University of Zurich UZH

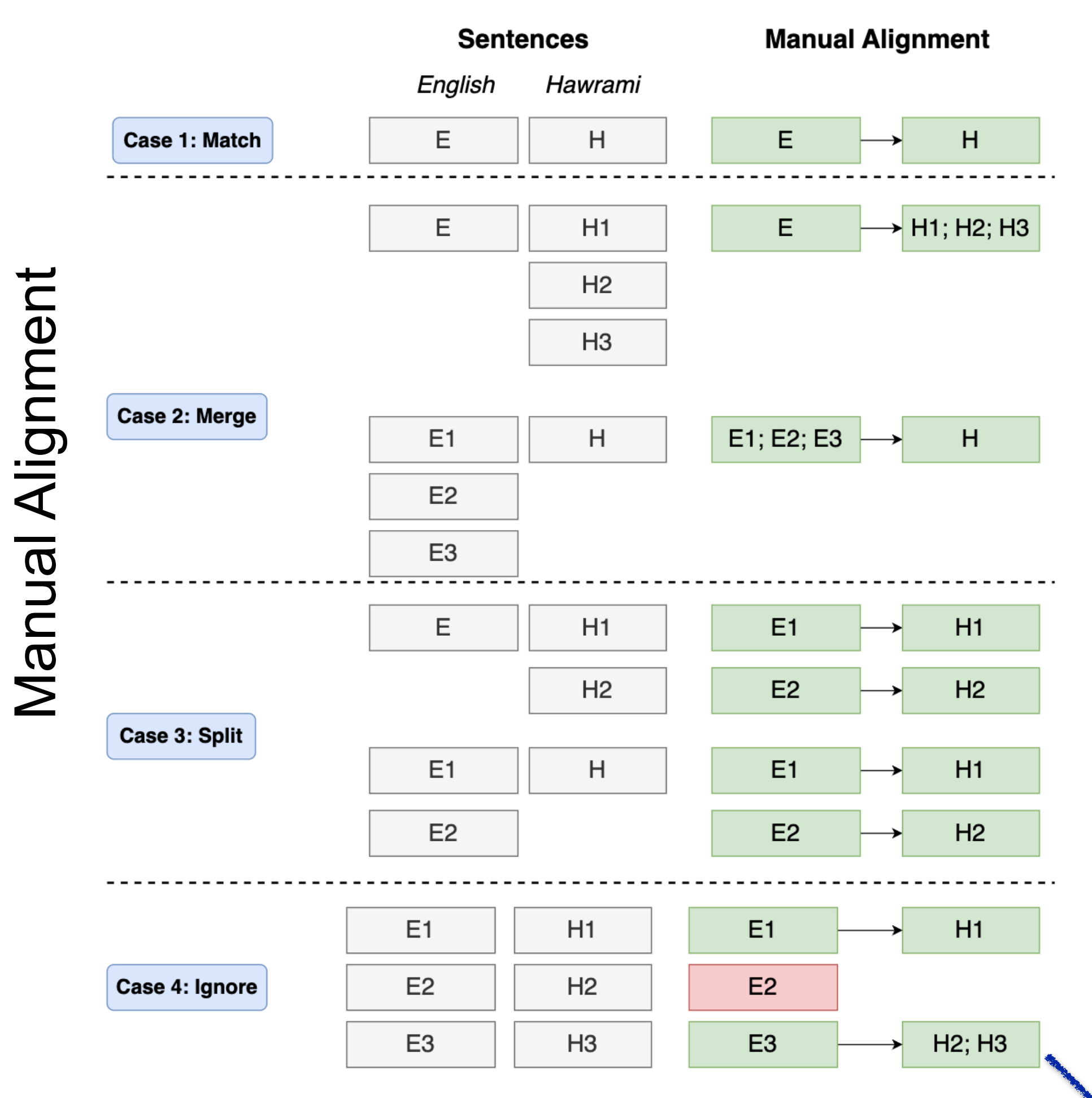
Department of Computational Linguistics

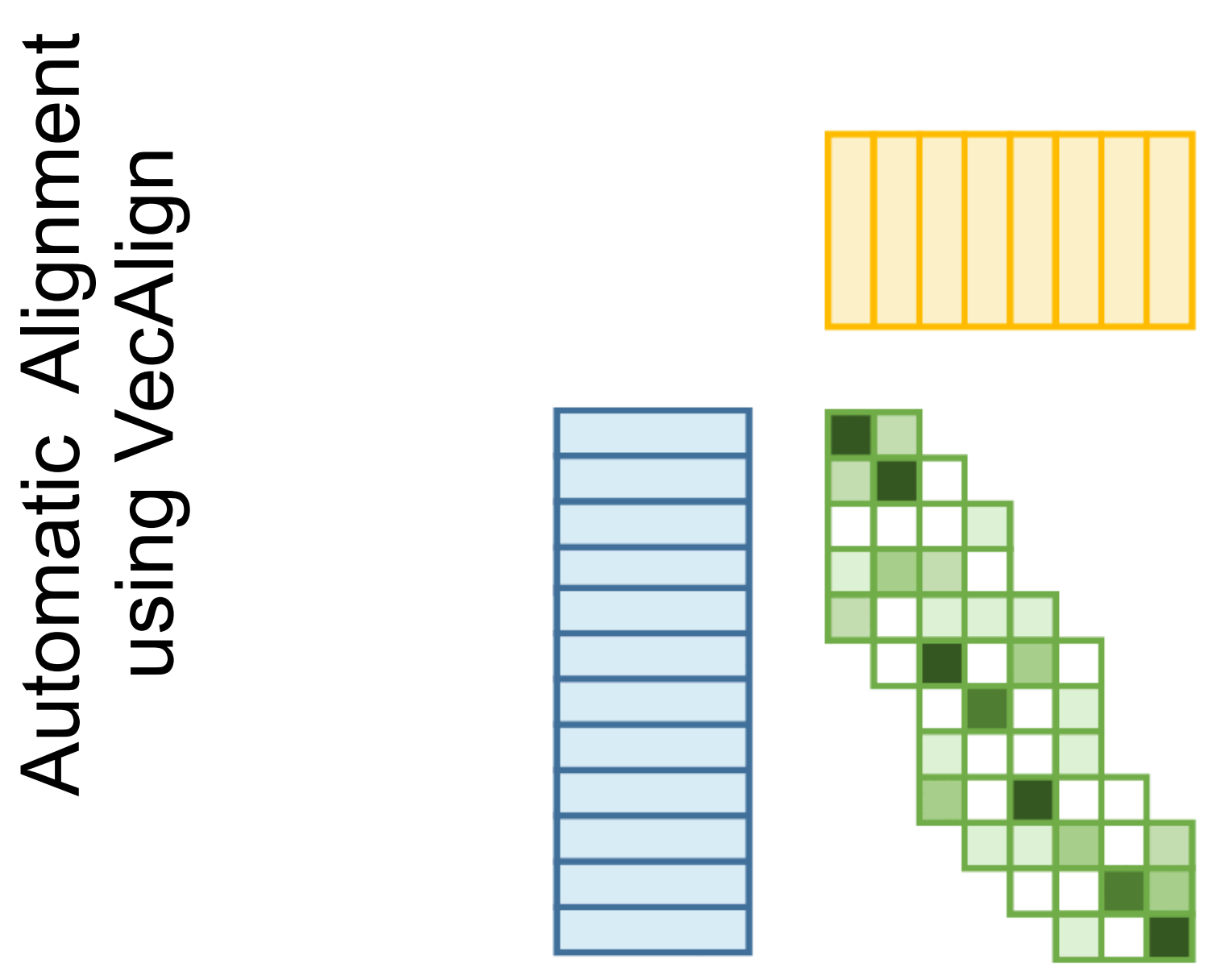
ACL 2025 VIENNA JULY 27 - AUGUST 1

22th International Conference on Spoken Language Translation
IWSLT 2025
Vienna, Austria
31 July – 1 August 2025

