



**University of  
Zurich** <sup>UZH</sup>

# Dialogues for Documenting Dialects

## Language and Speech Technology for Central Kurdish Varieties

**Sina Ahmadi**<sup>1</sup>, Daban Q. Jaff<sup>2</sup>, Md Mahfuz Ibn Alam<sup>3</sup>, Antonios Anastasopoulos<sup>3,4</sup>

1 University of Zurich, Switzerland

2 University of Erfurt, Germany

3 George Mason University, USA

4 Archimedes AI Research Unit, Greece

LREC-COLING  2024



## Table of Contents

- Background
- Methodology
- Experiments
- Conclusion

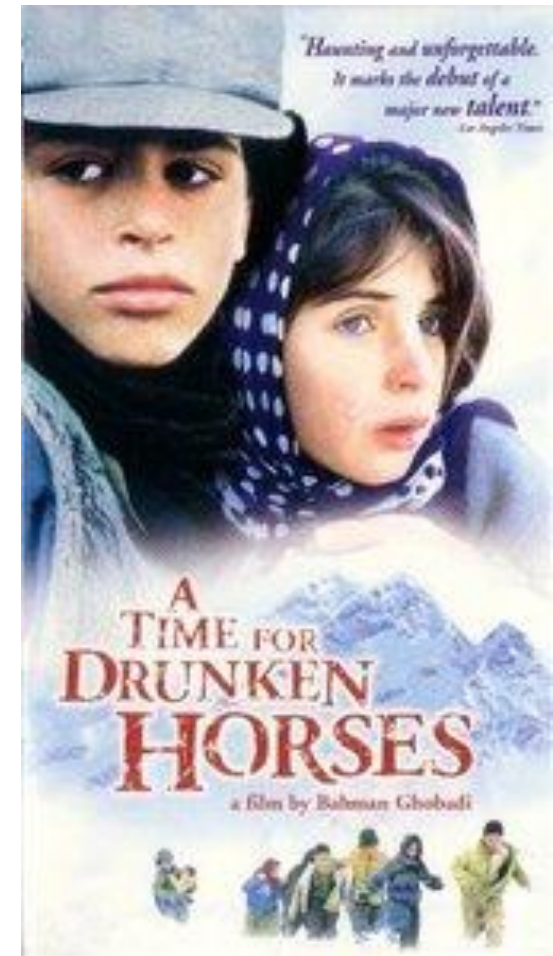


## Background

- Disparity between the speakers of various dialects of a language  
In language and speech technology (LST) development, priority is typically given to varieties and dialects with greater data representation
- Many studies have gone beyond the monolithic concept of a language (Ziems et al., 2022)
- LST for dialects and varieties is challenging (Zampieri et al., 2020):
  - Differences in written language: orthographic supremacy (Lew, 2012)
  - Lexical variations: more than 10 words for “hedgehog” in Kurdish!
  - Loanwords and terminologies (“*velo*” in Swiss German vs. “*Fahrrad*”)
  - typological variations
  - Lack of data

