University of Zurich[UZH]

Dialogues for Documenting Dialects

Language and Speech Technology for Central Kurdish Varieties

Sina Ahmadi[1], Daban Q. Jaff[2], Md Mahfuz Ibn Alam[3], Antonios Anastasopoulos[3,4]

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

- 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

- Possible solutions:

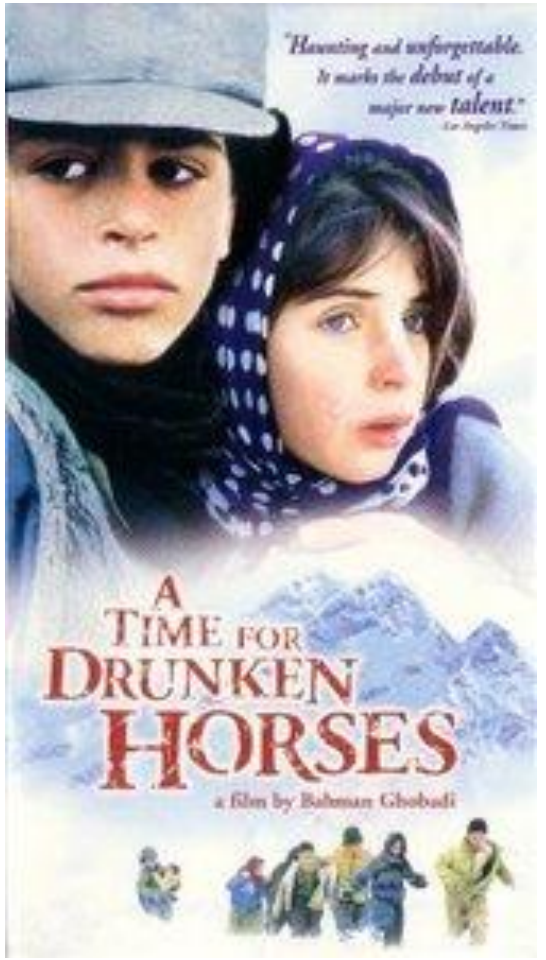
  a. Crawl the web → data paucity ❌

  b. Fieldwork → time and resources ❌

  c. Textbooks and articles → not available ❌

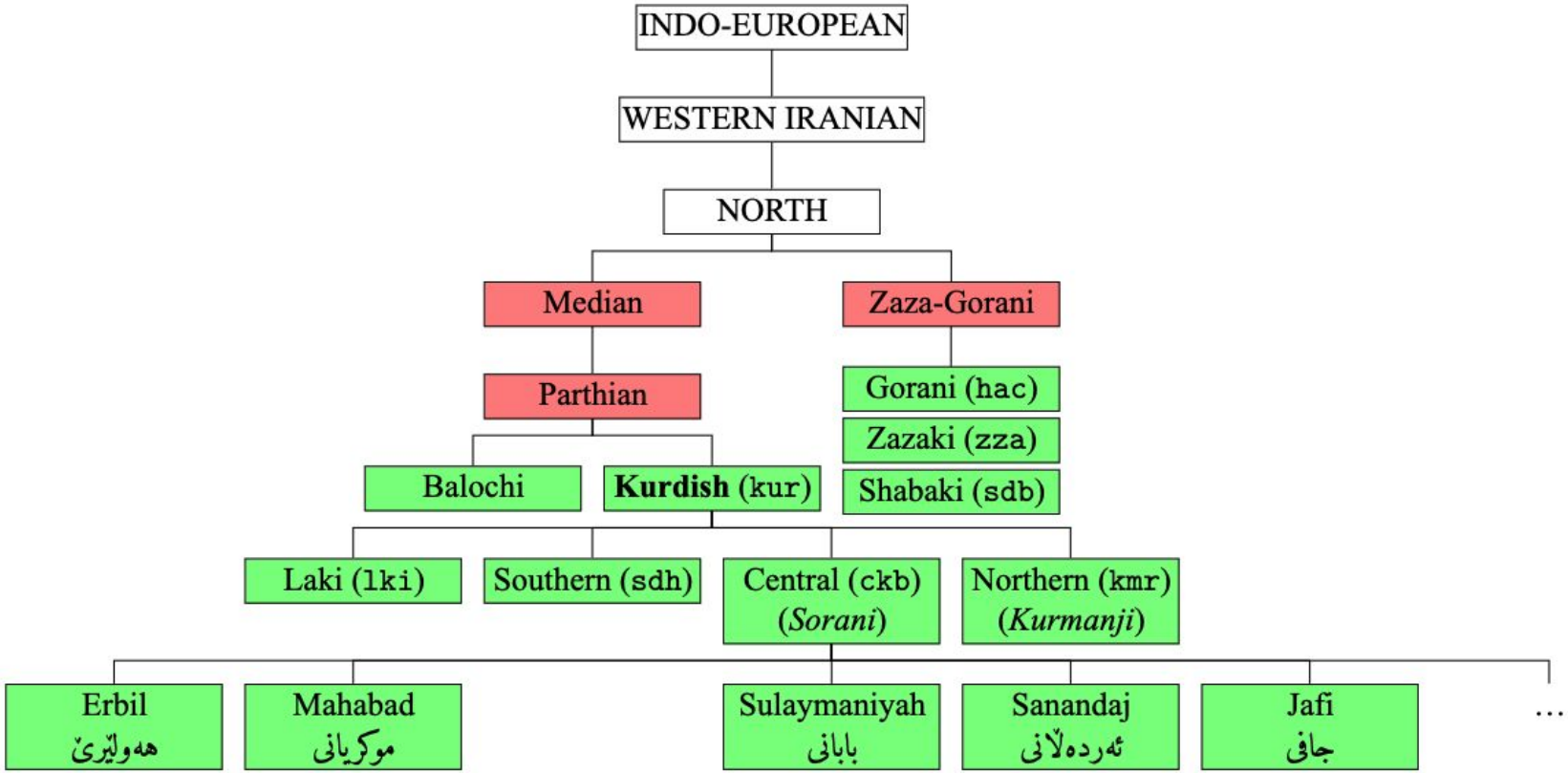
  d. Crowdsourcing → expertise ❌

  e. **Use dialogues in movies to document dialects!** ✅



Background: Central Kurdish Dialects

Kurdish, an Indo-European language spoken by over 30 million speakers, is considered a dialect continuum and known for its diversity in language varieties



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

Variety	# Utterances	length (hours)	Ave. tokens	Ave. length (seconds)	Speaker metadata (%)
Sulaymaniyah	115,083	64.44	9.06	2.39	78.1
Sanandaj	18,584	18.57	9.53	2.47	82.59
Erbil	39,674	11.2	7.78	1.68	89.62
Mahabad	9,410	4.3	8.45	2.2	52.28
Kalar	2,150	1.22	10.92	2.88	97.72
Sardasht	1,137	0.42	7.97	2.29	59.1
Total	186,038	100.15	8.95	2.32	76.56

