

Approaches to Corpus Creation for Low-Resource Language Technology: the Case of Southern Kurdish and Laki



Sina Ahmadi¹ Zahra Azin² Sara Belleli³ Antonios Anastasopoulos¹

¹George Mason University, USA ²Carleton University, Canada ³University of Tuscia, Italy

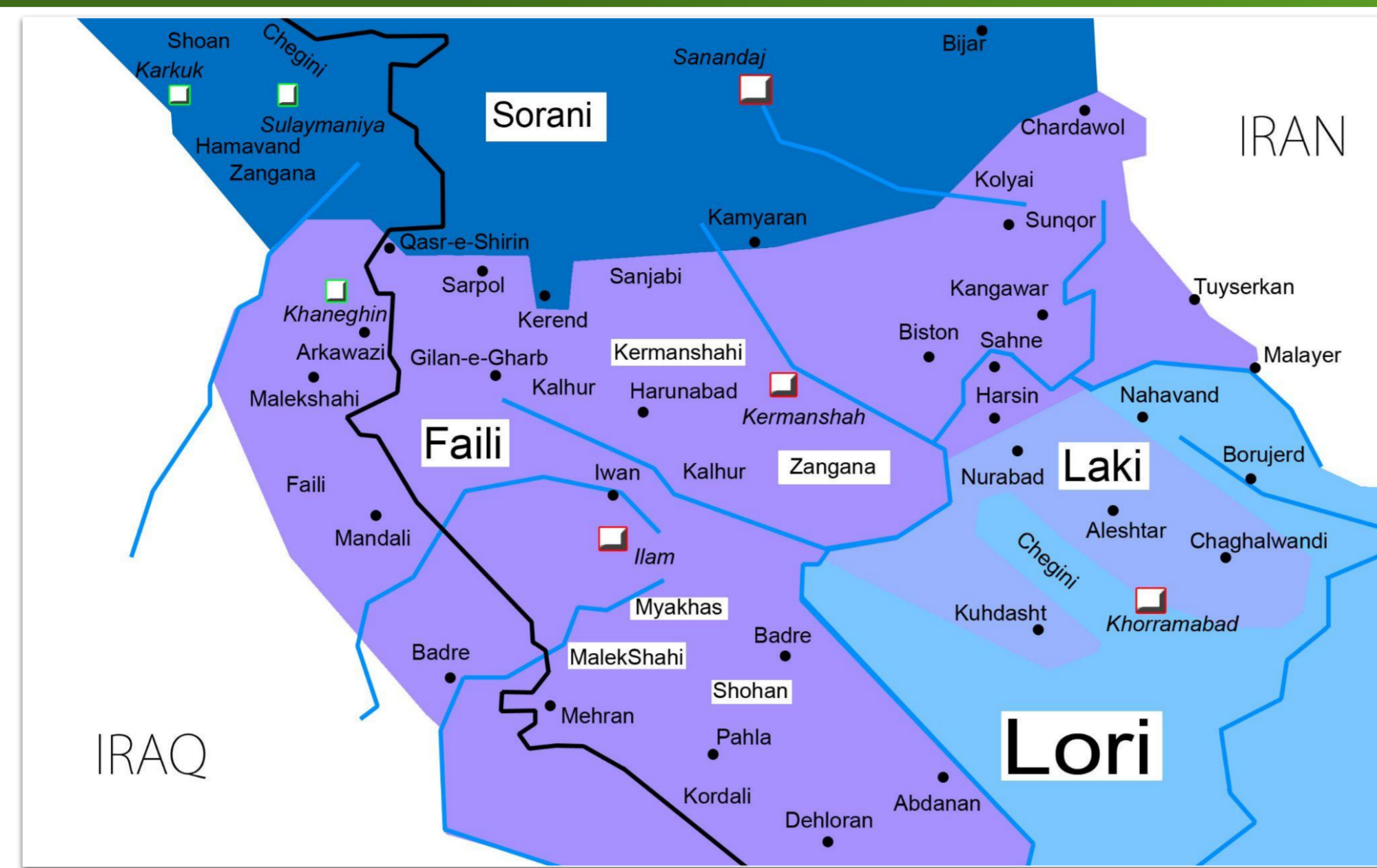
{sahmad46}@gmu.edu

Highlights



<https://github.com/sinaahmadi/KurdishLID>
Email us for any questions!