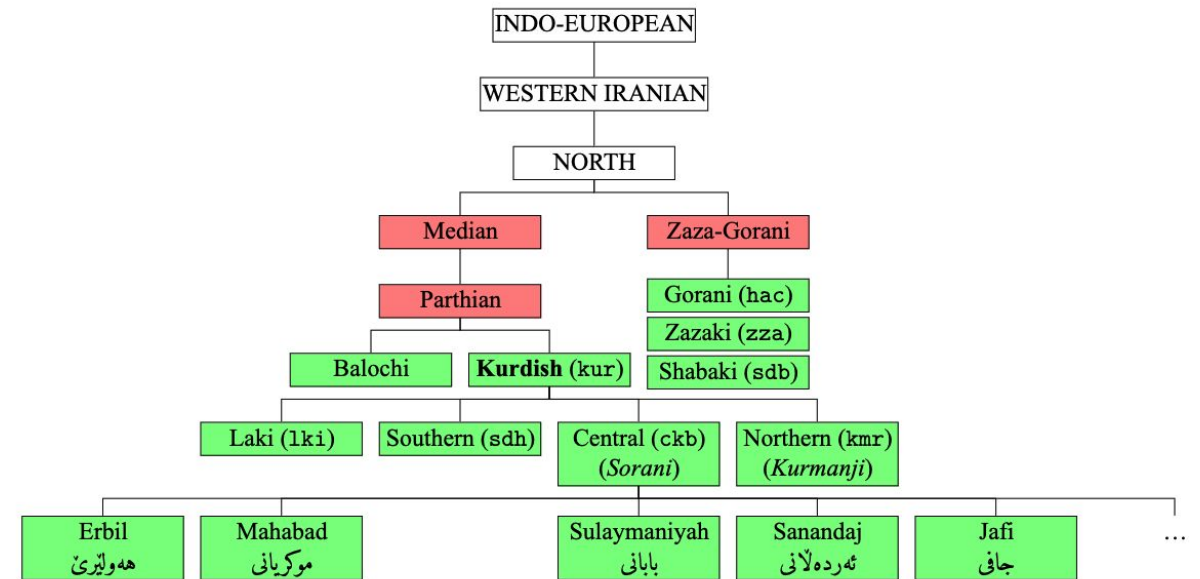
## Background: Creating a corpus for dialects

- Conditions:
  - a. A dialect continuum
  - b. Low-resourced language
  - c. You have €0 funding
  - d. Passionate volunteers 🥰
- Possible solutions:
  - a. Crawl the web → data paucity ❌
  - b. Fieldwork → time and resources ❌
  - c. Textbooks and articles → not available ❌
  - d. Crowdsourcing → expertise ❌
  - e. **Use dialogues in movies to document dialects!** ✅



## Background: Central Kurdish Dialects

Kurdish, an Indo-European language spoken by over 40 million speakers, is considered a dialect continuum and known for its diversity





## Methodology

### **CORDI – a text and audio corpus by transcribing movies and series.**

1. **Data Collection:** identify material and classify based on dialects  
Sulaymaniyah, Erbil, Kalar, Sanandaj, Mahabad and Sardasht
2. **Audio Transcription:** Using Amara (<https://amara.org/>) for transcription, native annotators were guided to transcribe dialogues while keeping meta-data for each utterance: (age, gender and dialect)
3. **Corpus Creation:** Downloading and converting content, then segmenting utterances according to the beginning and ending timecodes in the transcriptions
4. **Corpus Statistics:** 186,038 utterances among which 184,805 utterances are synchronized in text and audio.