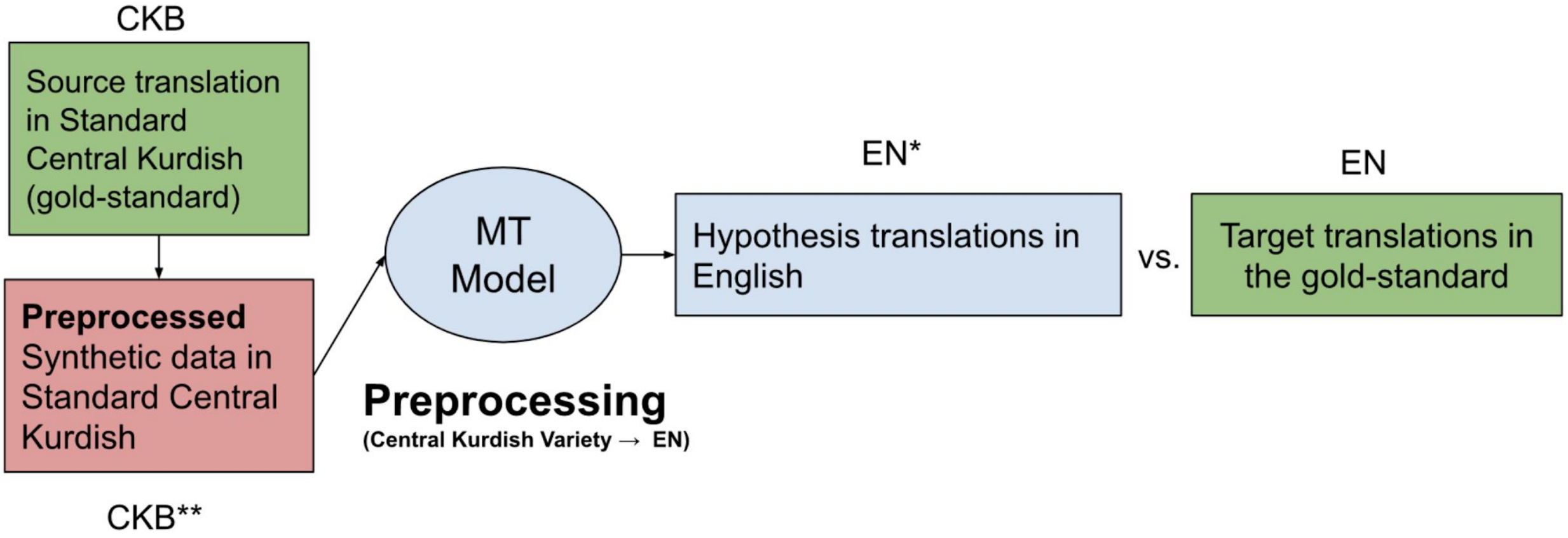
Experiments: Machine Translation

- 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?**

	English → Central Kurdish		Central Kurdish → English	
	NLLB Baseline	Google Baseline	NLLB Baseline	Google Baseline
Standard	1.5 (25.5)	3.5 (33.6)	11.4 (28.6)	22.4 (42.9)
Sulaymaniyah	1.5 (25.3)	3.3 (33.2)	10.6 (27.9)	22.7 (43.3)
Sanandaj	0.6 (19.9)	1.1 (24.7)	6.5 (21.6)	15.6 (35.9)
Erbil	1.2 (24.5)	3.1 (31.2)	10.4 (27.6)	21.2 (41.9)
Mahabad	0.6 (22.5)	2.5 (29.3)	7.1 (24)	18.6 (39)

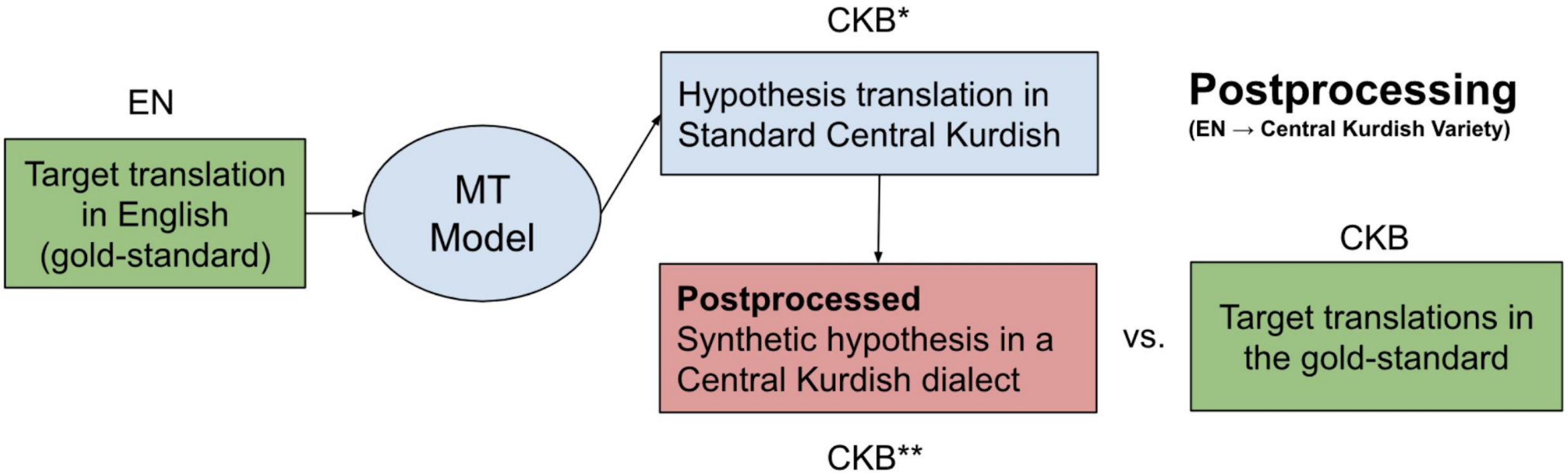
Experiments: Machine Translation - Standardization

