir**[ʒ]
Mohammad Amin Ghodrati[ɛ] **Kourosh Amini**[ə] **Zeynab Ashourinezhad**[ɾ] **Mana Baladi**[ə]
Farshid Ezzati[ə] **Alireza Ghasemifar**[ə] **Daryoush Hosseinpour**[ə] **Behrooz Abbaszadeh**[ʔ]
Amin Hassanpour[ə] **Bahaddin Jalal Hamaamin**[ø] **Saya Kamal Hama**[ʃ] **Ardeshir Mousavi**[ə]
Sarko Nazir Hussein[ø] **Isar Nejadgholi**[ə] **Mehmet Ölmez**[ə] **Horam Osmanpour**[ə]
Rashid Roshan Ramezani[ə] **Aryan Sediq Aziz**[ø] **Ali Salehi Sheikhalikelayeh**[ə]
Mohammadreza Yadegari[ɲ] **Kewyar Yadegari**[ɥ] **Sedighe Zamani Roodsari**[ɨ]

[α]University of Zurich, Switzerland
[ð]University of Kurdistan, Iran [ə]Independent Researcher [ʒ]Koya University, Iraq
[ʏ]University of Bamberg, Germany [ø]University of Halabja, Iraq [ʃ]University of Sulaimani, Iraq
[ɲ]Shahid Beheshti University, Iran [ɥ]Tabriz Islamic Art University, Iran [ɨ]Auburn University, USA
[ɛ]Qazvin IKI University, Iran [ɾ]University of Guilan, Iran
[ʔ]Algonquin College of Applied Arts and Technology, Canada

[*]sina.ahmadi@uzh.ch

Abstract

The Middle East is characterized by remarkable linguistic diversity, with over 400 million inhabitants speaking more than 60 languages across multiple language families. This study presents a pioneering work in developing the first parallel corpora for eight severely under-resourced varieties in the region—PARME, addressing fundamental challenges in low-resource scenarios including non-standardized writing and dialectal complexity. Through an extensive community-driven initiative, volunteers contributed to the creation of over 36,000 translated sentences, marking a significant milestone in resource development. We evaluate machine translation capabilities through zero-shot approaches and fine-tuning experiments with pretrained machine translation models and provide a comprehensive analysis of limitations. Our findings reveal significant gaps in existing technologies for processing the selected languages, highlighting critical areas for improvement in language technology for Middle Eastern languages.[1]

1 Introduction

The Middle East, also known as West Asia or Southwest Asia–a region roughly bounded by Turkey to the north, Iran to the east, Yemen to the south, and Egypt to the west–stands as a crucial crossroads of civilizations, where geopolitics, economics, and millennia of history converge. This region, home to over 400 million people, contains a linguistic diversity that often goes unrecognized beneath its complex cultural, religious, and political dynamics. While Arabic, Persian, Turkish, and Hebrew dominate official and administrative spheres, the region encompasses numerous distinct languages spanning multiple families, including Afroasiatic, Indo-European, Caucasian and Turkic. This rich linguistic landscape, frequently oversimplified due to prevailing political narratives, not

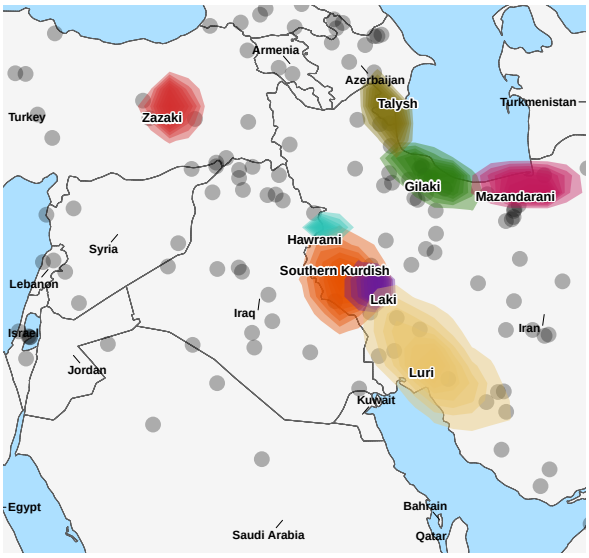

Figure 1: Distribution of languages in the Middle East according to Glottospace (Norder et al., 2022). Gray circles represent documented languages in the region, while colored areas show the approximate distribution of the varieties discussed in this paper. Talysh, Southern Kurdish and Hawrami are spoken across borders.

[1]Corpora and models: DOLMA-NLP/PARME

only reflects the region's diverse cultural heritage but also highlights a critical challenge: the limited availability of language technology for speakers of under-represented languages, perpetuating a cycle of digital exclusion and potentially deepening existing socio-economic disparities (Bird, 2020).

The challenges facing under-represented languages in the Middle East extend beyond mere technological limitations. UNESCO has identified 60 varieties in the region as endangered (Moseley, 2010), with many facing existential threats. These languages often struggle to maintain their status as living languages, suffering from diminishing prestige and declining intergenerational transmission. The lack of standardization presents another significant hurdle, with some languages having multiple competing orthographies but no widely accepted standard due to limited media presence and formal documentation. Limited access to educational resources and formal instruction in these languages further compounds the problem, as younger generations have fewer opportunities to develop literacy in their heritage languages (Sheyholislami and Vessey, 2024). In this context, language technology emerges as a potential lifeline, offering tools and resources that could help revitalize languages and prevent further erosion of linguistic diversity (Crystal, 2002).

This paper focuses on eight under-represented languages of the Middle East as specified in Figure 1: Luri Bakhtiari, Gilaki, Hawrami, Laki Kurdish, Mazandarani, Southern Kurdish, Talysh and Zazaki. Although spoken by speaker populations ranging from 300,000 to 5 million, these languages are severely under-represented, with Zazaki and Hawrami being classified as endangered (Moseley, 2010). Aware of the range of fundamental challenges that these languages face computationally, we implement a community-driven initiative to develop parallel corpora which are essential for evaluating and creating machine translation (MT) systems for these varieties. Additionally, we carry out various experimental evaluations on pretrained models for MT in both zero-shot and fine-tuning scenarios. Our findings indicate that substantial work remains necessary to develop effective MT systems for these languages. We believe that this work opens new research avenues and brings attention to under-explored problems in the context of Middle Eastern languages specifically, but also for low-resourced languages generally.

2 Background

The Middle East exhibits remarkable linguistic diversity, yet the region's language policies often present a concerning landscape of restrictions and oppression. Most nations have implemented predominantly monolingual policies favoring a single official language, justified through nationalist and religious ideologies (Miller, 2003). These policies frequently serve as instruments of discrimination against minority language communities and their speakers (Dubinsky and Starr, 2022). Systematic assimilation campaigns, known as "Arabization" (Absi, 1981), "Turkification" (Üngör, 2012), and "Persianization" (Haddadian-Moghaddam and Meylaerts, 2015), have been implemented throughout recent decades. These campaigns encompass various policies, including the alteration of place names (Jongerden, 2007, p. 31) and the establishment of education systems that exclude mother tongue learning.

These restrictive language policies in education and institutional support have severely impacted the non-official languages in the Middle East, particularly in writing and language development (Bahmany, 2024). This marginalization has resulted in both diminished social prestige and increasing language loss, as fewer parents teach these languages to their children (Fernandes, 2012). For instance, Zamani Roodsari (2023) notes how Gilaki usage varies significantly across demographic groups, with women, younger generations, and more highly educated individuals demonstrating a preference for Persian over their native language.

These languages face an additional fundamental challenge: the struggle for recognition as distinct languages (Shabani, 2021). From a linguistic perspective, many varieties in the Middle East exhibit distinct features in phonology, morphology, syntax, and mutual intelligibility that clearly differentiate them from dominant languages like Persian, Arabic, and Turkish. However, social and political factors often lead to their controversial classification as "dialects", reflecting a broader discourse that aims to diminish their linguistic legitimacy (McDermott and Nic Craith, 2019).

2.1 NLP for Middle Eastern Languages

Beyond the complex sociolinguistic landscape, the development of NLP tools and resources for Middle Eastern languages faces substantial challenges.

	Resources								Tools			
	Grammar	Corpus	UniMorph	UD	WordNet	NLLB	Wiktionary	Wikipedia	LID	MT	Spell checker	ASR
Arabic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hebrew	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Turkish	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Persian	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Northern Kurdish	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Central Kurdish	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
Southern Kurdish	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗
Mazandarani	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
Gilaki	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
Talysh	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Zazaki	✓	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗
Hawrami	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Laki	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Luri Bakhtiari	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 1: Overview of available resources such as Universal Dependencies (UD), and tools, such as language identification (LID) and automatic speech recognition (ASR), in language and speech technologies across our selected languages. ✓ indicates availability and ✗ indicates absence.

Despite ongoing efforts in language revival and ethnic identity recognition (Demir, 2017), NLP progress remains limited. Many of these languages have predominantly oral traditions, with relatively recent attempts to standardization, particularly in writing. The development of orthographies and technological support, including keyboard implementations, remains nascent. Consequently, speakers often resort to adapting the writing systems of dominant languages, creating "unconventional writing"–a phenomenon that poses significant computational challenges, as previously studied by Ahmadi and Anastasopoulos (2023). Needless to say, such challenges are common to many low-resourced languages and have been addressed through participatory research or funded initiatives, as in the context of African or Asian languages (Nekoto et al., 2020; Caswell et al., 2025, *inter alia*). These approaches offer potential pathways for Middle Eastern language communities to overcome similar resource constraints through participatory research and targeted investment.

Among the selected languages in this paper, some have achieved modest digital presence: Talysh, Zazaki, Mazandarani, and Gilaki maintain Wikipedia portals containing between 8,000 and 50,000 articles.[2] However, these varieties lack substantial data resources, especially parallel corpora, which are crucial for modern NLP applications. While biblical translations exist for some varieties, these limited, domain-specific texts are insufficient for developing robust, general-purpose MT systems. Previous work has focused on collecting text material to create monolingual corpora for Zaza-Gorani languages (Ahmadi, 2020) and Southern Kurdish (Ahmadi et al., 2023b), along with tools for language identification (LID) (Ahmadi et al., 2023a). Table 1 provides key information about our selected languages in terms of resources, such as grammars, monolingual corpora and Universal Dependencies (De Marneffe et al., 2021), and a few tools, such as spell checkers and automatic speech recognition. Appendix A provides additional details on the selected languages.

3 PARME

This study addresses a critical gap in MT by developing parallel corpora for low-resourced Middle Eastern languages–PARME, with the ultimate goal of facilitating MT development. As such, data collection lies at the core of our work.

Ideally, data collection would adhere to uniform standards, that is, translating sentences into the standard variety and orthography of the target languages. However, the low-resource languages in our study necessitate a more flexible approach due to fundamental challenges such as:

- The selected languages generally lack widely accepted standard varieties.
- They exhibit significant dialectal variation, often poorly documented and not readily distinguishable even to non-expert native speakers.
- Even when prioritizing a specific dialect, e.g., the dialect of major urban centers, securing sufficient participants remains challenging due to our reliance on volunteer contributors.

