Universität Zürich UZH

Department of Computational Linguistics

Parallel Corpora for Low-Resourced Middle Eastern Languages

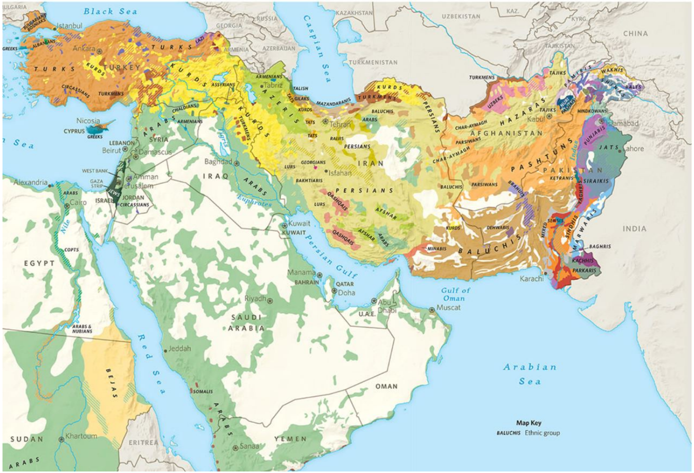
Sina Ahmadi Rico Sennrich

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 Sedighe Zamani Roodsari

The 63rd Annual Meeting of the Association for Computational Linguistics - ACL 2025
Vienna, Austria

Background: Languages in the Middle East

- Remarkable linguistic diversity in the Middle East
- 400+ million people speaking lots of “languages”
- Only a handful of those languages are officially recognized



Background: Under-represented Middle Eastern Languages

- 60 varieties in the region identified as **endangered** by UNESCO [Moseley, 2010]
- Many face **existential threats**
  - Systematic assimilation campaigns
  - Limited educational resources
  - Younger generations have **fewer opportunities** to develop literacy in heritage languages
  - Lack of standardization

> *"An endangered language will progress if its speakers can make use of electronic technology."*
>
> **— [Crystal, 2002] (Language Death)**

Background: Limited Technological Progress for Middle Eastern NLP

Little or no progress in language and speech technologies

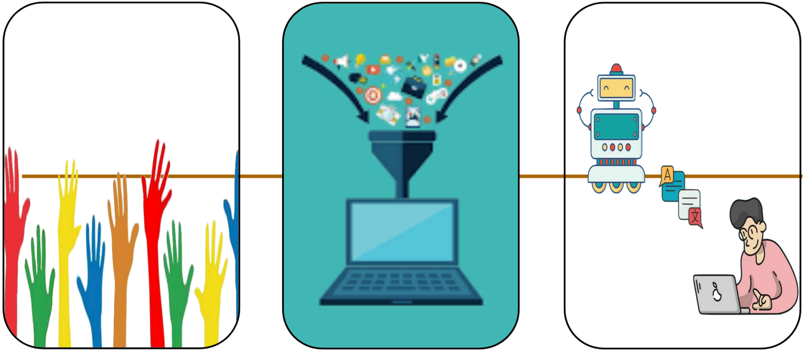
- Limited keyboard implementations and technological support
- Limited community support for resource development
- Lack of corpora, including parallel ones
- Only a few languages have Wikipedia portals
- 8,000–50,000 articles where available

	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	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

PARME: Community-Driven Approach

Objective: parallel corpora for under-represented Middle Eastern languages (PARME)

⇒ Community-driven participatory research [Nekoto et al., 2020]



Step 1
Community Building

Step 2
Data Collection

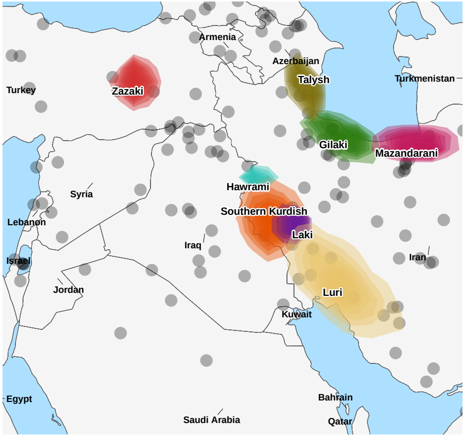
Step 3
NLP Development

PARME: Community Building

- Intensive campaign (Aug-Oct 2024)
- Twitter/X & LinkedIn for public announcements
- Direct outreach to academics and native speakers
- Cold messaging to published authors & experts
- Mixed Reception: enthusiasm but also skepticism

⇒ **45 volunteers for eight severely under-resourced languages**

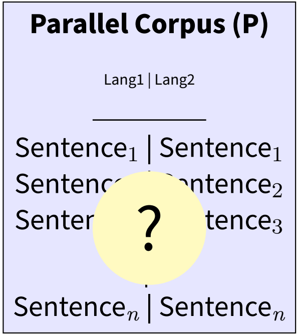
⇒ Each language is spoken by 0.3 to 5 million speakers



PARME: Data Collection

What sentences should the volunteers translate into their languages?