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



Experiments: Machine Translation - Dialectalization

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



Experiments: Machine Translation

	English → Central Kurdish Variety				Central Kurdish Variety → English			
	NLLB		Google		NLLB		Google	
	Baseline	postprocess	Baseline	postprocess	Baseline	preprocess	Baseline	preprocess
Standard	1.5 (25.5)		3.5 (33.6)		11.4 (28.6)		22.4 (42.9)	
Sulaymaniyah	1.5 (25.3)	1.6 (26)	3.3 (33.2)	3.5 (33.9)	10.6 (27.9)	11.2 (28.5)	22.7 (43.3)	25 (43.2)
Sanandaj	0.6 (19.9)	0.7 (22)	1.1 (24.7)	2.1 (27.3)	6.5 (21.6)	7.4 (22.8)	15.6 (35.9)	17.4 (35.9)
Erbil	1.2 (24.5)	1.3 (25.3)	3.1 (31.2)	3.4 (31.7)	10.4 (27.6)	11.4 (28.5)	21.2 (41.9)	23.3 (42.1)
Mahabad	0.6 (22.5)	0.8 (23.9)	2.5 (29.3)	3.5 (30.8)	7.1 (24)	8.6 (25.2)	18.6 (39)	19.8 (38.8)

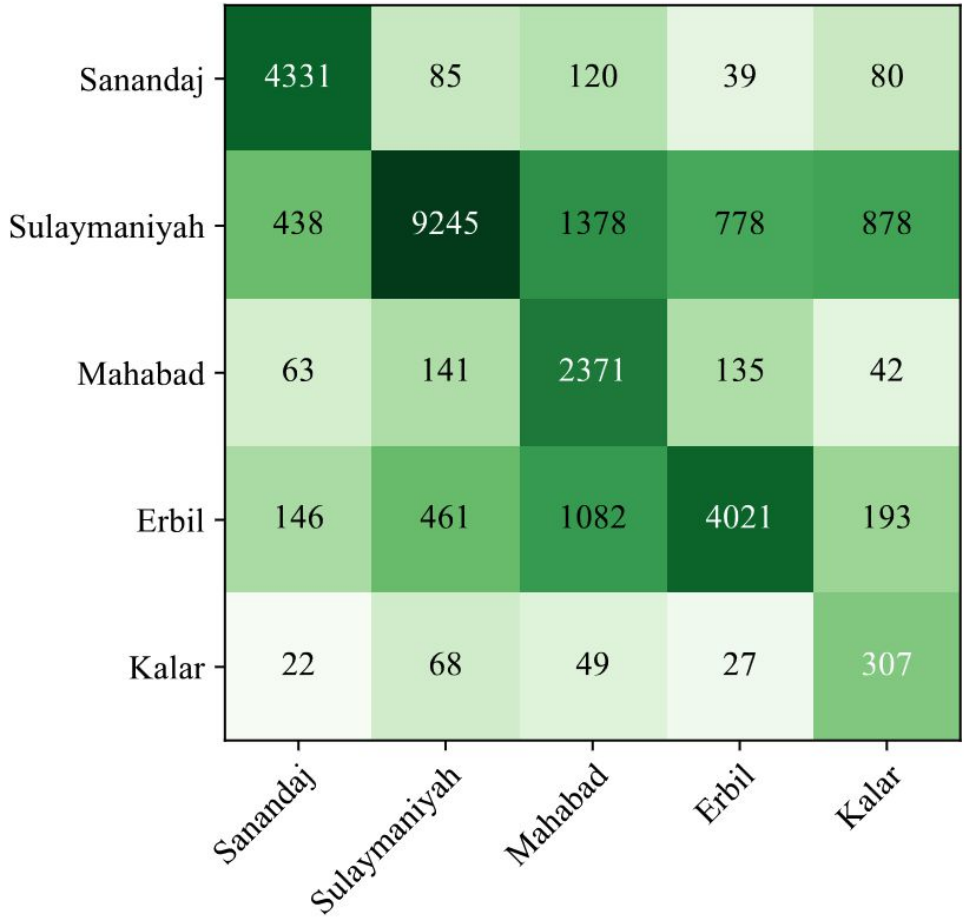
- Google Translate demonstrates increased resilience to dialectal variations, surpassing the established baseline.
- our postprocess and preprocess approaches yield modest quality improvements
- Still a lot of room for improvement

Experiments: Automatic Speech Recognition

Data	CV-Scratch	CV-PT-en	CORDI-Scratch	CORDI-PT-en	CORDI-PT-CV
Sulaymaniyah	125.42	112.11	**58.56**	62.9	60.97
Sanandaj	131.7	111.68	**58.98**	67.84	60.84
Erbil	117.67	114.4	72	74.96	**71.7**
Mahabad	128.43	116.03	**66.88**	68.32	69.69
CommonVoice	148.99	**85.6**	170.14	171.86	236.51

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



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.