Consequently, our methodology is shaped by not only technical constraints but also pragmatic obligations.

3.1 Translation Initiative

Our corpus creation relies on volunteer translators who are native speakers of the target languages with strong writing and translation skills.

[2] As of October 2024.

Given financial constraints, we offered potential co-authorship in this paper to volunteers making substantial contributions. To recruit translators, we leveraged both social media platforms and personal networks to build language-specific translation communities. Each translator was provided with source sentences in both English and another language that the translator is fluent in, accommodating varying levels of English proficiency while enabling potential cross-lingual learning. The translation task was primarily conducted using spreadsheets, though notably, one language activist used Instagram for collecting translations of challenging vocabulary, selecting optimal translations based on community feedback.

Although we initially targeted all marginalized languages in the Middle East, our translation campaign resulted in the 10-week participation of 45 volunteers to translate into eight languages specified in Table A.1, namely **Luri Bakhtiari** (BQI), **Gilaki** (GLK), **Hawrami** (HAC), **Laki** (LKI), **Mazandarani** (MZN), **Southern Kurdish** (SDH), **Talysh** (TLY) and **Zazaki** (ZZA). Additionally, we provided each translator with comprehensive guidelines in a few languages addressing key challenges (summarized in Appendix B).

3.2 Corpus Selection

In low-resource settings with data constraints, selecting the most effective data for training machine translation models is crucial to achieve good performance. Previous studies in domain adaptation and transfer learning have explored various strategies for targeted selection of sentence pairs, such as weighted diversity sampling using n-grams (Ambati et al., 2011), performance-based selection during training (van der Wees et al., 2017), semantic similarity-based selection (Sharami et al., 2021), and fine-tuning with curriculum learning (Mohiuddin et al., 2022). Building upon these findings and considering our resource-constrained conditions, we propose a data selection approach that aims to increase lexical diversity and semantic coverage while leveraging a bilingual corpus in English as the high-resource language, and another language familiar to the translator, e.g., Persian or Turkish.

Our data selection objectives are threefold: (i) maximizing vocabulary coverage of the target language, (ii) ensuring diversity in sentence length, and (iii) enhancing the semantic richness of the selected sentences. To achieve these goals, we employ a data cleaning process that filters out sentences containing named entities, ellipses, or code-switching, and removes less informative cases, such as sentences containing URLs. We believe that this preprocessing step helps to focus on the general vocabulary and reduces noise in the selected data. Data preprocessing is further described in Appendix C.

Formally, given a general-domain large bilingual corpus P containing source sentences s in a language and source translations t in English, we first randomly sample a subset of the corpus as C:

$$C_k = \{(s_i, t_i) \in P \mid \text{valid}(s_i, t_i)\} \quad (1)$$

where $\text{valid}(s_i, t_i)$ ensures compliance with filtering criteria and k refers to the batch number, so $k = 1$ for the first batch. For subsequent batches, we find valid sentences again and compute for each new candidate pair (s'_j, t'_j) in P:

$$\overline{D}_i = \frac{1}{|C_{k-1}|} \sum_{i \in C_{k-1}} \text{Levenshtein}(s_i, s'_j) \quad (2)$$

then, we calculate the semantic similarities based on the source translation t as

$$\overline{\mathcal{S}}_i = \frac{1}{|C_{k-1}|} \sum_{j \in C_{k-1}} \text{cosine}(\mathcal{E}(t_i), \mathcal{E}(t'_j)) \quad (3)$$

where C_{k-1} represents all previously selected pairs and $\mathcal{E}$ represents the semantic similarity function providing sentence embeddings. The diversity score for each candidate pair i is computed as the ratio of edit distance $\overline{D}_i$ to semantic similarity $\overline{\mathcal{S}}_i$ in the embedding space, defined as:

$$\text{score}_i = \frac{\overline{D}_i}{\overline{\mathcal{S}}_i} \quad (4)$$

Finally, we rank all sentence pairs according to their score_i, and then select only the n top-scoring sentence pairs. This formulation rewards sentences with higher edit distances while penalizing those with high semantic similarities. This process continues until the corpus reaches a specific size. In our case, we use Mizan Farsi-English corpus (Kashefi, 2018) given that the translators of the selected languages are familiar with Farsi, the Persian variety spoken in Iran. Our batches are of size 3,000 and the selection continues until it reaches 15,000 sentences. We use sentence embeddings in Sentence-BERT (Reimers and Gurevych, 2019) (`all-MiniLM-L6-v2` model). We also remove punctuation marks in both sentences when calculating the edit distance.

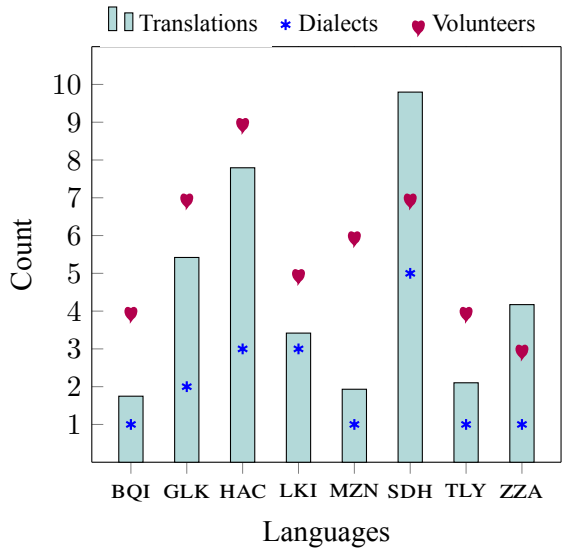

Figure 2: Number of translated sentences ($\times 10^3$) with the number of dialects and volunteers per language. 36,384 sentences are translated by 45 volunteers in eight languages in 10 weeks.

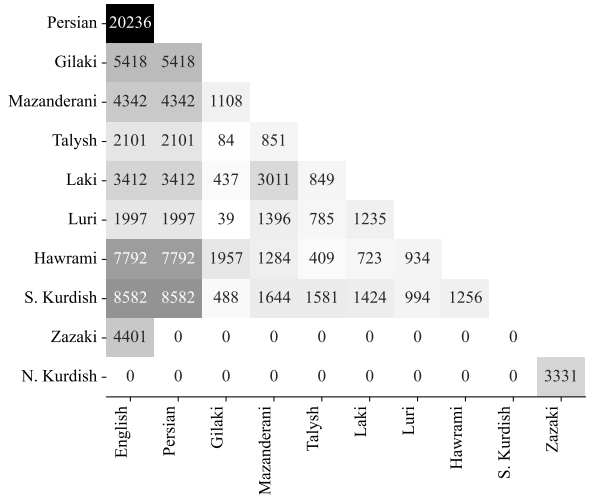

Figure 3: Cross-lingual coverage matrix showing parallel sentence distribution across our selected languages. Each cell indicates the number of shared sentences between language pairs. All translations are parallel with references in English and Persian, except for Zazaki having references in English and Northern Kurdish.

3.3 Quality Control

To ensure corpus quality, translators were permitted to skip sentences containing inconsistencies, spelling errors, code-switching, or named entities. These criteria helped maintain focus on cleanly translatable content while avoiding potential sources of confusion or inconsistency. Our quality control mechanisms included marking empty translations for exclusion from the final corpus and also, requiring translators to maintain consistent orthography throughout their work, and preserving source text as-is, to maintain parallel alignment integrity. Additionally, at least two translators were assigned to check the translations and assess quality at the end of the translation initiative.

3.4 Corpora Statistics

Figure 2 illustrates the size of PARME which overall contains **36,384** translation pairs with Luri Bakhtiari (BQI) and Southern Kurdish (SDH) having the least and the most number of translations, respectively. All the translations are parallel with references in English and Persian, except Zazaki which has references in English and Northern Kurdish. An analysis of length distributions across the languages, shown in Figure C.1 for all the parallel corpora and in Figure C.2 for the test sets, reveals that Persian sentences typically contain more tokens (8-12) than English (4-8), while translations closely follow English token patterns. This systematic difference persists across all languages, suggesting variations in how these languages and their orthographies encode information at the word level. In the case of Zazaki (ZZA), a perfectly overlapping distribution of the reference languages (English and Northern Kurdish), all written in a Latin-based orthography, can be observed. Additionally, we provide a coverage matrix in Figure 3 to show the number of parallel sentences cross-lingually.

3.5 Evaluation Set

Given that rigorous evaluation is crucial for assessing MT performance, creating a representative evaluation test set requires careful considerations. In cases where a standardized form of the language is unavailable, the test set could uniformly represent all major dialects. As such, we set up a methodical approach to dataset splitting with sentences in the test set selected sequentially from the parallel corpora according to the following criteria:

A. data contamination is avoided, i.e., if a sentence is translated into more than one dialect, they are both included in the same split;
B. selected sentences are consistently written in one orthography;
C. for languages lacking a standard form, a balanced distribution of dialects is ensured;
D. and finally, sentences that are translated across the highest number of languages in PARME are prioritized to be included in the test set.

This process yields evaluation sets of approximately 1,000 instances per language. Our selection criteria help reduce some confounding factors in MT evaluation, such as inconsistent orthographies that can affect tokenization. However, since the evaluation set is not fully multi-parallel, cross-lingual comparisons should be interpreted with appropriate caution. Further details on dataset split creation are provided in Appendix C.

4 Methodology

Neural machine translation (NMT) systems traditionally use millions of parallel sentences for effective training from scratch. Given our limited parallel data, we instead adopt a transfer learning approach, leveraging existing models pre-trained on related languages. Recent work has demonstrated the effectiveness of various transfer learning techniques for low-resource machine translation, including parameter-efficient model adaptation (Chronopoulou et al., 2023), fine-tuning on synthetic data (Sant et al., 2024), and large language models (LLMs) (Moslem et al., 2023). Building on these insights, we fine-tune the No Language Left Behind (NLLB) model (Team et al., 2024), a multilingual translation system supporting 203 languages, including languages related to our low-resource languages.

For evaluation, we employ BLEU (Papineni et al., 2002) which calculates n-gram precision with a brevity penalty to account for translation length. We also use chrF (Popović, 2015) for our baseline system which computes character-level n-gram F-scores, offering sensitivity to both word-level accuracy and morphological variations in the translations. We use the SacreBLEU implementation (Post, 2018) for both metrics.[3] While more recent metrics such as COMET (Rei et al., 2020) and its reference-free variant COMET-QE (Kocmi et al., 2022) have shown promise, the extreme data scarcity in our target languages precludes their use.