- Sample randomly? → limited resources and time
- Instead, *select sentences strategically* [Ambati et al., 2011]
  - Select sentences from a parallel corpus in Lang1-Lang2
  - **Lang1:** Language familiar to translator (Farsi, Turkish, Arabic or Kurdish)
  - **Lang2:** High-resource language (English)
  - Maximize **lexical diversity**
  - Enhance **semantic richness**



Parallel Corpus (P)

Lang1 | Lang2

Sentence$_1$ | Sentence$_1$

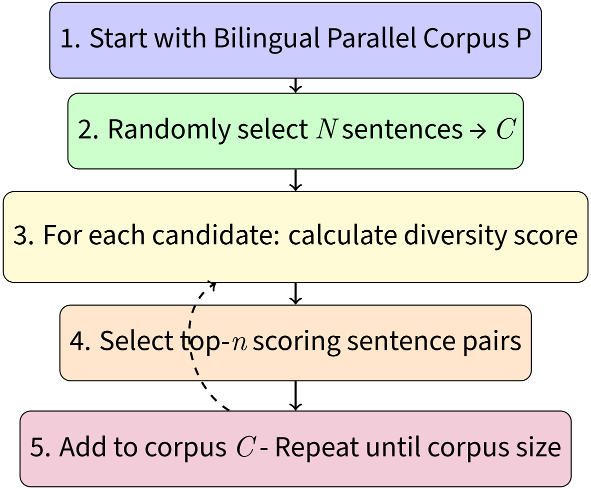
Sentenc ntence$_2$

Senter tence$_3$

?

Sentence$_n$ | Sentence$_n$

PARME: Strategic Data Selection



1. Start with Bilingual Parallel Corpus P

2. Randomly select N sentences → C

3. For each candidate: calculate diversity score

4. Select top-n scoring sentence pairs

5. Add to corpus C - Repeat until corpus size

$N, n = 3000$ / corpus size: 15,000

$\Rightarrow$ **a trilingual corpus (Lang1, Lang2, X)**

Diversity Score Method:

D = Edit distance on Lang1 sentences

S = Semantic similarity on Lang2 sentences

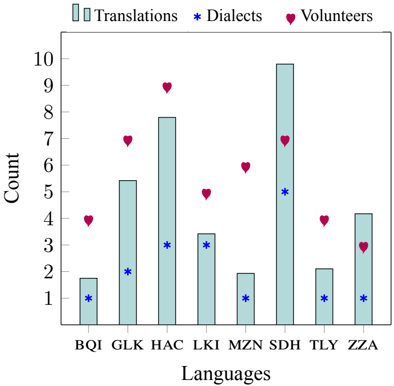
$$\text{score}_i = \frac{\overline{D}_i}{\overline{S}_i}$$

$D\uparrow$, $S\downarrow \Rightarrow$ score $\uparrow$

Rewards lexically diverse sentences with different semantic content

PARME: Manual Translation

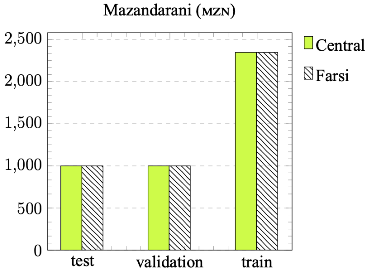
- Manually translate on spreadsheets
- Follow translation guidelines
  - Consistent orthography (per translator)
  - Translate into the standard variety, otherwise into your dialect
- At least two translators checked translations and assessed quality
- ⇒ **36,384 translation pairs** in eight languages, 18 varieties and seven orthographies