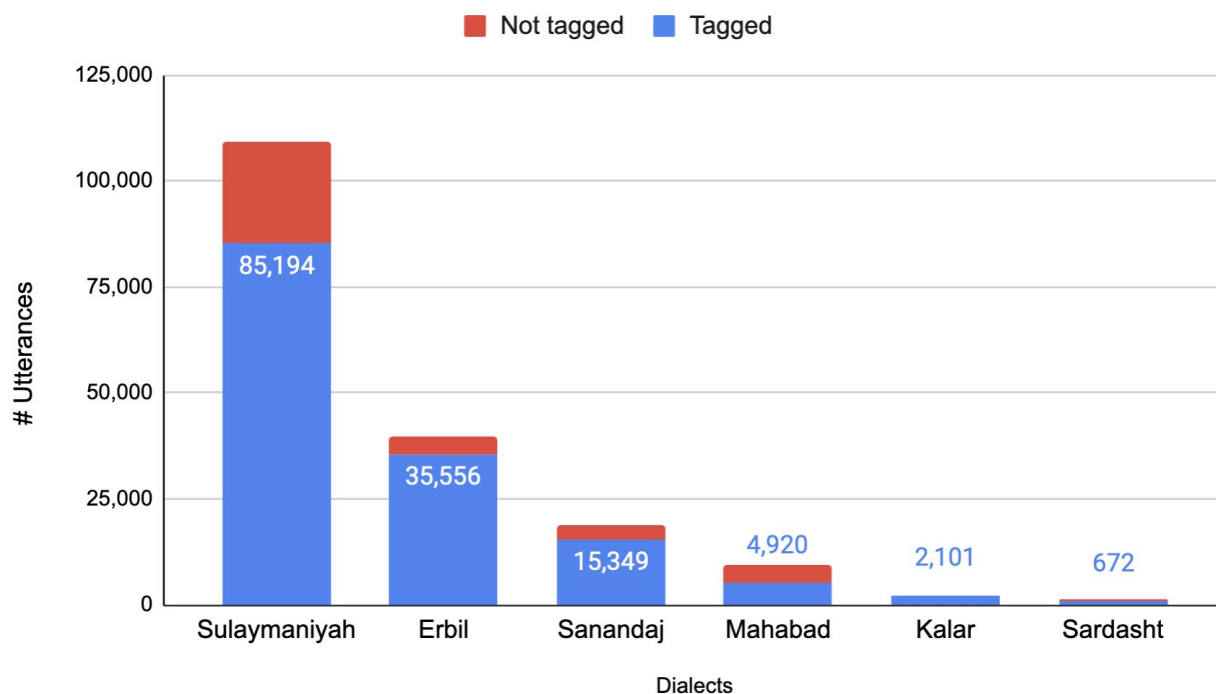


## Methodology



## Methodology: Corpus Statistics

- Over 180,000 utterances in