4.1 Baselines

We evaluate NLLB's zero-shot performance using two model variants: `3.3B` and its `600M` distilled variant. Since NLLB's training data does not include language indicator tokens for our selected languages, we assess the model's performance by leveraging linguistically and geographically proximate languages supported in NLLB. These include Arabic (ARA), Turkish (TUR), Central Kurdish (CKB), Northern Kurdish (KMR), Farsi (PES) and English (ENG).

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

Language	New Token	Initialization
Luri Bakhtiari (BQI)	`bqi_Arab`	Farsi
Gilaki (GLK)	`glk_Arab`	Farsi
Hawrami (HAC)	`hac_Arab`	C. Kurdish
Laki (LKI)	`lki_Arab`	C. Kurdish
Mazandarani (MZN)	`mzn_Arab`	Farsi
Southern Kurdish (SDH)	`sdh_Arab`	C. Kurdish
Talysh (TLY)	`tly_Arab`	Farsi
Zazaki (ZZA)	`zza_Latn`	N. Kurdish

Table 2: Initialization of our fine-tuned model for the selected languages by adding new tokens to NLLB.

Transliteration To mitigate the impact of divergent writing systems (Perso-Arabic-based and Latin-based scripts), we experiment with a transliteration pipeline as well. For Southern Kurdish, Laki, Hawrami, and Zazaki varieties, we extend the rule-based transliteration system developed by Ahmadi (2019) with variety-specific character mappings, e.g., <ۆ> to <ö>. For the remaining languages, we employ phonetic-based character mapping, converting each character to its closest phonological equivalent in the Latin script, e.g., <ۆ> to <o>.

4.2 Fine-tuning

Aiming to fine-tune a single multilingual model for our selected languages, we extend NLLB's language support by leveraging embeddings from linguistically similar languages in a systematic token-based approach. The extension process consists of two key steps. First, we expand the tokenizer's vocabulary by adding new language-specific tokens, e.g., `sdh_Arab` for Southern Kurdish, while preserving the existing token structure. Second, we initialize each new language token's embedding by cloning the embeddings from closely-related languages provided in Table 2, e.g., Southern Kurdish is initialized based on Central Kurdish embeddings. This initialization strategy leverages the model's pre-existing knowledge of those related languages, particularly benefiting from the writing systems. This way, our initialized fine-tuned models efficiently extend the model's multilingual capabilities without requiring extensive architectural changes or complete retraining, making it suitable for incorporating low-resource languages.

4.3 Experimental Setup

Datasets We explore two experimental setups:

- **Base Setup** ($\mathcal{M}_{\text{base}}$): We utilize the train sets, each containing sentences paired with English translations. For languages lacking dedicated training data, namely (BQI and TLY), we re-purpose their validation sets as training data due to resource constraints.
- **Augmented Setup** ($\mathcal{M}_{\text{aug}}$): Given that our parallel corpora include translations in languages besides English, we leverage these additional translations to expand our training data. Specifically, we use NLLB to translate sentences into English from Northern Kurdish for Zazaki and from Farsi for all other languages.

For fine-tuning NLLB, we merge all parallel sentences and convert to JSON Lines format. Each entry contains source sentences with our newly introduced tokens paired with their English translations, specified by `eng_Latn`. The distribution of data across both setups is illustrated in Figure C.3.

Implementation Focused on X→English translation direction, we implement our fine-tuning pipeline using the NLLB tokenizer from Huggingface Transformers. Our implementation builds upon the NLLB architecture, which is based on the M2M100 encoder-decoder architecture (Fan et al., 2021). Given the multilingual context, it is important to handle multilingual tokenization by explicitly setting the source language token while fixing English (`eng_Latn`) as the target language. Based on the language token provided in the train set, the source language is specified during preprocessing and tokenization ensuring proper handling of multiple source languages.

Hyper-parameters We explore three main configurations for fine-tuning:

- CONFIG I: Initial setup with a batch size of 8, low learning rate of 3e-5, and 20 epochs
- CONFIG II: Enhanced setup with a batch size of 16, increased learning rate of 5e-4, and 50 epochs
- CONFIG III: maintaining CONFIG II's parameters but training for 100 epochs

More details on the hyper-parameters and training are provided in Appendix D.

5 Experiments

5.1 Quantitative Analysis

Baseline Initial evaluation of NLLB's zero-shot capabilities using proximate languages demonstrates significant limitations, with BLEU scores ranging from 0.2 to 1.98. When averaging across all test sets, using the Farsi language indicator token results in the highest BLEU score (1.39), followed by Central Kurdish (1.215). This poor performance highlights the substantial gap in current multilingual models' ability to handle these underresourced varieties. The choice of language indicator token only has a small effect, with differences explainable by language similarity: Central Kurdish token yielding marginally better results for Southern Kurdish and Hawrami input, Farsi performs best for Luri Bakhtiari, Gilaki, Laki, Mazandarani, and Talysh, and Northern Kurdish works best for Zazaki–likely reflecting similarities in orthographies and lexical overlap. Notably, transliterating the input text to match the script of the indicated language generally deteriorates performance. The highest BLEU scores per language in the baselines are presented under the $\text{Baseline}_{\text{BEST}}$ in Table 4 with the complete baseline results provided in Appendix E.

Fine-tuning Fine-tuning results in Table 3 demonstrate substantial improvements across different configurations and data setups. Our initial configuration (CONFIG I) achieves an average BLEU score of 5.41, which improves remarkably to 7.40 in CONFIG II. The mean BLEU scores of the base setup ($\mathcal{M}_{\text{base}}$) mostly outperform the augmented setup ($\mathcal{M}_{\text{aug}}$), suggesting that the quality of the training data is more important than quantity for these languages. We observe that increasing the number of epochs (from 20 in CONFIG I to 100 in CONFIG III) and adjusting learning rates and warmup ratios positively impacts performance. Further fine-tuning our best model ($\mathcal{M}_{\text{base}}$ based on CONFIG II) for an additional 50 epochs yields the highest average BLEU score of 7.50 in CONFIG +, henceforth, referred to as $\mathcal{M}_{\text{best}}$.

Looking at individual languages, we find that the $\mathcal{M}_{\text{aug}}$ setup particularly benefits languages closely related to Farsi, while Hawrami, Laki, and Southern Kurdish achieve the most promising results with BLEU scores exceeding 10 points. Notably, Zazaki reaches its peak performance of 6.19 BLEU points with the augmented setup.

	CONFIG I		CONFIG II		CONFIG III		CONFIG +
	$\mathcal{M}_{\text{base}}$	$\mathcal{M}_{\text{aug}}$	$\mathcal{M}_{\text{base}}$	$\mathcal{M}_{\text{aug}}$	$\mathcal{M}_{\text{base}}$	$\mathcal{M}_{\text{aug}}$	$\mathcal{M}_{\text{base}}$
BQI	4.38	**4.66**	3.40	4.22	3.01	4.58	3.25
GLK	2.73	3.54	3.47	**4.14**	2.46	3.86	3.67
HAC	8.23	8.63	16.41	12.31	14.46	9.48	**16.54**
LKI	6.33	5.43	9.75	6.79	10.03	6.61	**10.31**
MZN	5.23	5.31	5.30	**5.66**	4.53	5.66	5.51
SDH	9.93	9.85	**11.46**	9.09	10.22	9.81	10.84
TLY	3.01	3.35	6.23	6.44	6.82	3.62	**6.98**
ZZA	3.45	3.81	3.17	4.11	2.20	**6.19**	2.89
Avg	5.41	5.57	7.40	6.59	6.72	6.23	**7.50**

Table 3: BLEU scores across different hyper-parameter configurations and dataset setups. While $\mathcal{M}_{\text{aug}}$ dataset setup is beneficial for languages closely-related to Farsi, the $\mathcal{M}_{\text{base}}$ setup outperforms the others.

To evaluate the significance of our results, we also conduct paired bootstrap resampling (Koehn, 2004) using Sacrebleu with the following specific comparisons: (1) each augmented model against its corresponding base model within the same configuration; (2) CONFIG III against CONFIG I; and (3) CONFIG II against both CONFIG I and CONFIG III. Our results conclusively demonstrate that CONFIG II significantly outperforms both CONFIG I and CONFIG III ($p < 0.001$). The comparison between augmented and base models revealed that augmentation significantly decreased performance in CONFIG II and CONFIG III ($p < 0.001$ and $p = 0.012$).

5.2 Qualitative Analysis

Analyzing the quality of translations in our best performing model, presented in Table E.2 in appendix, reveals several interesting patterns in the model's translation behavior:

- First, we observe that the generated translations are generally grammatically well-formed and capture the core meaning of the source sentences. However, the model sometimes fails to select precise word translations, as illustrated by the Southern Kurdish word پرسه (*pirse*) meaning 'funeral' being incorrectly confused with پرس (*pirs*) as 'question'.
- Second, we note inconsistencies between Farsi and English reference translations. For instance, in the Southern Kurdish example, while the source text refers to a 'funeral', the Farsi reference uses واقعه 'event'. Since translators may have translated similar instances from the Farsi references, this mismatch could explain translations not perfectly matching the English references.

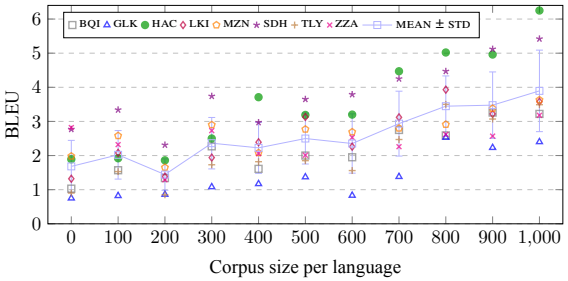

Figure 4: BLEU scores with varying amount of training data. With only 1000 samples per language, an average BLEU score of 3.89 is achieved vs. 1.68 in zero-shot evaluation. The mean BLEU score of our top running model is 7.50.

- Finally, our fine-tuned model sometimes generates more sophisticated vocabulary than the references (e.g., 'disagreeable' instead of 'unwell' in Gilaki), resulting in lower BLEU scores despite potentially acceptable translations. Trainable metrics might better capture such semantic equivalences that n-gram based metrics miss.

5.3 Ablation

As an ablation study, we address the impact of two factors in two questions:

Q1: How does corpus size impact translation quality? We investigate the relationship between training data size and model performance in our low-resource setting by incrementally sampling sentences from our training sets. Starting from zero, we gradually increase the sample size up to 1,000 sentences per language—representing less than half of the training data for BQI and TLY, and less than one-third for other languages. Following our initial methodology, we fine-tune NLLB on these limited samples.