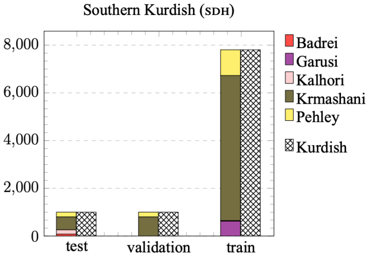


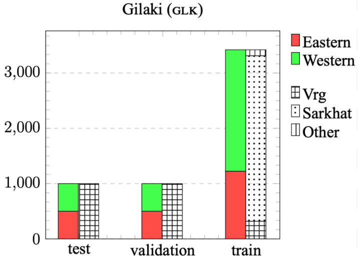
	English	Persian	Gilaki	Mazanderani	Talysh	Laki	Luri	Hawrami	S. Kurdish	Zazaki
Persian	20236									
Gilaki	5418	5418								
Mazanderani	4342	4342	1108							
Talysh	2101	2101	84	851						
Laki	3412	3412	437	3011	849					
Luri	1997	1997	39	1396	785	1235				
Hawrami	7792	7792	1957	1284	409	723	934			
S. Kurdish	8582	8582	488	1644	1581	1424	994	1256		
Zazaki	4401	0	0	0	0	0	0	0	0	
N. Kurdish	0	0	0	0	0	0	0	0	0	3331

PARME: Evaluation Set

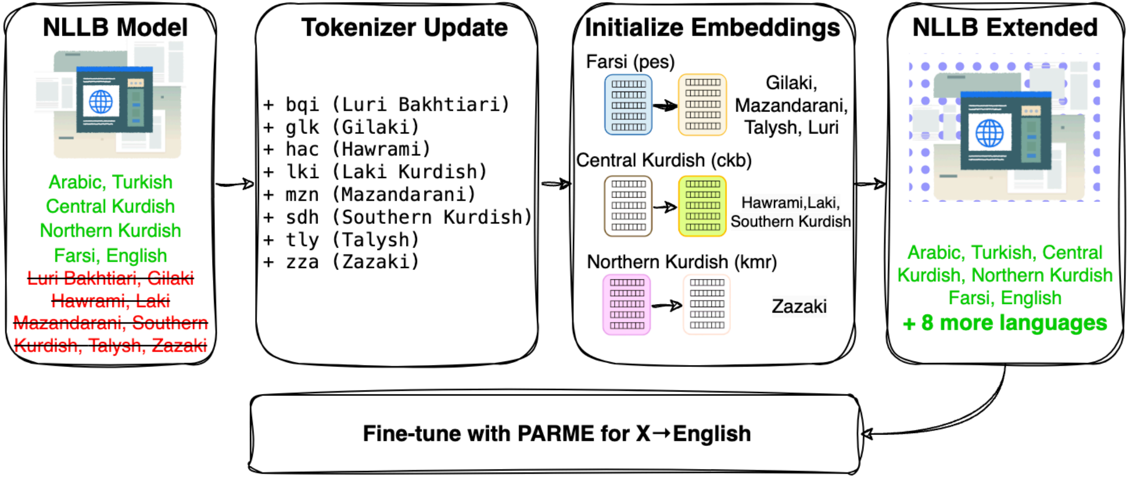
- **Create representative test sets for multi-script multi-dialectal translations** by
  1. Avoiding data contamination
  2. Maintaining orthographic consistency
  3. Representing all dialects
  4. Prioritizing cross-linguality
- ⇒ **Approximately 1,000 test instances per language**


Mazandarani (MZN)


Southern Kurdish (SDH)


Gilaki (GLK)

Experiments: NLLB Fine-tuning



NLLB Model

Arabic, Turkish
Central Kurdish
Northern Kurdish
Farsi, English
~~Luri Bakhtiari, Gilaki~~
~~Hawrami, Laki~~
~~Mazandarani, Southern Kurdish, Talysh, Zazaki~~

Tokenizer Update

```
+ bqi (Luri Bakhtiari)
+ glk (Gilaki)
+ hac (Hawrami)
+ lki (Laki Kurdish)
+ mzn (Mazandarani)
+ sdh (Southern Kurdish)
+ tly (Talysh)
+ zza (Zazaki)
```