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

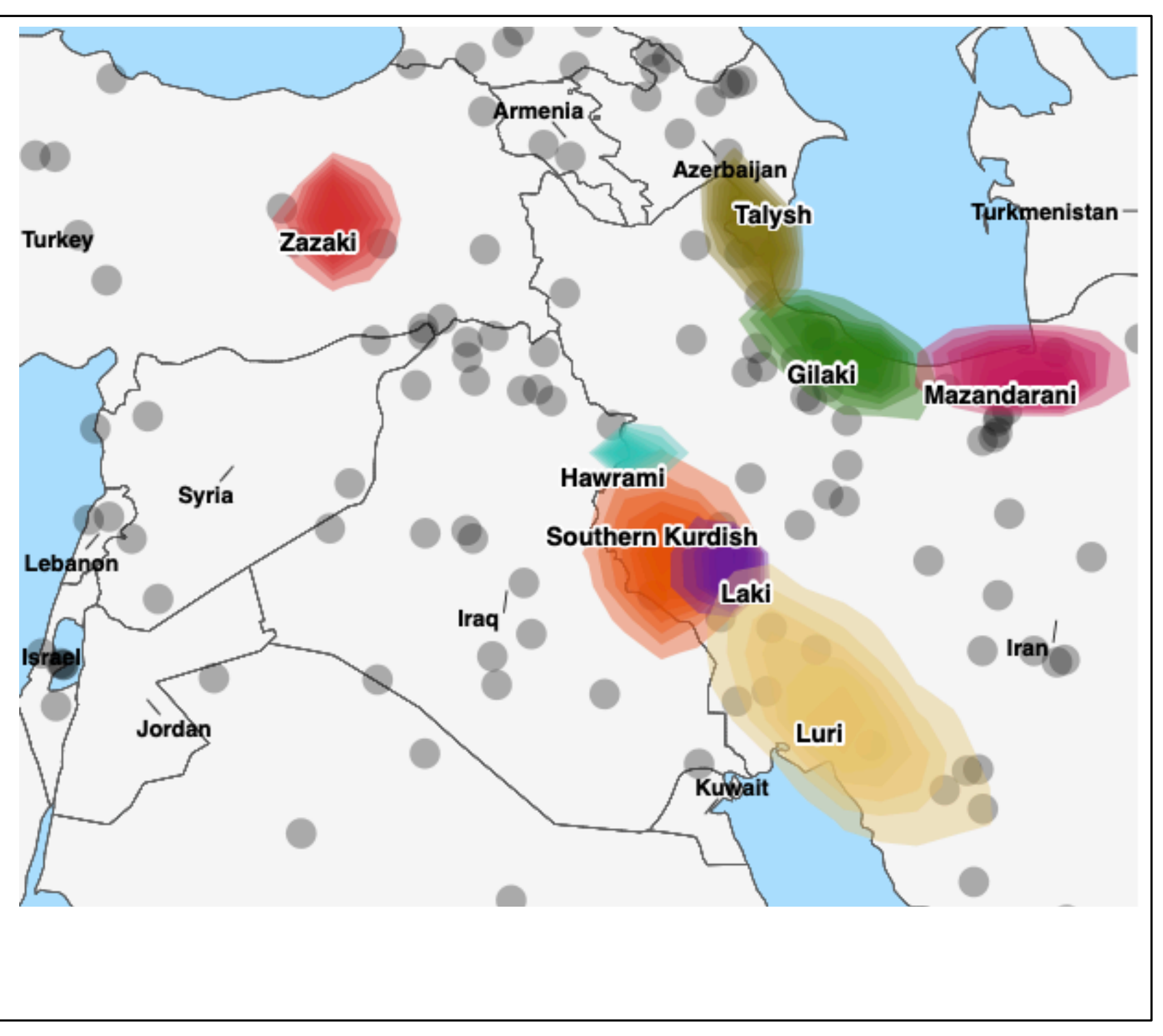
Sina Ahmadi, Razhan Hameed, Rico Sennrich
University of Zurich, Switzerland & Vox AI, Netherlands