Our analysis, presented in Figure 4, reveals several key patterns. The mean of the highest BLEU scores per language in the zero-shot baselines (with no language-specific training data) achieves 1.68, while training with 1,000 sentences per language yields 3.89 BLEU. We observe that performance improvements are not consistently proportional to data volume: most languages show fluctuating performance up to 600 samples, suggesting a period of model instability, followed by more substantial improvements, particularly for Hawrami and Southern Kurdish. In fact, Hawrami and Southern Kurdish are two of the best-performing languages, and also the two languages with the most training data. Even if we control for the

	$\mathcal{M}_{1000}^{-\text{GLK}}$	$\mathcal{M}_{1000}^{-\text{MZN}}$	$\mathcal{M}_{1000}^{-\text{TLY}}$	$\mathcal{M}_{1000}^{-\text{HAC}}$	$\mathcal{M}_{1000}^{-\text{LKI}}$	$\mathcal{M}_{1000}^{-\text{SDH}}$	$\mathcal{M}_{1000}^{-\text{BQI}}$	$\mathcal{M}_{1000}^{-\text{ZZA}}$	$\text{Baseline}_{\text{BEST}}$	$\mathcal{M}_{1000}$	$\mathcal{M}_{\text{BEST}}$
GLK	**1.44**	2.39	**1.98**	2.37	2.45	2.47	2.17	**2.99**	0.75	2.40	3.67
MZN	**3.23**	**2.05**	**3.23**	3.95	3.46	3.72	3.31	**3.98**	1.98	3.63	5.51
TLY	3.18	3.22	**1.37**	3.48	**2.88**	3.16	3.28	**3.79**	0.90	3.49	6.98
HAC	**5.65**	5.48	5.57	**3.19**	5.59	4.88	**4.45**	5.39	1.89	6.25	16.54
LKI	2.88	**3.27**	3.43	**3.92**	**2.83**	3.78	3.66	3.67	1.32	3.58	10.31
SDH	**5.34**	4.89	5.12	5.11	5.12	**3.68**	**4.71**	4.66	2.77	5.42	10.84
BQI	2.83	3.24	2.91	**2.73**	3.37	**3.39**	**1.39**	3.28	1.03	3.22	3.25
ZZA	**2.73**	3.15	2.99	2.67	**3.39**	2.83	2.95	**0.44**	2.82	3.18	2.89
Avg	3.41	3.46	3.33	3.43	3.64	3.49	3.24	3.53	1.68	3.89	7.50

Table 4: BLEU scores of models excluding a language from the fine-tuning data, e.g., $\mathcal{M}_{1000}^{-\text{GLK}}$ excludes Gilaki. The worst, the second worst and best models are respectively specified in red, orange and greenish-blue.

amount of training data, they tend to be the languages with the highest BLEU scores, which indicates positive cross-lingual transfer from other languages in the base NLLB model.

There is a clear gap between models trained on 1,000 samples (3.89 BLEU) and those trained on the full dataset (shown in Table 3, ranging from 5.41 to 7.50 BLEU), with differences of 1.52 BLEU points (5.41 - 3.89) to 3.61 BLEU points (7.50 - 3.89). While significant, this improvement is less dramatic than might be expected given the substantial increase in training data volume, suggesting that while data size is important, factors such as data quality and hyperparameters also play crucial roles in determining overall performance.

Q2: How does excluding individual languages affect cross-lingual performance? We examine cross-lingual transfer effects by training a series of models on 1,000 samples per language while systematically excluding individual languages, e.g., $\mathcal{M}_{1000}^{-\text{GLK}}$ represents a model trained without any Gilaki data. We evaluate each model's performance across test sets, with results presented in Table 4.

Our analysis reveals several intriguing patterns. The mean BLEU score remains relatively stable at approximately 3.50 across different configurations. As anticipated, excluding a language's data typically impacts its own translation performance most significantly (highlighted in red along the diagonal and the second worst, in orange). However, we observe unexpected beneficial effects when certain languages are excluded. For example, the translation quality for Gilaki, Mazandarani, and Talysh improves notably when Zazaki is excluded from training (specified in greenish-blue). The results also illuminate the role of linguistic proximity. Closely related languages generally demonstrate strong interdependence, exemplified by how the exclusion of Talysh negatively affects performance on Gilaki and Mazandarani. Conversely, we observe a counterintuitive pattern where Laki's performance improves most significantly when Hawrami is excluded, despite their close linguistic relationship. This unexpected finding suggests potential interference between similar linguistic features, orthographic differences, or data quality imbalances between these varieties, a phenomenon that warrants further investigation.

6 Conclusion

This paper presents a systematic approach to developing parallel corpora for eight severely under-resourced languages of the Middle East. Our methodology, which leverages native speaker volunteers for translating carefully curated sentences, has resulted in the creation of trilingual corpora comprising 36,384 translation pairs. These resources represent the first publicly available parallel datasets for several of these languages, contributing not only to the advancement of language technologies and MT specifically, but also enabling broader cross-linguistic research. We conduct a comprehensive evaluation of these corpora using NLLB, a state-of-the-art NMT system. The baseline results reveal significant challenges, with X→English BLEU scores below 1 for multiple languages, highlighting the complexity of the task. Through extensive fine-tuning experiments on various multilingual models, we investigate the impact of different experimental configurations on translation performance. Our best model achieves a mean X→English BLEU score of 7.50, with Hawrami having the highest performance at 16.54, demonstrating the potential for improving translation quality in these under-resourced languages. We believe that our findings open new avenues for research in low-resource NLP.

Limitations The limitations of this work open several promising research avenues. While we focus on X→English translation direction, the reverse English→X direction remains unexplored. Our preliminary experiments on English→X translation following our same fine-tuning methodology indicates poor performance on NMT. The relatively modest size of our parallel corpora, though unprecedented for these languages, inherently constrains NMT performance. This limitation could be addressed through data augmentation techniques leveraging LLMs or back-translation approaches (Sennrich et al., 2016). The limited availability of native speakers prevents exploration of alternative corpus selection strategies and a more comprehensive translation validation and assessment based on different dialects and orthographies. Furthermore, despite implementing rigorous quality control measures, some translations may remain suboptimal due to the scarcity of qualified bilingual speakers in these low-resource languages. A systematic investigation of how dialectal variation and competing orthographic standards impact translation quality remains an important direction for future work, especially given the ongoing standardization efforts in several of these languages.

Ethics Statement Our research prioritizes ethical considerations in both data collection and community engagement. The parallel corpora underwent screening to ensure the removal of all personal information and sensitive content, maintaining privacy and data security standards. A cornerstone of this work is the invaluable contribution of over 40 volunteers across eight languages. Given the absence of financial support, we established a clear framework emphasizing that these resources and models would serve the broader community, with full acknowledgment of contributors' efforts. The mobilization of contributors presented unique challenges within the complex geopolitical landscape of the Middle East. We addressed initial concerns through extensive dialogue with volunteers and community stakeholders, clearly articulating the project's scope as an open-source initiative designed to elevate the status of these languages and advance technological accessibility for their communities. Our commitment to ethical practices extended to offering co-authorship to volunteers, recognizing their crucial role in preserving and advancing their native languages with technology.

Acknowledgements

This work was supported by the Swiss National Science Foundation through the MUTAMUR project (no. 213976). We gratefully acknowledge the support of the Stanford Initiative on Language Inclusion and Conservation in Old and New Media (SILICON) at Stanford University, with particular thanks to Thomas S. Mullaney, Audrey Gao, and the entire SILICON team for their valuable support. We extend our sincere appreciation to the numerous individuals who contributed to the success of this initiative, including those who helped promote the project across social media platforms. Special recognition goes to Milad Mirzaei, Razhan Hameed, Özcan Yilmaz, Ferhat Aydın, Rahim Gholamvaisi and Darya Valadbeigi for their contributions. We are deeply grateful to the reviewers for their constructive feedback on this paper as well as authors, publishers, and volunteers whose dedication made this work possible.

References

Samir Abu Absi. 1981. Language-in-education in the Arab Middle East. *Annual review of applied linguistics*, 2:129–143.

Sina Ahmadi. 2019. A rule-based Kurdish text transliteration system. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–8.

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

Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023a. PALI: a language identification benchmark for Perso-Arabic scripts. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 78–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Sina Ahmadi and Antonios Anastasopoulos. 2023. Script normalization for unconventional writing of under-resourced languages in bilingual communities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14466–14487, Toronto, Canada. Association for Computational Linguistics.

Sina Ahmadi, Zahra Azin, Sara Belelli, and Antonios Anastasopoulos. 2023b. 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.

Vamshi Ambati, Stephan Vogel, and Jaime G. Carbonell. 2011. Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of Machine Translation Summit XIII: Papers, MTSummit 2011, Xiamen, China, September 19-23, 2011*.

Erik Anonby and Ashraf Asadi. 2014. *Bakhtiari studies: Phonology, text, lexicon*. Acta Universitatis Upsaliensis.

Erik Anonby and Mortaza Taheri-Ardali. 2018. Bakhtiari. *The Languages and Linguistics of Western Asia: An Areal Perspective*, 6:445.

Erik John Anonby. 2003. Update on Luri: How many languages? *Journal of the Royal Asiatic Society*, 13(2):171–197.

Victoria Arakelova. 2022. The Talishis on opposite banks of the Araxes river: Identity issues. *Iran and the Caucasus*, 26(4):407 – 417.

Sevda Arslan. 2019. Language, religion, and emplacement of Zazaki speakers. *Journal of Ethnic and Cultural Studies*, 6(2):11–22.

Zeinab Ashouri Nejad. 2024. Thematic and comparative dictionary of Southern Taleshi. Master's thesis, Faculty of Literature and Humanities, University of Guilan.

Leila Rahimi Bahmany. 2024. Monolingualism in Iran: The Politics of Writing in Azeri Turkish. *Iranian Studies*, 57(2):329–334.

Hassan Bashirnezhad. 2018. A socio-linguistic analysis on the status and usage of Mazandarani and Persian in Mazandaran. *Zabanpazhuhi (Journal of Language Research)*, 10(27):119–145.

Hassan Bashirnezhad. 2023. Mazandarani: Current status and future prospects. *Iranian and Minority Languages at Home and in Diaspora*, 1:37.

Sara Belelli. 2022. *The Laki variety of Harsin: grammar, texts, lexicon*. University of Bamberg Press.

Sara Belelli et al. 2019. Towards a dialectology of Southern Kurdish: Where to begin. *Current issues in Kurdish linguistics*, 1:73.

Îbrahim Bingöl. 2020. *Rêzimana Zazakî ya devoka Gimgimê*. Lêkolînên ziman. Avesta.

Steven Bird. 2020. Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020*, pages 3504–3519. Association for Computational Linguistics (ACL).

Habib Borjian. 2004. Māzandarān: Language and people (the state of research). *Iran & the Caucasus*, pages 289–328.

Habib Borjian. 2019. Mazandarani: A typological survey. *Typology of Iranian languages*, pages 79–101.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diané, Koulako Moussa Doumbouya, Djibrila Diané, and Solo Farabado Cissé. 2025. SMOL: professionally translated parallel data for 115 under-represented languages. *CoRR*, abs/2502.12301.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

David Crystal. 2002. *Language death*. Cambridge University Press.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Ipek Demir. 2017. Shedding an ethnic identity in diaspora: de-Turkification and the transnational discursive struggles of the Kurdish diaspora. *Critical Discourse Studies*, 14(3):276–291.