Initialize Embeddings

Farsi (pes) → Gilaki, Mazandarani, Talysh, Luri

Central Kurdish (ckb) → Hawrami,Laki, Southern Kurdish

Northern Kurdish (kmr) → Zazaki

NLLB Extended

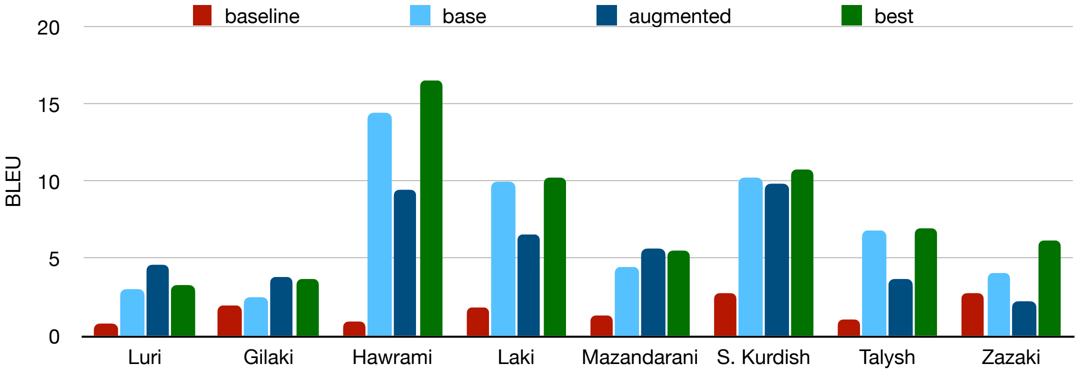
Arabic, Turkish, Central Kurdish, Northern Kurdish
Farsi, English
+ 8 more languages

Fine-tune with PARME for X→English

* No Language Left Behind covering 200 languages [Team et al., 2024]

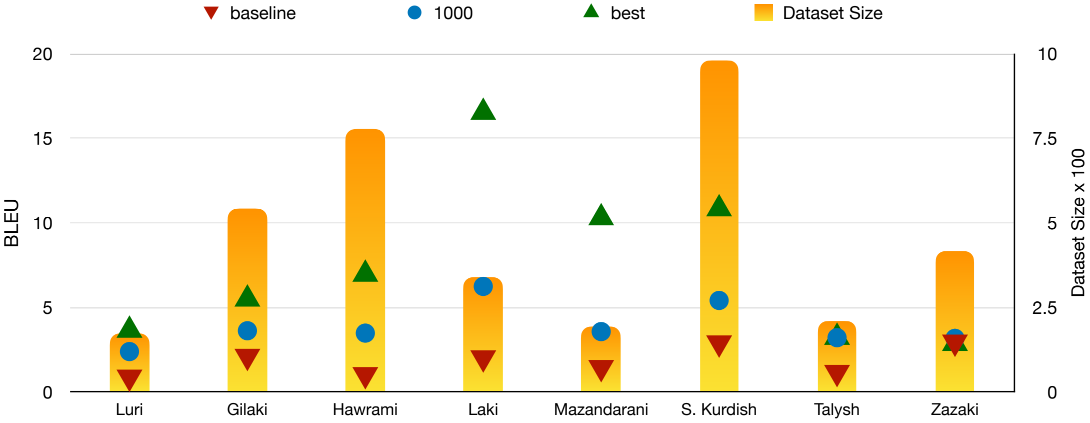
Experiments: Results

- Baseline (zero-shot) performance is poor
- Fine-tuning on PARME X-EN → substantial improvements
- Augmented fine-tuning on X-EN & X-(Lang1→EN) → mixed performance
- Augmentation hurts languages initialized from CKB while beneficial for PES & KMR
- **Best performance** → fine-tuning on PARME X-EN for longer



Experiments: Data Size Impact

- Are performance improvements consistently proportional to data quantity?
- ⇒ Data quality and hyperparameters matter as much as quantity



Conclusion

- **First parallel corpora** for 8 severely under-resourced ME languages (18-23M speakers)
- **Community-driven success:** Native speakers volunteering in participatory research
- **Substantial MT improvements** but much room for improvement
- **Reproducible framework**
- **Opening new research avenues** in NLP for ME languages
- **Language preservation:** Digital resources can enhance language prestige and vitality

"Technology as a pathway to linguistic diversity preservation and digital inclusion"

Thank You!

Acknowledgments:

- **Heartfelt gratitude to > 40 volunteers who actively participated in this initiative.**
- Swiss National Science Foundation (MUTAMUR project)
- Stanford SILICON Initiative (`https://silicon.stanford.edu`)

Contact: `sina.ahmadi@uzh.ch`

Resources: `https://github.com/DOLMA-NLP/PARME`

Questions?

References

Ambati, V., Vogel, S., and Carbonell, J. G. (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*.

Crystal, D. (2002).
Language death.
Cambridge University Press.

Moseley, C. (2010).
Atlas of the World's Languages in Danger.
UNESCO.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020).
Participatory research for low-resourced machine translation: A case study in African languages.
In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

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