Southern Kurdish and Laki



- ❑ Southern Kurdish as one of the main branches of Kurdish
- ❑ The classification of Laki is still debated
- ❑ Both languages:
 - ❑ Face various **discriminatory language policies**
 - ❑ Sociolinguistic effects on language attitudes and heritage language maintenance
 - ❑ Lack of standardization
 - ❑ Limited linguistic resources
 - ❑ Speakers have inadequate technology access
 - ❑ Few available digital resources are available
 - ❑ There are no processing tools
 - ❑ Lack of funding

Highlights

- ❑ **Task: Corpus Creation for Low-Resource Language Technology**
- ❑ Southern Kurdish and Laki
- ❑ Understanding the impact of sociolinguistic factors on underrepresented languages
- ❑ Data collection approaches for low-resource languages
- ❑ Preservation and promotion of endangered languages

Data Collection and Corpus Creation

Radio Shows

- ❑ Dialects of Kermanshah (Iran)
- ❑ Local radio broadcaster
- ❑ 18 handwritten scenarios of radio shows in the form of dialogue
- ❑ Educational, cultural and daily topics
- ❑ Manually typing the content
- ❑ Original scenarios in the Persian script
- ❑ Central Kurdish Perso-Arabic script

News Articles

- ❑ Crawling a news website that publishes articles in Feyli
- ❑ 15,985 articles were collected in HTML and converted to text
- ❑ Preprocessing the raw text
 - ❑ unifying character encoding
 - ❑ removing private information
 - ❑ categorizing articles by topic
- ❑ Integrating metadata including source, topic, title, and date of publication for each article

Fieldwork

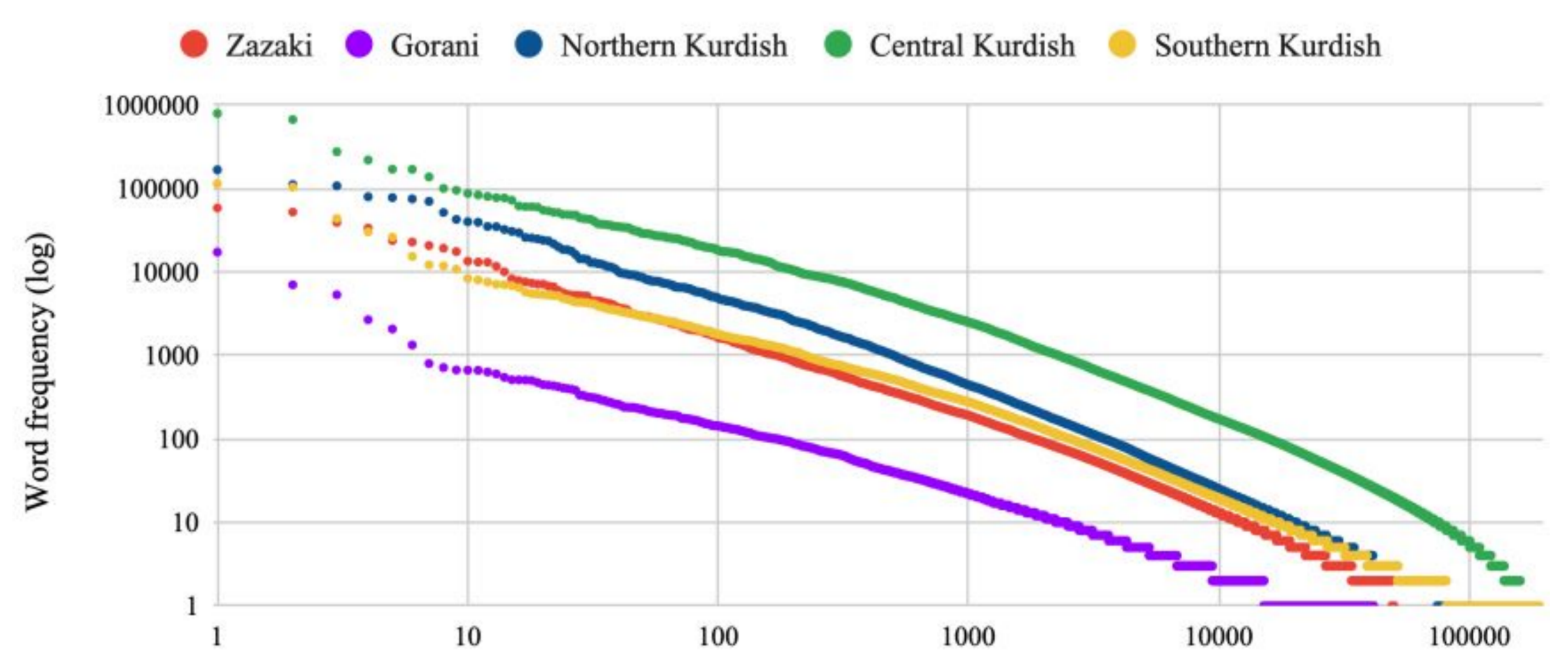
- ❑ Fieldwork conducted to document Laki language in Harsin city, western Iran
- ❑ 7 traditional narratives, 4 speakers
 - ❑ 5 folktales
 - ❑ 2 anecdotes
 - ❑ in monologue form
- ❑ Texts manually transcribed using a conventional transcription system
- ❑ One text interlinearized with morpheme-by-morpheme glosses