Stanley Dubinsky and Harvey Starr. 2022. Weaponizing language: Linguistic vectors of ethnic oppression. *Global Studies Quarterly*, 2(2).

Ethnologue. 2023. Ethnologue (online). Dallas, Tex. Summer Institute of Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Ismaïl Kamandâr Fattah. 2000. *Les dialectes kurdes méridionaux: étude linguistique et dialectologique*. Acta Iranica: Encyclopédie permanente des études iraniennes. Peeters.

Desmond Fernandes. 2012. Modernity and the linguistic genocide of Kurds in Turkey. *International Journal of the Sociology of Language*, 2012(217):75–98.

Behrouz Ghesmatpour, Ali Reza Gholi Famian, Seifollah Molaei Pashaei, and Narges Banoo Sabouri. 2020. Computational dialectometry of variations in Taleshi language varieties in the Southwestern

Caspian Coast. *Semi-Annual Journal of Linguistic Science*, undefined(undefined).

Esmaeil Haddadian-Moghaddam and Reine Meylaerts. 2015. What about translation? Beyond "Persianization" as the language policy in Iran. *Iranian Studies*, 48(6):851–870.

Joost Jongerden. 2007. *The settlement issue in Turkey and the Kurds: An analysis of spatial policies, modernity and war*, volume 102. Brill.

Shuan Osman Karim and Saloumeh Gholami, editors. 2024. *Gorani in its Historical and Linguistic Context*. De Gruyter Mouton, Berlin, Boston.

Omid Kashefi. 2018. MIZAN: A large Persian-English parallel corpus. *CoRR*, abs/1801.02107.

Zia Khoshsirat. 2018. The origin of the Gilaki causative suffix -be(:)-. Master's thesis, University of Kentucky. Available since 25 July 2018.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

David Lockhart Robertson Lorimer. 1922. *The Phonology of the Bakhtiari, Badakhshani, and Madaglashti dialects of Modern Persian, with vocabularies*, volume 6. Royal Asiatic Society.

Parvin Mahmoudveysi and Denise Bailey. 2018. Hawrāmī of Western Iran. *The World of Linguistics*, page 533.

Philip McDermott and Máiréad Nic Craith. 2019. Linguistic recognition in deeply divided societies: Antagonism or reconciliation? *The Palgrave Handbook of Minority Languages and Communities*, pages 159–179.

Catherine Miller. 2003. Linguistic policies and the issue of ethno-linguistic minorities in the Middle East. *Islam in the Middle Eastern studies: Muslims and minorities*, pages 149–174.

Seyyed-Abdolhamid Mirhosseini. 2015. Loving but not living the vernacular: A glimpse of Mazandarani-Farsi linguistic culture in Northern Iran. *Language Problems and Language Planning*, 39(2):154–170.

Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale, and Shafiq R. Joty. 2022. Data selection curriculum for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1569–1582. Association for Computational Linguistics.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Sietze Norder, Laura Becker, Hedvig Skirgård, Leonardo Arias, Alena Witzlack-Makarevich, and Rik van Gijn. 2022. `glottospace`: R package for the geospatial analysis of linguistic and cultural data. *Journal of Open Source Software*, 7(77):4303.

Fatih Ozek, Bilgit Saglam, and Charlotte Gooskens. 2023. Mutual intelligibility of a Kurmanji and a Zazaki dialect spoken in the province of Elazığ, Turkey. *Applied Linguistics Review*, 14(5):1411–1449.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Lisbon, Portugal. 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.

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

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.

Jahandoust Sabzalipour. 2020. *Sarkhat: Proposed Guidelines for the Alphabet and Gilaki Writing System*. Gilan Studies Research Institute, University of Guilan. Approved by a group of researchers.

Esat Şanlı. 2022. The potentials and challenges of Zazaki translation for language revitalisation. *Kurdish Studies Archive: Vol. 10 No. 2 2022*, page 141.

Aleix Sant, Daniel Bardanca, José Ramom Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier Garcia Gilabert, Pablo Gamallo, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Training and fine-tuning NMT models for low-resource languages using Apertium-based synthetic corpora. In *Proceedings of the Ninth Conference on Machine Translation*, pages 925–933, Miami, Florida, USA. 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Mansour Shabani. 2021. The distinction between language and dialect: Exploring the state of Gilaki variety. volume 12, pages 121–148. Alzahra University.

Pourmostafa Roshan Javad Sharami, Dimitar Sterionov, and Pieter Spronck. 2021. Selecting parallel in-domain sentences for neural machine translation using monolingual texts. *Computational Linguistics in the Netherlands Journal*, 11:213–230.

Jaffer Sheyholislami and Rachelle Vessey. 2024. Language policy and discourse in the public sphere: the discursive construction of language and multilingualism as policy objects. In *The Routledge Handbook of Language Policy and Planning*, pages 201–212. Routledge.

Karli Storm. 2024. "diffuse support" and authoritarian regime resilience: Azerbaijanism vis-à-vis Azerbaijan's Talysh minority. *Caucasus Survey*, 1(aop):1–26.

Karli-Jo Storm. 2023. Discursive-technical landscaping and policing the body (politic) in Azerbaijan: a case study of Talysh activists. *Geografiska Annaler: Series B, Human Geography*, pages 1–20.

Sahar Taghipour. 2024. *Case and Phi-Agreement in Laki: Parametrizing Split-Ergativity in Kurdish*. Ph.D. thesis, University of Toronto.

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

Terry Lynn Todd. 1985. *A grammar of Dimili (also known as Zaza)*. Ph.D. thesis, University of Michigan.

Uğur Ümit Üngör. 2012. Untying the tongue-tied: Ethnocide and language politics. *International journal of the sociology of language*, 2012(217):127–150.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Gernot Windfuhr. 1988. Baktīārī tribe: The Baktīārī dialect. *Encyclopædia Iranica*, 3(5):553–60.

Grûba Xebate ya Vateyî. 2009. *Ferhengê Kirmanckî (Zazakî)-Tirkî: Kırmancca (Zazaca)-Türkçe Sözlük*. Weşanxaneyê Vateyî.

Sedighe Zamani Roodsari. 2023. Gilaki: From language regimes into minoritizing. *Studies in Linguistics and Literature*, 7:2023.

Sima Zolfaghari. 2023. The Bakhtiari language: Maintenance or shift? a diachronic survey on the status of Bakhtiari in the city of Masjed Soleiman between 1996–2020. *Iranian and Minority Languages at Home and in Diaspora*, 1:111.

A Selected Languages

Language (ISO 639-3)	Country	Population
Luri Bakhtiari (BQI)	Iran	4-5 M
Gilaki (GLK)	Iran	3-4 M
Hawrami (HAC)	Iran / Iraq	0.3 M
Laki (LKI)	Iran / Iraq	0.68 M
Mazandarani (MZN)	Iran	4-5 M
Southern Kurdish (SDH)	Iran / Iraq	4-5 M
Talysh (TLY)	Iran / Azerbaijan	1-2 M
Zazaki (ZZA)	Turkey	1-2 M

Table A.1: Low-resourced Middle Eastern languages included in our parallel corpora. Population estimates are based on (Ethnologue, 2023; Arakelova, 2022).

The languages addressed in this paper represent severely under-resourced linguistic varieties that have remained largely unexplored in computational linguistics and language technology. These languages face multiple challenges: absence of standardized forms, fragmented writing systems, and limited documentation in linguistic literature. Many exist in complex sociolinguistic environments where multiple dialects and orthographic conventions coexist. Beyond these, they face existential threats through systematic sociolinguistic discrimination, endangering their survival and intergenerational transmission. Below, we provide a brief overview of each language variety included in our study.

A.1 Luri Bakhtiari

Luri Bakhtiari (BQI in ISO 693-3, also referred to as Bakhtiari) is spoken by a population exceeding one million (estimated far more) speakers across southwestern Iran, primarily distributed across the provinces of Lorestan, Chaharmahal and Bakhtiari, Khuzestan, and Isfahan, with significant populations in urban centers including Masjedsoleyman, Shahr-e Kord, Dorud, and Aligudarz (Anonby and Taheri-Ardali, 2018). The traditional Bakhtiari society has been characterized by long-distance nomadic patterns, with seasonal migrations between summer and winter residences (Windfuhr, 1988). Despite significant cultural shifts due to modernization, including forced sedentarization and formal education in Persian (Anonby and Asadi, 2014, p. 15) causing domain-specific language shift and intergenerational language loss, the Bakhtiari language has endured as a vital cultural element. The cultural and linguistic heritage of Bakhtiari, known as "Loriyāti" among native speakers, remains integral to community identity.

Bakhtiari belongs to the Southwestern branch of West Iranian languages within the Indo-European family. Linguistically, it is found in the Luri continuum between Persian and Kurdish, with particularly close structural affinities to Southern Luri (Anonby, 2003). While historically sometimes classified as a Persian dialect (Lorimer, 1922), the Bakhtiari language exhibits distinctive features in phonology, morphosyntax, and lexicon that differentiate it from regional languages including Northern or Southern Luri, Persian, and Kurdish (Anonby and Asadi, 2014). Dialect variation within Bakhtiari is traditionally categorized according to geographical distribution. Additionally, urban migration, particularly to metropolitan areas such as Ahvaz, has contributed to dialectal diversification (Zolfaghari, 2023).

Writing System As primarily an oral language, Bakhtiari has historically utilized the Perso-Arabic script. Documentation of written Bakhtiari is sparse, limited to select excerpts in grammar books, interlinear glosses (Anonby and Taheri-Ardali, 2018), and reported text samples (Anonby and Asadi, 2014, p.91). Recent standardization efforts have emerged, notably the Pāpêrik orthography, which augments the Perso-Arabic script with additional graphemes. While this system is supported by a dedicated keyboard layout, its adoption remains limited within the community.

A.2 Gilaki

Gilaki is an Indo-European language belonging to the Caspian subgroup of the Northwestern Iranian branch, with a speaker population exceeding 4 million Gilaks distributed along the southern Caspian Sea, chiefly in Gilan. While sharing linguistic features with related languages including Mazandarani and Talyshi, as well as Persian, Gilaki exhibits distinctive characteristics in its morphology, syntax, vocabulary, and phonological system. Gilaki is divided into four varieties:

- Western Gilaki spoken in western Gilan, e.g. Fuman (فۆمن)
- Central Gilaki spoken in Rasht (رشت) – the capital of the Gilan province and around
- Eastern Gilaki spoken in eastern Gilan, primarily in Lahijan (لاجۆن)
- Southern Gilaki in other southern regions, as in Rudbarat (رۆدبارات)

For a detailed discussion of Gilaki dialect classification and nomenclature, see (Khoshsirat, 2018, p. 56).