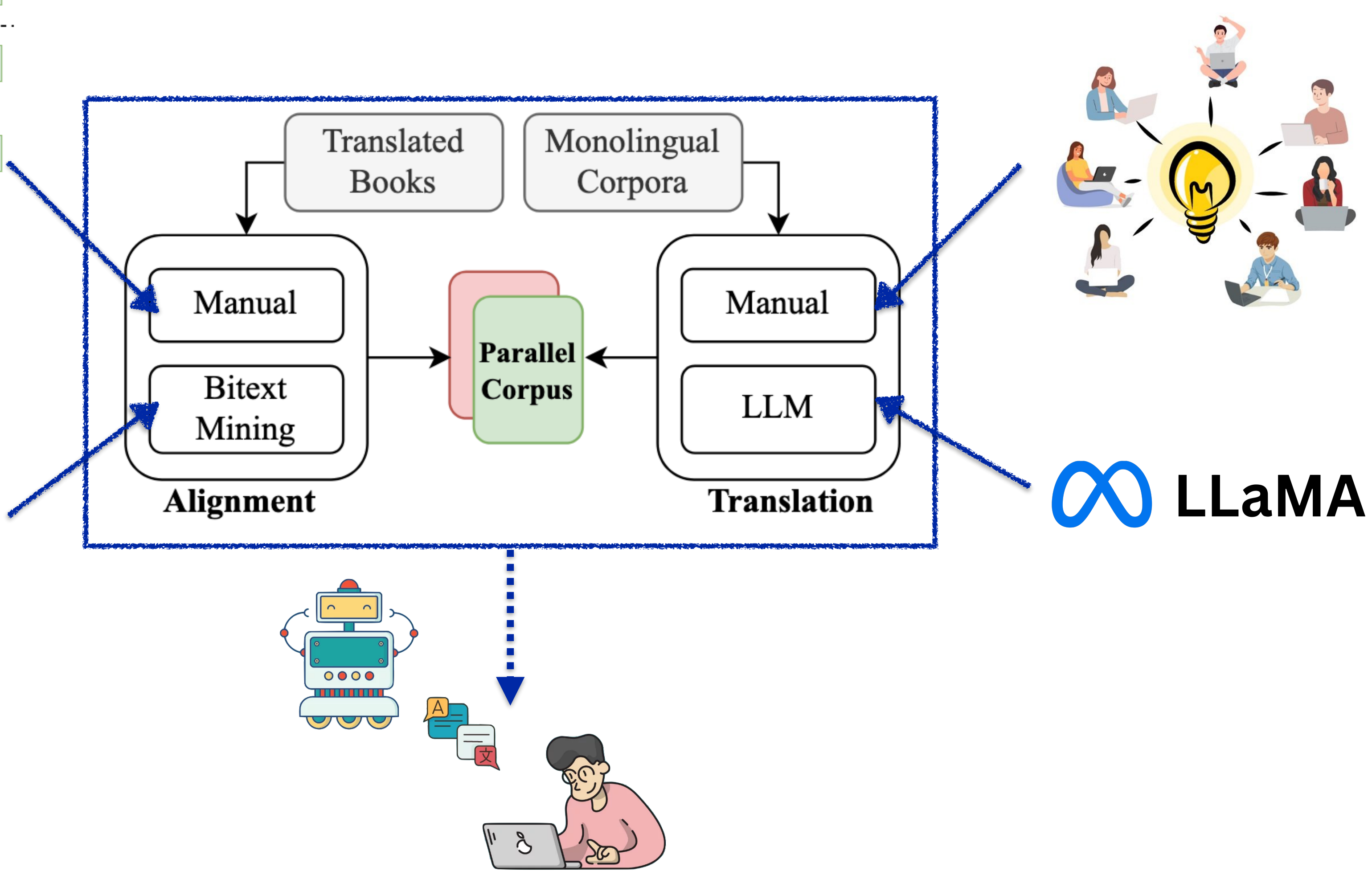




Motivation

- Remarkable linguistic diversity in the Middle East
- 400+ million people speaking lots of "languages"
- Only a handful of those languages are officially recognized
- 60 varieties identified as endangered by UNESCO
- Limited technological support
- Limited community support for resource development
- Lack of corpora, including parallel ones
- **Objective: extract sentences to create parallel corpora**





Machine Translation: fine-tuning NLLB

Data Collection

- **Manual Translation (PARME)**: Native speakers translate 25,334 English sentences into 8 Middle Eastern languages through participatory research
- **Sentence Alignment**: Align 25 translated books/articles to original English texts using **manual expert alignment (M)** and automatic **Vecalign (V)** > 25,203 pairs
- **LLM Augmentation**: Few-shot prompting with Gemini-2.0-flash and LLaMa on monolingual corpora > 221,774 pairs
- **Final Dataset: 272,311 total sentence pairs across PARME (P), Manual (M), Vecalign (V), and LLM (L) sources with varying coverage per language**