Our Contributions

- ❑ 3 approaches to corpus creation for Southern Kurdish and Laki
- ❑ Fieldwork
- ❑ Crawling
- ❑ Local broadcasters
- ❑ Quantitative and qualitative analysis
- ❑ Language identification

Analysis

- ❑ Issues in script normalization
- ❑ Inconsistent orthographies
- ❑ Code-switching
- ❑ Different topics and genres of text
- ❑ Cross-dialectal/lingual analyses



Zipfian distribution of Pewan corpora of Northern and Central Kurdish, a corpus of Zaza-Gorani and our corpus of Southern Kurdish

Takeaways

- ❑ Paucity of resource: **Tip of the iceberg**
- ❑ Everyone has a role:
 - ❑ Policymakers
 - ❑ Native speakers
 - ❑ NLPers
- ❑ Make it open-source!

A Downstream Task: Language Identification

Turkish	600	10	1	0	0	1	1	0
Zazaki	0	1163	6	7	0	0	0	0
Northern Kurdish	0	22	1151	12	4	17	1	0
Central Kurdish	0	5	12	1170	3	8	0	0
Southern Kurdish	0	0	12	7	583	22	0	0
Gorani	0	0	16	4	6	548	1	1
Persian	0	0	2	0	4	4	595	4
Arabic	0	0	0	0	0	0	2	595
	Turkish	Zazaki	Northern Kurdish	Central Kurdish	Southern Kurdish	Gorani	Persian	Arabic

(a) Classification with language codes

Turkish (L)	594	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zazaki (W)	1	579	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zazaki (L)	1	17	541	0	0	0	0	0	3	6	0	0	0	0	0	0	0	0
Northern Kurdish (A)	0	0	0	561	5	6	15	0	0	0	0	0	0	0	0	0	0	0
Central-Kurdish (A)	0	0	0	6	586	4	12	0	0	0	0	0	0	0	0	0	0	3
Southern Kurdish (A)	0	0	0	22	8	583	23	0	0	0	0	0	0	0	0	0	0	0
Gorani (A)	0	0	1	11	1	7	546	0	0	0	1	0	0	0	0	0	0	0
Central Kurdish (L)	3	2	8	0	0	0	0	0	0	0	596	5	0	0	0	0	0	0
Northern Kurdish (L)	0	1	31	0	0	0	0	0	0	1	589	0	0	0	0	0	0	0
Persian (A)	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	595	7	0
Arabic (A)	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	4	590	0
	Turkish (L)	Zazaki (W)	Zazaki (L)	Northern Kurdish (A)	Central-Kurdish (A)	Southern Kurdish (A)	Gorani (A)	Central Kurdish (L)	Northern Kurdish (L)	Persian (A)	Arabic (A)							

(b) Classification with language and script codes