Writing System Gilaki has a documented writing tradition spanning over five centuries, though contemporary written usage remains limited. Current orthographic standardization efforts have produced two notable systems: Sarkhat (Sabzalipour, 2020), which prioritizes compatibility with Persian orthography, and Vrg[4], which introduces specific graphemes for Gilaki phonemes. Both systems are being actively promoted through online educational content, with growing communities of users contributing written materials. While recent initiatives have attempted to introduce Latin-based script alternatives, particularly on Wikipedia[5], these efforts have not gained sustained traction.

A.3 Hawrami

Hawrami (also referred to as Gorani) is an Indo-European language with approximately 300,000 speakers (estimated far more) residing in the Hawraman region, which spans the border between Iran and Iraqi Kurdistan (Mahmoudveysi and Bailey, 2018). The language is primarily spoken in a mountainous area encompassing parts of Kurdistan and Kermanshah provinces (Iran) and Halabja (Iraq). Despite its relatively small speaker population, Hawrami has maintained a distinct linguistic identity and has historically been significant in the region's literary traditions.

Due to the complex linguistic and cultural interactions between Kurdish and Gorani communities and their shared Kurdish identity, Hawrami has occasionally been classified as a Kurdish dialect. However, it is more commonly categorized within the Gorani language group, which, together with Zazaki and Shabaki, constitutes the Zaza-Gorani family within the Northwestern Iranian languages. The precise classification remains a subject of ongoing scholarly debate, as detailed in (Karim and Gholami, 2024).

Writing System Hawrami utilizes the Kurdish Perso-Arabic script and follows its orthographic conventions. However, it incorporates several distinctive graphemes that are not found in Kurdish orthography, including <ڎ> (U+068E) and <ۆ> (U+06CB).

[4] https://v6rg.com

[5] https://glk.wikipedia.org

A.4 Laki Kurdish

Laki is primarily spoken in the Kermanshah and Lorestan provinces of Iran, with additional speaker communities in Chamchamal and Erbil regions of Iraqi Kurdistan. The linguistic classification of Laki presents unique challenges due to its historical conflation with Southern Kurdish. However, recent research supports its unique features that can be defined as a distinct language, despite sharing characteristics with both Kurdish and Lurish languages (Belelli, 2022). Laki exhibits interesting linguistic constructions, notably ergativity, contributing to its unique typological profile (Taghipour, 2024).

Writing System Laki employs the Kurdish Perso-Arabic script, distinguished by the characteristic grapheme <ۏ> (U+06CF).

A.5 Mazandarani

Mazandarani (also known as Mazanderani, Mazani, or Tabari) is a northwestern Iranian language spoken primarily along the eastern Caspian coastline in Iran, predominantly in Mazandaran Province (Borjian, 2004). With approximately four million speakers, it belongs to the Caspian language group and shares significant linguistic features with neighboring languages, particularly Gilaki (Borjian, 2019; Mirhosseini, 2015). Recent sociolinguistic studies indicate a concerning decline in language use, with younger generations increasingly favoring Persian (Bashirnezhad, 2018, 2023).

Writing System Mazandarani has a documented literary tradition extending over five centuries. Its writing system adopts the Perso-Arabic script with minor modifications from Persian orthography, notably the distinctive diacritic <ۜ> (U+02C7). It also has its Wikipedia portal.[6]

A.6 Southern Kurdish

Southern Kurdish, despite being spoken by over 4 million people, has received significantly less academic and technological attention compared to other Kurdish varieties, particularly Central and Northern Kurdish (Ahmadi et al., 2023b). The seminal work of Fattah (2000) provides a comprehensive dialectological classification of Southern Kurdish varieties predominantly spoken in Iran. These varieties include Garusi spoken mainly in

[6] https://mzn.wikipedia.org

Bijar (Kurdistan province), Pehley (also known as Feyli), Badrei and Khezli spoken in Ilam province, and Kolyai, Sanjabi, Krmashani (Kirmaşanî), and Kalhuri spoken in Kermanshah province, with the latter also extending into Iraqi Kurdistan, mainly spoken in the Khanaqin district. In our computational resource development efforts, Southern Kurdish presented unique challenges due to its substantial dialectal variation, a complexity well-documented in linguistic and dialectological research (Belelli et al., 2019).

Writing System Southern Kurdish employs the Kurdish Perso-Arabic script, distinguished by the unique grapheme <ۆ> (U+06CA).

A.7 Talysh

Talysh is spoken along the southwestern Caspian coast, divided between Azerbaijan and Iran by the Araxes River (Arakelova, 2022). Despite some influence from Persian and local Turkic dialects, Talysh has retained many unique linguistic features. It has three main dialects–Southern and Central spoken in Iran, and Northern spoken in Azerbaijan–with relatively minor differences due to the historically compact settlement of its speakers (Storm, 2024). Recent work has focused on dialectology and comparative study of Talysh (Ashouri Nejad, 2024; Ghesmatpour et al., 2020). While exact speaker numbers are disputed, conservative estimates suggest around 1 million Talysh speakers in Azerbaijan (though official statistics claim only 112,000) and approximately 700,000 in Iran. Talysh has faced aggressive assimilation policies, with activists facing persecution for promoting their language and culture (Storm, 2023). Recent years have seen increased activism around Talysh language rights and cultural identity in both countries.

Writing System Northern Talysh in Azerbaijan uses a standardized Latin-based script, enabling digital literacy through platforms like Wikipedia[7] and online publications. On the other hand, Talysh varieties spoken in Iran typically employ a Persian-based script with diacritical marks when written, though it lacks standardization and remains primarily oral.

[7] https://tly.wikipedia.org

A.8 Zazaki

Zazaki (also known as Dimlî or Kirmanckî) is an Indo-European language that, together with Gorani (to which Hawrami belongs), forms the Zaza-Gorani subgroup. While linguistically distinct, Zazaki speakers in Turkey and Hawrami speakers in Iran and Iraq often identify as Kurdish, though this remains a complex sociopolitical issue (Arslan, 2019). Zazaki and Northern Kurdish (Kurmanji) have a high mutual-intelligibility (Ozek et al., 2023). Among the languages discussed in this paper, Zazaki has received the most extensive linguistic documentation, including comprehensive grammar books (Todd, 1985; Bingöl, 2020), dictionaries (ya Vateyî, 2009) and corpora (Ahmadi, 2020). Despite this academic attention, there have been ongoing calls for language revitalization efforts (Şanlı, 2022).

Writing System The language employs two distinct orthographic systems: one used on the Zazaki Wikipedia portal[8], influenced by Turkish orthography, and another based on Bedirxan's Kurdish orthography, which is also used in Northern Kurdish.

B Translation Guidelines

This appendix presents the detailed guidelines provided to translators participating in our parallel corpus creation project. The primary objective is to create parallel corpora for under-represented Middle Eastern languages.

Orthographic Considerations For languages with limited written traditions, we emphasize the importance of using existing orthography when available, regardless of its popularity among speakers. In cases where multiple writing systems exist, as in Zazaki and Gilaki, translators are instructed to select one system and maintain consistency throughout their translations. For languages lacking standardized orthography, translators are asked to develop and document their systematic approach.

Language Standardization Given that many target languages lack a standardized form, translators are advised to work within their most familiar dialect. This approach acknowledges the reality of linguistic variation while ensuring translation quality through the translator's expertise in their chosen

[8] https://diq.wikipedia.org

variety. If a standard form known, translators are asked to translate accordingly.

Lexical Choice Vocabulary selection is guided by the necessity of language preservation where native words are prioritized over loanwords. Although in everyday life, modern speakers might use certain words from other languages, like Arabic, Persian or Turkish, the task should reflect the potential of the language the most and not how the language under assimilation would be used. For instance, a Kurdish speaker might naturally use 'تسلط' (*tasaluṭ*) meaning 'authority', a loanword from Arabic (also used in Persian), while not using the native Kurdish word 'دەسەلّات' (*desellat*) for the same meaning. To aid decision-making, translators are encouraged to consider historical language use, often using the heuristic of whether a monolingual elder would recognize and use the term or not.

Terminologies Translating technical and specialized terminology presents unique challenges, particularly for low-resource languages. When encountering technical terms, e.g., 'computer' or 'hard disk', translators are advised to follow a two-step approach. If a native coinage exists in the target language–such as خێراکار (*xêrakar*) for 'computer' in Kurdish–that term should be used. Otherwise, in the absence of established native terminology, borrowing from other languages is acceptable, provided the borrowed terms conform to the target language's orthographic conventions. For example, in Gilaki, the word "computer" is borrowed as کامپیوتر (*kampyuter*), following Gilaki orthographic patterns despite its English origin. This approach maintains orthographic consistency while accommodating modern technical vocabulary.

Additional Resources Translators are encouraged to consult available dictionaries and language resources. This way, they maintain a dedicated glossary of new terminology and could compare both source languages when necessary for clarity. This approach supports consistent decision-making while building valuable supplementary resources for future work.

Translation Protocol To ensure corpus quality, translators were permitted to skip sentences containing morphosyntactic inconsistencies, spelling errors, code-switching, or named entities. These criteria help maintain focus on cleanly translatable content while avoiding potential sources of confusion or inconsistency.

C Datasets

As the first step in creating PARME, we preprocess both source and target sentences with character normalization, encoding standardization, removal of excessive characters such as elongations or *Tatweel* (U+0640), and punctuation normalization. This requires a set of language-dependent processing steps as well.

To split the datasets into train, validation and more importantly, test sets, we apply the following conditions on our parallel corpora:

1. **Avoiding Data Contamination:** The primary criterion is to eliminate data contamination by ensuring no overlap of source sentences across the training, validation, and test sets. This is especially critical as our data often includes multiple translations of the same sentence in various dialects. Any overlap would compromise the integrity of model evaluation.

2. **Consistency in Orthography:** For languages with multiple orthographic systems, we filter the test set to include sentences written in only one orthography. This minimizes the impact of orthographic variations on translation performance, particularly in cases where transliteration between orthographies is non-trivial. We believe that the orthographic unification should be carried out as a separate task, preferably on the training or validation set.

3. **Dialectal Diversity:** Aiming to reflect the full range of dialectal variation in the selected languages, we prioritize a uniform distribution of dialects in the test set if a standard form of the language is not known. This ensures the test set represents all dialects as long as the previous conditions are satisfied.