six dialects (> 100 hours of dialogue)

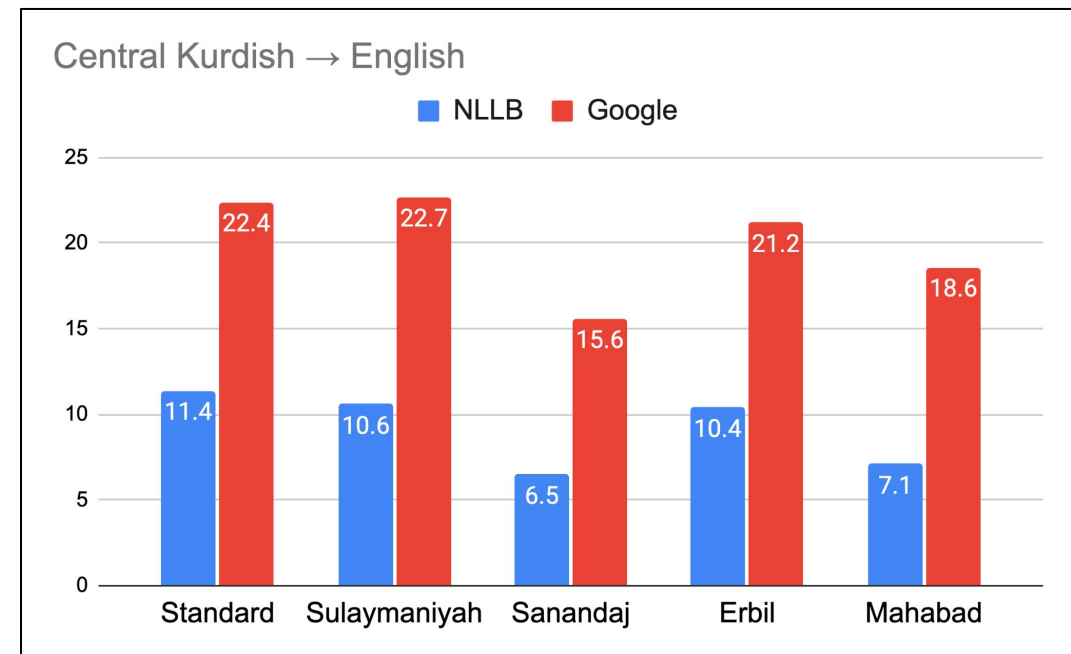
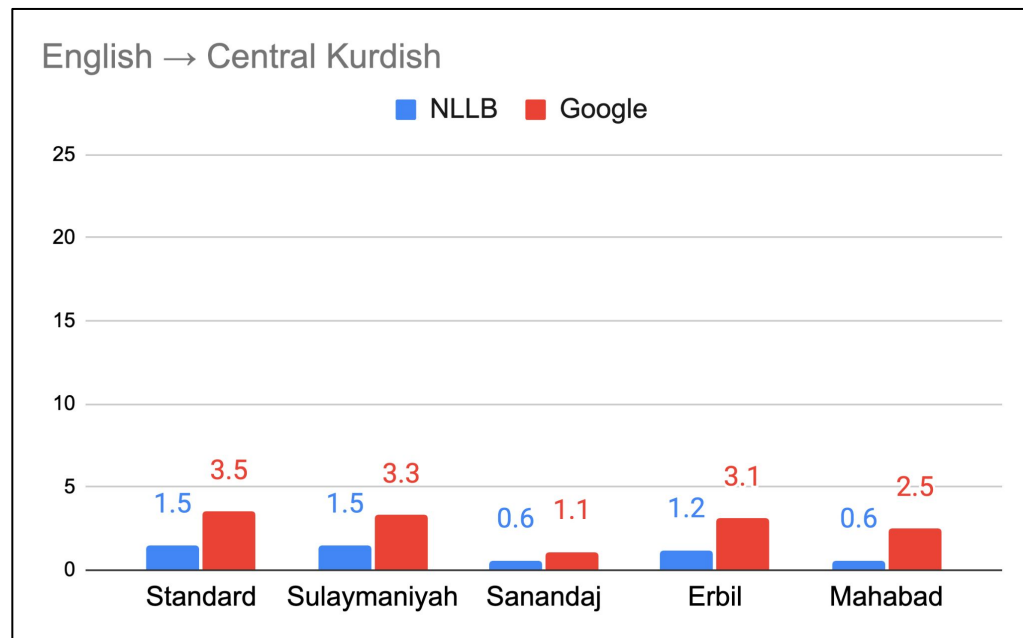