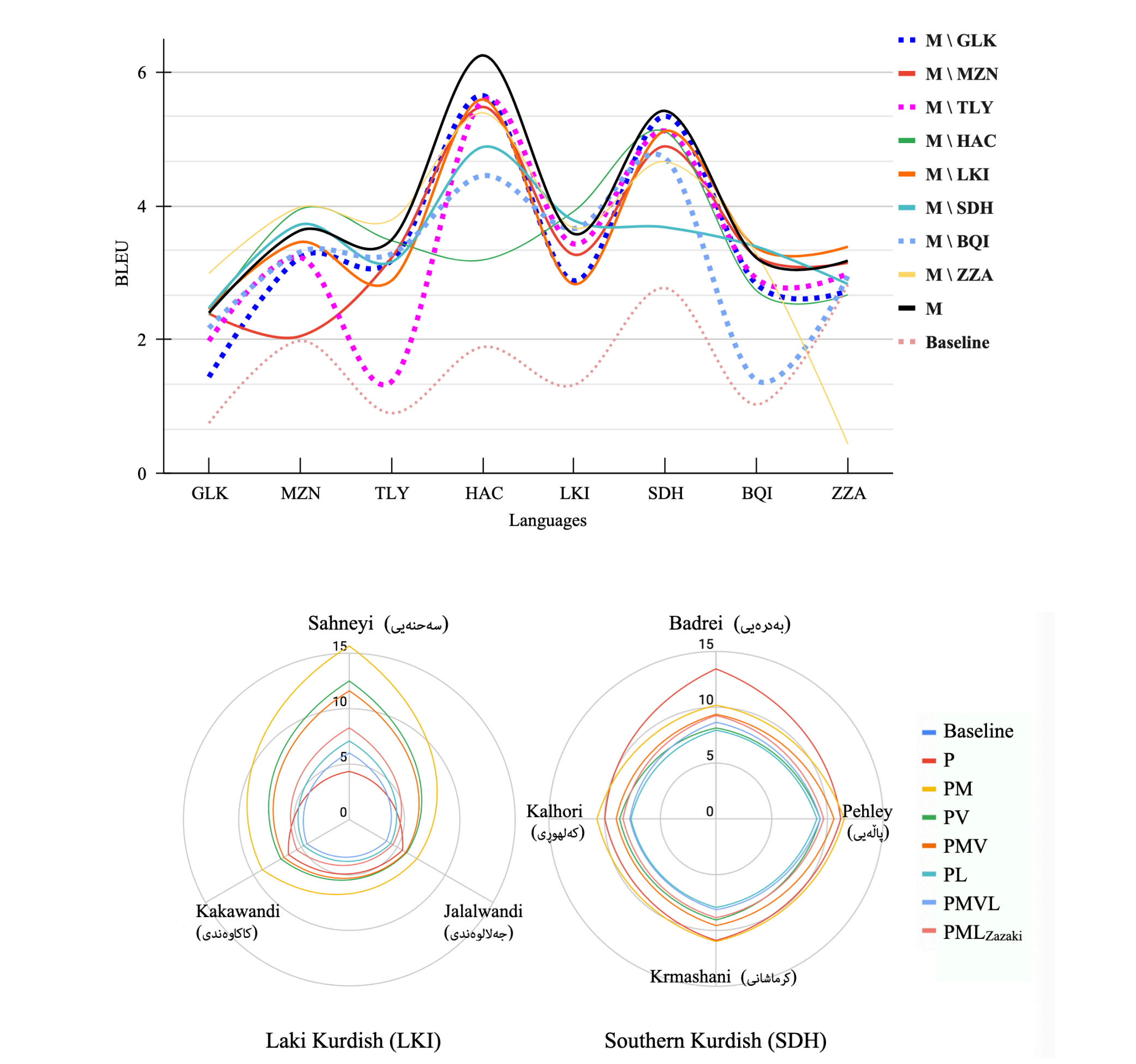
Language	P	M	V	L
Luri Bakhtiari (BQI)	999	0	0	0
Gilaki (GLK)	3420	999	1391	22467
Hawrami (HAC)	5796	7050	8367	49987
Laki Kurdish (LKI)	1487	1220	0	0
Mazandarni (MZN)	2345	0	0	49328
Southern Kurdish (SDH)	7806	3681	2495	49992
Talysh (TLY)	1107	0	0	0
Zazaki (ZZA)	2374	0	0	50000
Sum	25,334	12,950	12,253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

Strategic data curation is key: carefully selected small datasets outperform synthetic datasets for low-resource languages.

Experimental Results

- Fine-tuned NLLB (600M) with related language embeddings across different data combinations
- **Quality > Quantity**: Manual alignment (PM) achieves highest average BLEU (7.38) despite being smaller than LLM dataset (PL: 5.64)
- **Best Performance:** Hawrami reaches 15.46 BLEU, significant improvement over 0.9 baseline
- **Cross-linguistic Interference**: Adding data for one language can hurt others
- **Dialectal Variation**: Performance varies significantly within dialects
- **Overall Improvement: All languages show substantial gains over baseline, with average BLEU increasing from 1.68 to 7.38 (PM)**

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



Conclusion

- Manual alignment outperforms other datasets, achieving 7.38 vs 5.64 average BLEU
- Adding data for one language can hurt others in multilingual settings
- Dialectal variation matters: Performance varies significantly across varieties
- There are significant performance variation across different varieties in MT

Swiss National Science Foundation