4. **Cross-Lingual Alignment:** Where possible, we prioritize sentences that had also been translated into other selected languages. This enables cross-lingual analysis, such as evaluating models across different languages on a shared test set. While full alignment across all

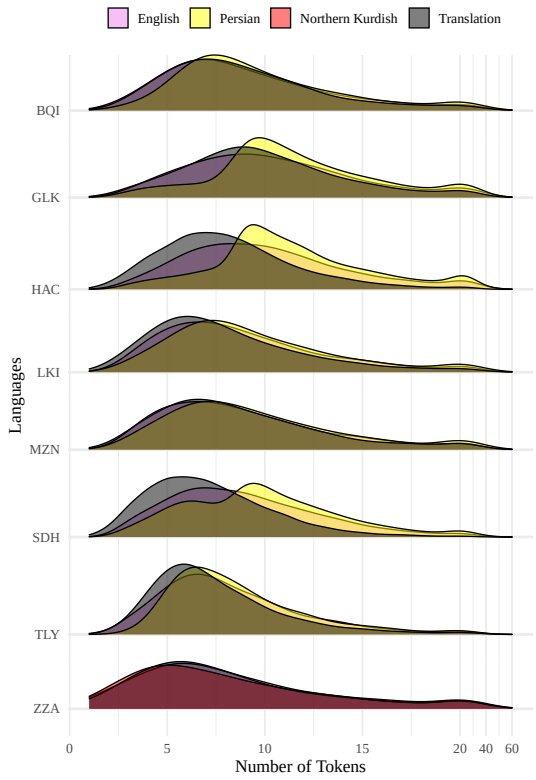

Figure C.1: Distribution of sentence lengths across test sets with a Gaussian density estimation. Higher peaks indicate more frequent token counts and overlapping curves show identical length in the reference languages and the translations.

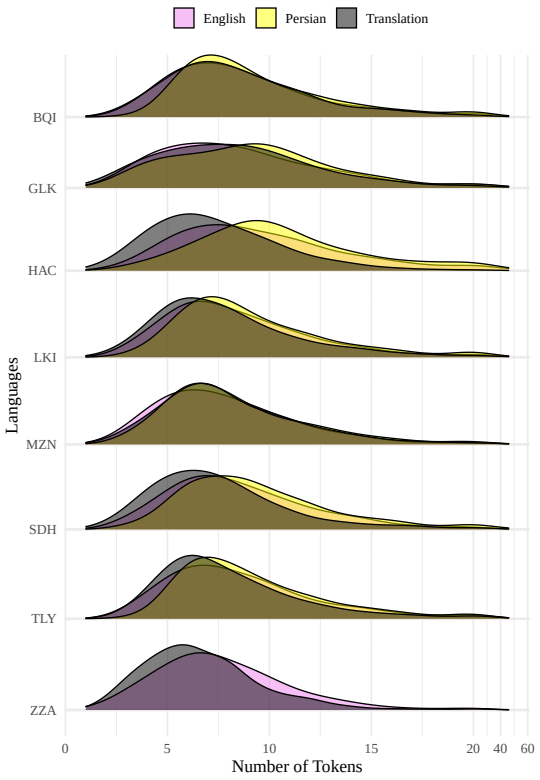

Figure C.2: Distribution of sentence lengths across test sets. Each curve represents the frequency of sentences containing different numbers of space-separated tokens in English, Persian, Northern Kurdish, and their translations. The analysis spans eight different languages.

languages is not feasible due to resource limitations, sentences meeting this criterion were prioritized during selection, provided they satisfy the other conditions.

Following the construction of the test set, the validation set, similarly containing 1,000 instances, is created using the same rules. This approach ensures that the validation set maintained high quality while complementing the test set effectively. All the remaining sentences are then added to the training set. Sentences that are included in the validation and training sets might have met a fewer number of the conditions.

Figures C.1 and C.2 show sentence length distributions across test sets and complete datasets, with ridge plots highlighting token patterns and overlapping curves indicating similarities between reference languages and translations. Meanwhile, Figures C.3 and C.4 illustrate the sentence distribution across languages in train/validation sets and base/augmented setups, alongside detailed dialect and orthography distributions.

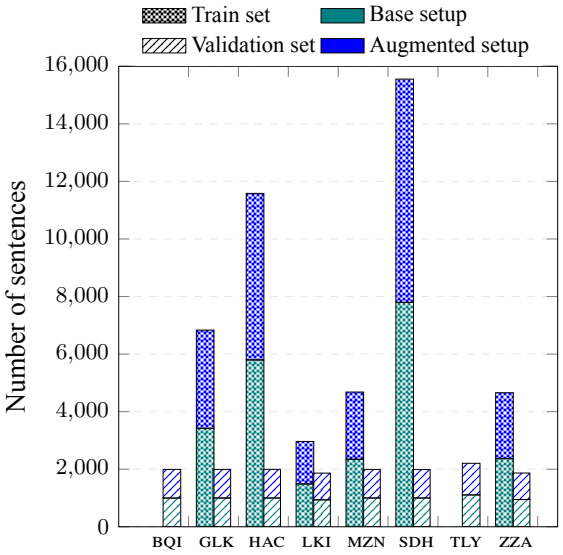

Figure C.3: Number of sentences in our collected parallel corpora based on languages in the train and validation sets (shown by colors) and the base and augmented setups (shown by patterns). Due to lack of data, Luri Bakhtiari (BQI) and Talysh (TLY) only have validation and test sets.

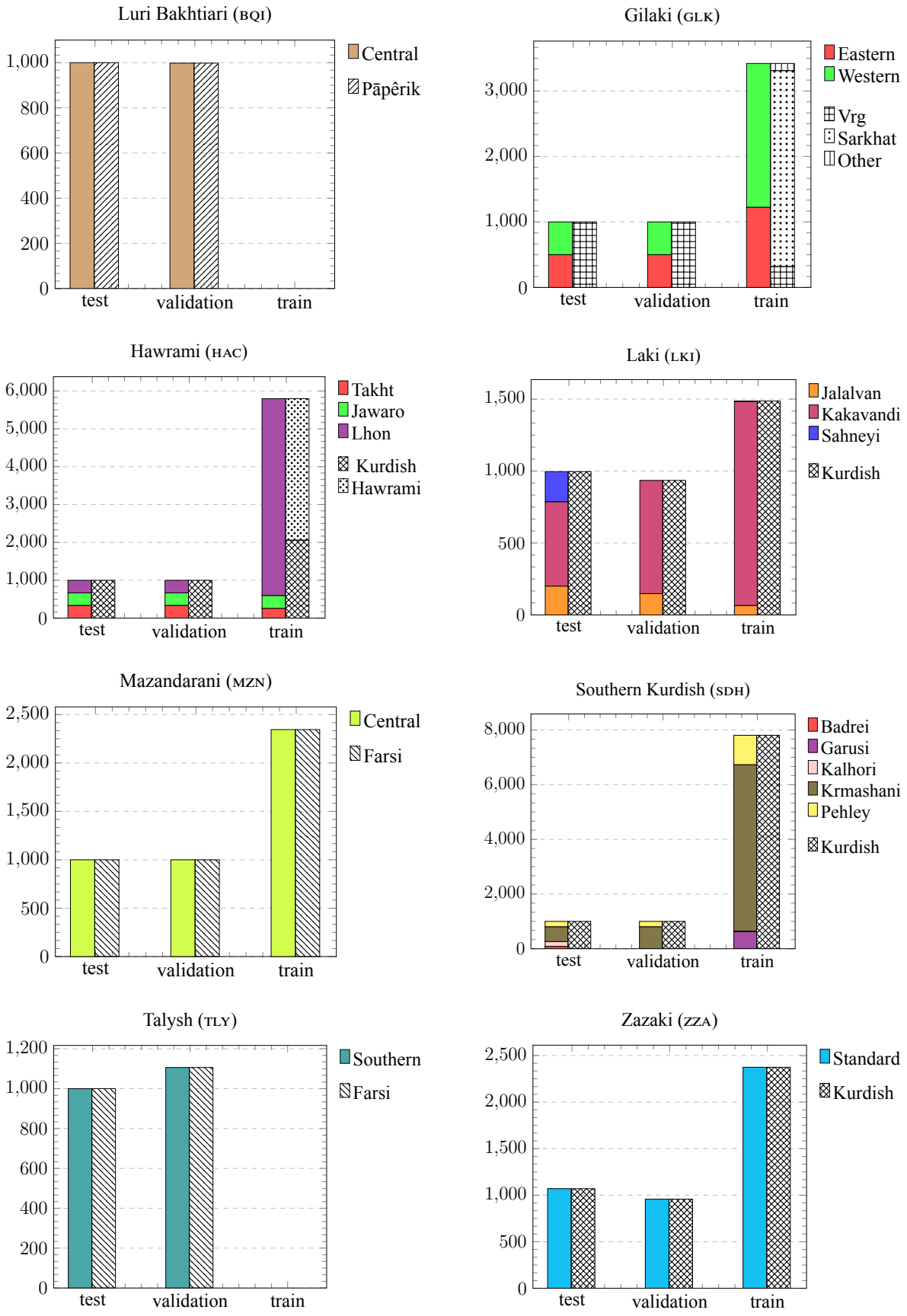

Figure C.4: Distribution of sentences across datasets and languages: solid colors indicate dialects, patterns indicate orthographic variants for each language. Our goal is to have a representative multi-dialectal test set in one orthography. There is no training set for Luri Bakhtiari and Talysh due to limited amount of data.

D Hyper-parameters Details

This appendix provides the complete hyper-parameter configurations used in our fine-tuning experiments. All experimental runs share the following base configuration:

- Base model: NLLB with multilingual tokenization
- Weight decay: 0.01
- Evaluation strategy: Per-epoch evaluation
- Model selection: Best checkpoint based on validation BLEU
- Save best model per epoch (max. 2 checkpoints)

We conducted several experimental runs with progressively refined configurations:

D.1 Configuration I (CONFIG I)

- Batch size: 8
- Gradient accumulation steps: 4
- Learning rate: 3e-5
- Training epochs: 20
- Warmup ratio: 0.1
- Maximum sequence length: 128 (source and target)
- Beam size: 5

D.2 Configuration II (CONFIG II)

- Batch size: 16
- Learning rate: 5e-4
- Training epochs: 50
- Warmup ratio: 0.15
- Maximum sequence length: 128 (source and target)
- Beam size: 5

D.3 Configuration III (CONFIG III)

Similar to Configuration II but with 100 epochs.

D.3.1 Improved Configuration: (CONFIG +)

- Base model: $\mathcal{M}_{\text{base}}$ trained with CONFIG I
- Batch size: 16
- Gradient accumulation steps: 4
- Learning rate: 2e-4
- Training epochs: 50
- Warmup ratio: 0.2
- Maximum sequence length: 256 (source and target)
- Beam size: 8

D.4 Training Resources

All models are trained on NVIDIA RTX 3090 GPUs (24GB VRAM) with training times ranging from 5.2 to 77.1 hours. The base models trained on 25,315 samples while augmented variants used 50,380 samples. Training throughput remained between 18-27 samples/second, with compute requirements scaling from 1.37e17 to 1.37e18 FLOPs across configurations. Improved configuration's losses reduces substantially with longer training, reaching 0.01 for 100-epoch runs.