Variety	Ave. tokens	Ave. length (seconds)
Sulaymaniyah	9.06	2.39
Sanandaj	9.53	2.47
Erbil	7.78	1.68
Mahabad	8.45	2.2
Kalar	10.92	2.88
Sardasht	7.97	2.29
<b>Total</b>	<b>8.95</b>	<b>2.32</b>



## Experiments: Machine Translation

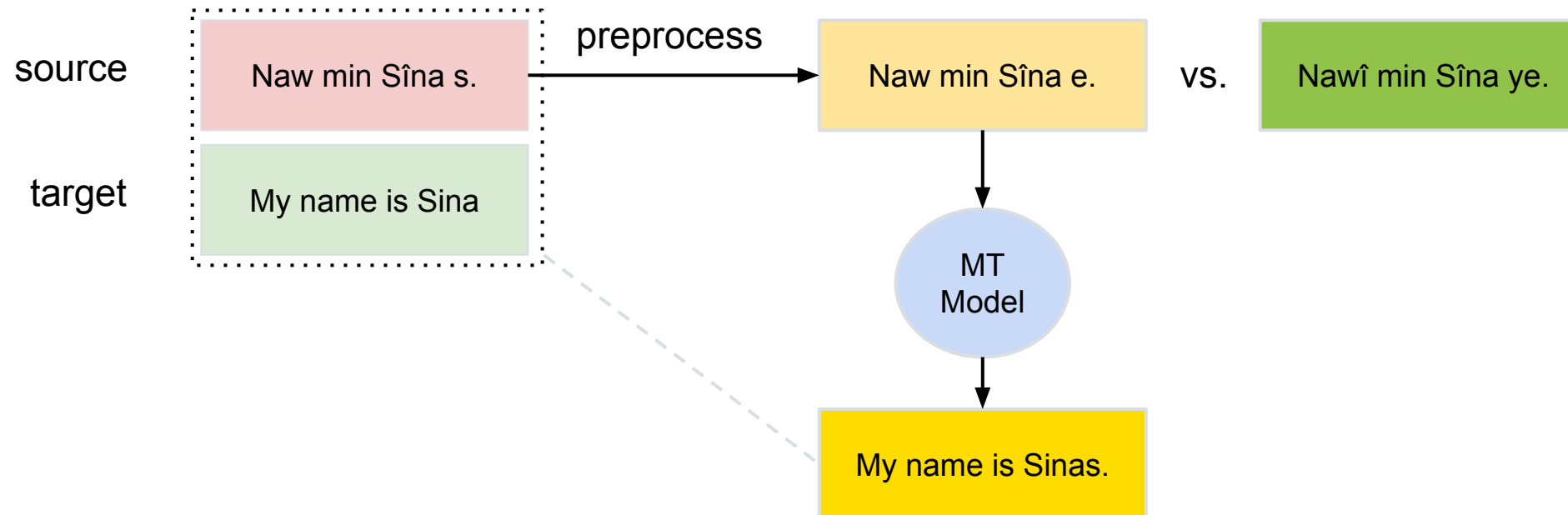
- Creating a parallel corpus containing 300 sentences in four sub-dialects and English translation
- Google Translate and Bing Microsoft Translator support Northern and Central Kurdish
- Previous research has targeted Northern and Central Kurdish (Ahmadi et al. (2022), Ahmadi and Masoud (2020), and Amini et al. (2021))
- **How existing models perform on Central Kurdish (sub)dialects?**