CONFIG	Model	Time	Loss	S/s	FLOPs
I	$\mathcal{M}_{\text{base}}$	5.2	3.07	27.0	1.4e17
	$\mathcal{M}_{\text{aug}}$	10.4	2.56	27.0	2.7e17
II	$\mathcal{M}_{\text{base}}$	14.7h	0.34	23.9	3.4e17
	$\mathcal{M}_{\text{aug}}$	16.9	0.49	24.8	4.1e17
III	$\mathcal{M}_{\text{base}}$	27.9h	0.22	25.2	6.9e17
	$\mathcal{M}_{\text{aug}}$	77.1	0.28	18.2	1.4e18
+	$\mathcal{M}_{\text{base}}$	19.1	0.01	18.4	6.8e17

Table D.1: Fine-tuning details for each model and configuration. Time is provided in hour and S/s refers to samples per second.

The results demonstrate the effectiveness of our improved configuration (CONFIG+), achieving the highest average BLEU score of 7.50 in just 19.1 hours of training compared to the more resource-intensive CONFIG III $\mathcal{M}_{\text{aug}}$ which yields only 6.23 BLEU despite requiring four times longer (77.1 hours). Even our lightweight CONFIG I $\mathcal{M}_{\text{base}}$ achieves a respectable 5.41 BLEU in just 5.2 hours. While Figure D.1 shows training loss consistently decreasing to 0.001 by epoch 50 with gradually increasing validation loss (typical in sequence-to-sequence tasks), BLEU scores remain robust throughout training (14.21 to 14.57), with consistent generation lengths (14.0-14.9 tokens), indicating the model learns effective translation patterns without suffering from performance-degrading overfitting.

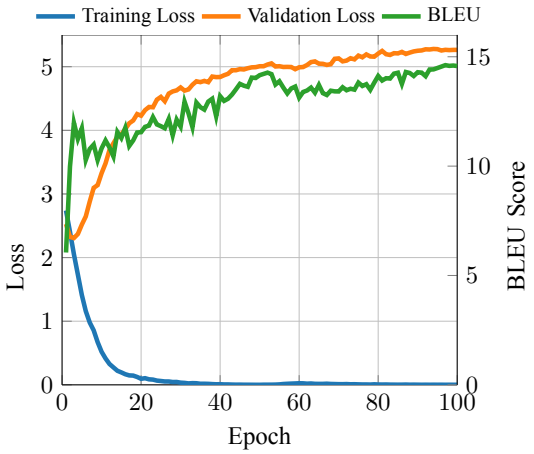

Figure D.1: Fine-tuning progress according to training and validations losses along with the BLEU Score over epochs for $\mathcal{M}_{\text{base}}$ in CONFIG II (up to epoch 50) and for $\mathcal{M}_{\text{base}}$ in CONFIG + for the succeeding epochs.

E Results

E.1 Baseline

	nllb-200-distilled-600M						nllb-200-3.3B					
	ENG_L	TUR_L	KMR_L	CKB_A	PES_A	ARB_A	ENG_L	TUR_L	KMR_L	CKB_A	PES_A	ARB_A
BQI_A	0.74	0.70	0.72	0.58	1.03	0.42	0.58	1.26	0.66	0.70	1.56	0.26
GLK_A	0.56	0.64	0.61	0.52	0.75	0.28	0.60	0.76	0.57	0.53	0.80	0.33
HAC_A	0.73	1.01	1.13	1.89	1.84	0.69	1.50	1.12	1.24	2.02	1.86	0.63
LKI_A	0.42	0.37	0.36	1.20	1.32	0.25	0.58	0.67	0.57	1.35	1.27	0.47
MZN_A	1.07	1.04	1.00	1.30	1.98	0.57	1.41	1.54	0.70	0.71	2.05	0.63
SDH_A	1.10	1.28	2.18	2.77	2.30	0.92	1.77	1.64	2.19	2.86	2.60	1.00
TLY_A	0.43	0.44	0.49	0.66	0.90	0.30	0.35	0.52	0.24	0.43	1.00	0.17
ZZA_L	0.45	0.81	2.82	0.80	1.00	1.09	0.43	1.03	2.96	1.59	1.26	0.30
AVE	0.69	0.77	1.16	1.21	1.39	0.56	0.90	1.07	1.14	1.27	1.55	0.47

(a) Evaluation based on BLEU ↑

	nllb-200-distilled-600M						nllb-200-3.3B					
	ENG_L	TUR_L	KMR_L	CKB_A	PES_A	ARB_A	ENG_L	TUR_L	KMR_L	CKB_A	PES_A	ARB_A
BQI_A	15.02	13.46	14.95	14.57	15.49	13.46	14.65	15.27	14.26	15.45	16.69	13.09
GLK_A	13.75	13.26	14.63	14.21	15.20	13.44	14.90	14.49	14.87	14.42	15.17	13.33
HAC_A	15.12	13.67	16.99	19.13	17.46	14.90	17.05	15.46	18.20	19.93	17.98	13.96
LKI_A	13.45	11.32	13.28	15.69	15.02	12.63	14.35	14.32	15.25	16.87	15.42	12.65
MZN_A	16.19	15.38	16.47	15.18	18.41	14.45	16.16	17.70	16.17	14.90	19.19	14.39
SDH_A	15.78	14.36	17.05	19.41	17.32	14.70	17.39	16.31	18.85	19.89	18.15	13.99
TLY_A	14.43	13.98	13.33	14.24	15.82	14.30	15.25	14.00	13.24	15.44	16.08	13.52
ZZA_L	13.81	15.76	22.07	16.45	16.22	18.69	14.50	17.67	22.41	19.45	16.26	12.86
AVE	14.69	13.9	16.09	16.11	16.37	14.57	15.53	15.65	16.65	17.04	16.87	13.47

(b) Evaluation using chrF ↑

	BLEU ↑				chrF ↑			
	ENG_L	TUR_L	KMR_L	ARB_L	ENG_L	TUR_L	KMR_L	ARB_L
BQI_L	0.17	0.10	0.05	0.09	10.78	8.85	9.13	9.51
GLK_L	0.06	0.05	0.04	0.06	10.77	9.55	9.47	10.82
HAC_L	0.16	0.34	0.38	0.27	11.84	11.76	14.78	13.82
LKI_L	0.19	0.17	0.23	0.14	11.46	11.60	14.33	14.44
MZN_L	0.10	0.03	0.01	0.03	9.72	8.68	8.49	8.72
SDH_L	0.17	0.21	1.00	0.23	11.91	11.85	14.93	13.52
TLY_L	0.14	0.10	0.08	0.16	10.00	9.26	8.99	11.07
ZZA_A	1.20	1.58	0.83	0.98	17.99	18.80	15.71	15.63
AVE	0.27	0.32	0.32	0.245	11.80	11.29	11.98	12.191

(c) Input languages' script transliterated to match that of the setup language using `nllb-200-distilled-600M`

Table E.1: Zero-shot evaluation of NLLB on our selected languages as the baseline based on BLEU (a) and chrF (b). Given that none of our selected languages have been seen by the model, we set different proximate languages for inference (second row). Transliteration to match the input and inference scripts (c) has little effect on performance.

E.2 Translation Examples

Language	System	Example 1	Example 2
Luri Bakhtiari (BQI)	S	بدبختی یو نه که مو زنده یی نه غلوه دوست داروم	وه دنگ پای زمان گوش ایْدَه
	R_{ENG}	I've loved life too much, shamefully much.	listening to the echoing footsteps of years.
	R_{PES}	من زندگی را زیاد دوست دارم. بدبختی همین است.	به طنین صدای پای زمان گوش فرا می داد.
	T	*the misfortune of not loving me alive and well*	*he's been listening to her all the time.*
Gilaki (GLK)	S	زماتي كي مي بالانه موَ سرأ گيفت فر بوَخوَره	چوَن ناخوَش بوَ.
	R_{ENG}	even as the hairs on my arms began to shrivel.	because she had been ill
	R_{PES}	حتی زمانی که موهای روی بازوانم شروع به فر خوردن کرد.	زیرا بیمار بود
	T	*When I put my arms on my hair, it was very cold.*	*that was all disagreeable to me.*
Hawrami (HAC)	S	من ئیزن نمەڎو كه ماجەرا پی جۆرە تەمام بۆ	گوێش گیرتەی جە دەنگدانەوەو هەنگامەکا ساڵانی
	R_{ENG}	I won't let it end like this.	listening to the echoing footsteps of years.
	R_{PES}	من اجازه نخواهم داد که ماجرا به این شکل پایان یابد.	به طنین صدای پای زمان گوش فرا می داد.
	T	*I don't expect it's happening.*	*listening to the footsteps of the young man*
Laki (LKI)	S	ڤه دوشمین و هەرەشه دەستوور دامێیه بێن	سه ئەو ڤه ڕاسییا ئەسیر بۆ
	R_{ENG}	he ordered me in with an oath	so truly was he captivated.
	R_{PES}	با دشنامی تهدیدآمیز فرمانم داد	کلاید به راستی اسیر شده بود.
	T	*I have given orders, with expressions of the highest acknowledgment.*	*so he was actually a prisoner*
Mazanderani (MZN)	S	اما اسا خله ویشتر وره بمونسسی	مه دل اینجه همه ی وسسه تنگ بونه
	R_{ENG}	you're much more her now	I'll miss everyone here.
	R_{PES}	اما حالا خیلی بیشتر به او شبیه شدی	من دلم برای همه در اینجا تنگ می شه
	T	*but now you've become enormously attached to it.*	*I miss everything here.*
Southern Kurdish (SDH)	S	تەقەلا ئەڕای ئەوڕەسین له مەبەست کەیوانوو	ئێ پرسه خەوەر ڕووژه.
	R_{ENG}	trying to grasp the old lady's meaning	this funeral is the news of the day.
	R_{PES}	سعی کرد که قصد بانوی سالخورده را درک کند.	واقعه امروزی اسباب صحبت امشب است.
	T	*I intend to get there.*	*the question is a feature on the war.*
Talysh (TLY)	S	يَواشينَه زونو نَه ايشتيمَه.	ايدَفَه هنی ايله ايسبيێه كاملیا گُلْش ويگِتَه
	R_{ENG}	I raised myself gently upon my knees	she received a white camellia again
	R_{PES}	آهسته به قوت زانو برخاستم	بار دیگر یک گل کاملیای سفید دیگر دریافت کرد
	T	*I knocked it out.*	*suddenly a spider came along.*
Zazaki (ZZA)	S	O rî ra kî ê mînetdarê dersimizan ê.	Di hebî kitabê min kenê çap bibê.
	R_{ENG}	For that reason they are thankful to the Dêrsimis	Two of my books are about to get published
	T	*That's the way it's supposed to be.*	*I'm going to print my book here.*

Table E.2: Translation of sentences from our selected languages into English. Two examples are provided per languages with source (S), reference in English (R_{ENG}) and Farsi (R_{PES}) along with machine translation output (T).