## Experiments: Machine Translation - Standardization

Using rules, convert sentences in a dialect to Standard Central Kurdish (\*\* synthetic sentences)

- Apply morphosyntactic rules
- Map Vocabulary
- Replace Terminology

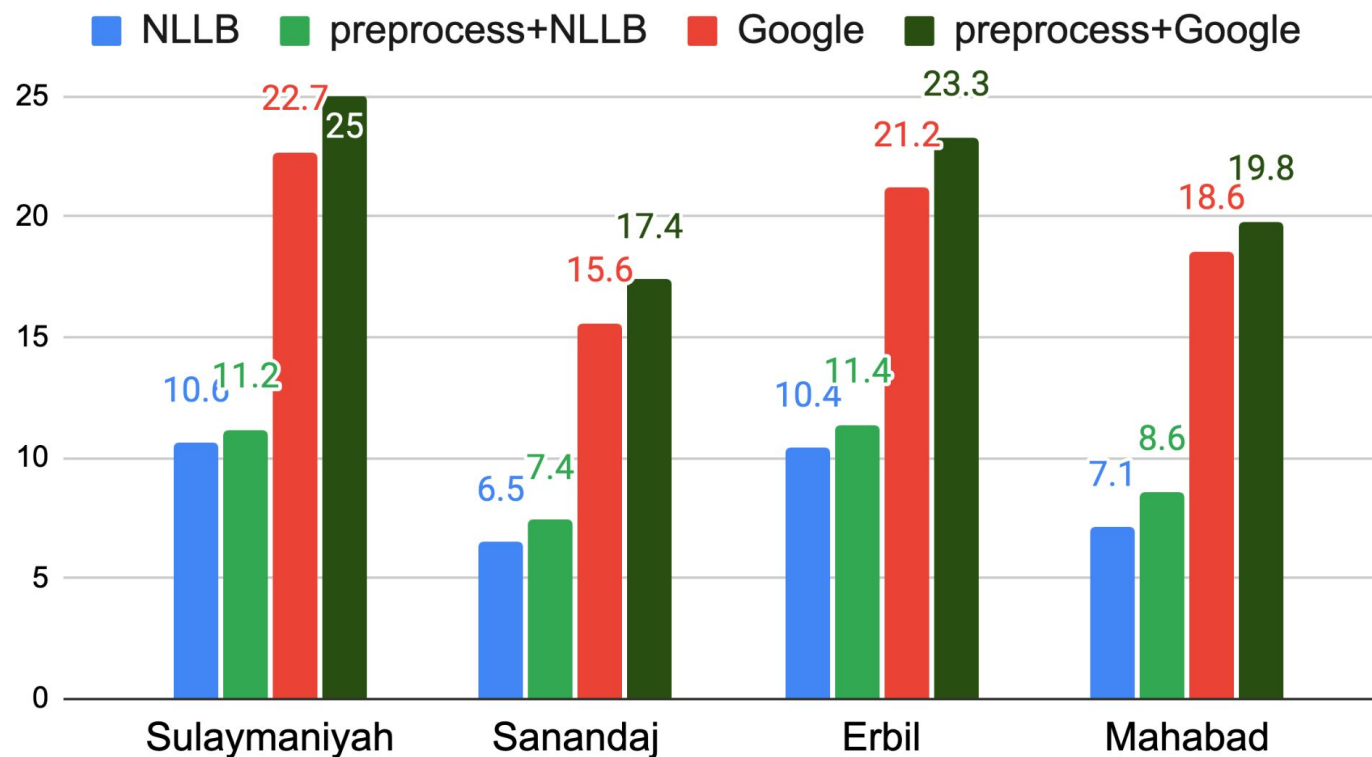




## Experiments: Machine Translation - Standardization

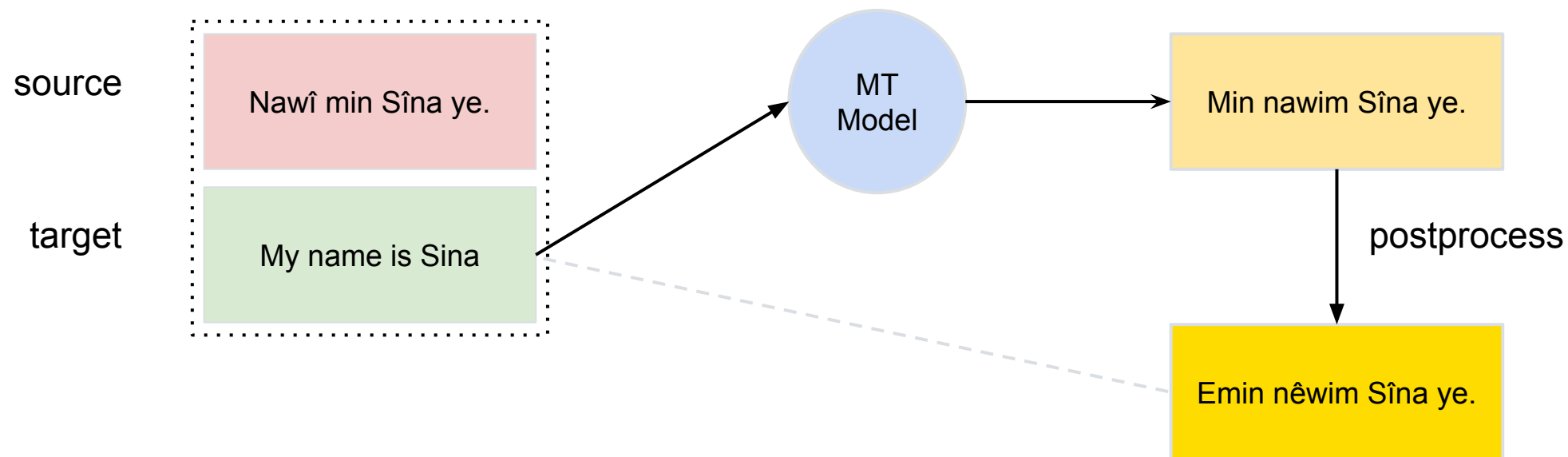
Using rules, convert sentences in a dialect to Standard Central Kurdish (\*\* synthetic sentences)

Preprocessing + Central Kurdish → English



## Experiments: Machine Translation - Dialectalization

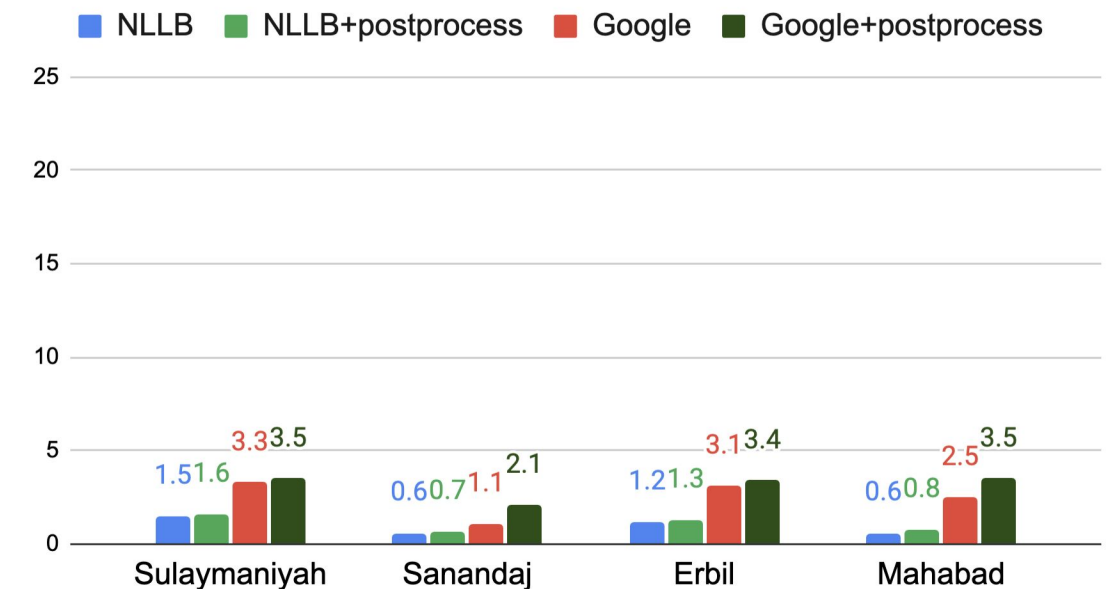
Using rules, convert sentences from Standard Central Kurdish into one of the dialects



## Experiments: Machine Translation - Dialectalization

Using rules, convert sentences from Standard Central Kurdish into one of the dialects

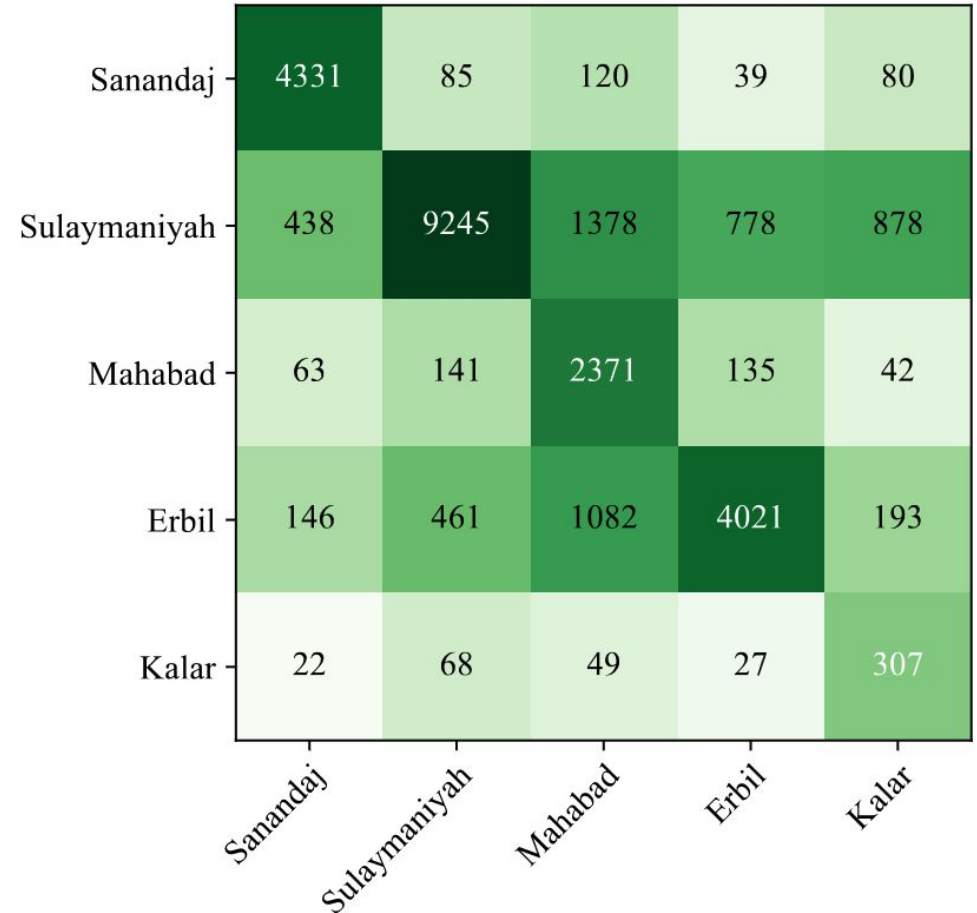
English → Central Kurdish + Postprocessing



- Google Translate demonstrates increased resilience to dialectal variations, surpassing the established baseline.
- our postprocess and preprocess approaches yield modest quality improvements
- Still a lot of room for improvement

## Experiments: Language Identification (LID)

- Use CORDI for training and testing LID
- Performance:
  - Detecting dialect: fastText predicts the language (Central Kurdish) with 0.94 F1
  - Detecting subdialect: our model predicts subdialects with 0.76 F1
- models confuse sentences in subdialects with other varieties, notably Southern Kurdish and Gorani



Sanandaj	4331	85	120	39	80
Sulaymaniyah	438	9245	1378	778	878
Mahabad	63	141	2371	135	42
Erbil	146	461	1082	4021	193
Kalar	22	68	49	27	307
	Sanandaj	Sulaymaniyah	Mahabad	Erbil	Kalar





## Conclusion

- Present a novel approach for creating an audio and text corpus for Central Kurdish subdialects called CORDI
- existing models for MT and LID exhibit suboptimal performance when subjected to evaluation on subdialects
- our resources pave the way for further advances in Kurdish NLP
- additional advancements are imperative to address nonstandard NLP effectively

**This project received funding of**



Many low-resourced languages face financial constraints and Kurdish is regrettably no exception.



# Heartfelt gratitude to the 39 volunteers who actively participated in the transcription and annotation tasks from June 2021 to April 2022.





## References

- Ahmadi, S. (2020, November). KLPT–Kurdish language processing toolkit. In Proceedings of second workshop for NLP open source software (NLP-OSS) (pp. 72-84).
- Ahmadi, S., & Masoud, M. (2020, December). Towards machine translation for the Kurdish language. In Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages (pp. 87-98).
- Ahmadi, S., Hassani, H., & Jaff, D. Q. (2022). Leveraging multilingual news websites for building a kurdish parallel corpus. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5), 1-11.
- Amini, Z., Mohammadamini, M., Hosseini, H., Mansouri, M., & Jaff, D. (2021). Central Kurdish machine translation: First large scale parallel corpus and experiments. arXiv preprint arXiv:2106.09325.
- Robert Lew. *How can we make electronic dictionaries more effective?* Oxford University Press, 2012
- Vaibhav, V., Singh, S., Stewart, C., & Neubig, G. (2019). Improving robustness of machine translation with synthetic noise. arXiv preprint arXiv:1902.09508.
- Zampieri, M., Nakov, P., & Scherrer, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6), 595-612.
- Ziems, C., Held, W., Yang, J., Dhamala, J., Gupta, R., & Yang, D. (2022). Multi-VALUE: A framework for cross-dialectal English NLP. arXiv preprint arXiv:2212.08011.