FieldMatters - EACL 2023 - Dubrovnik, Croatia

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

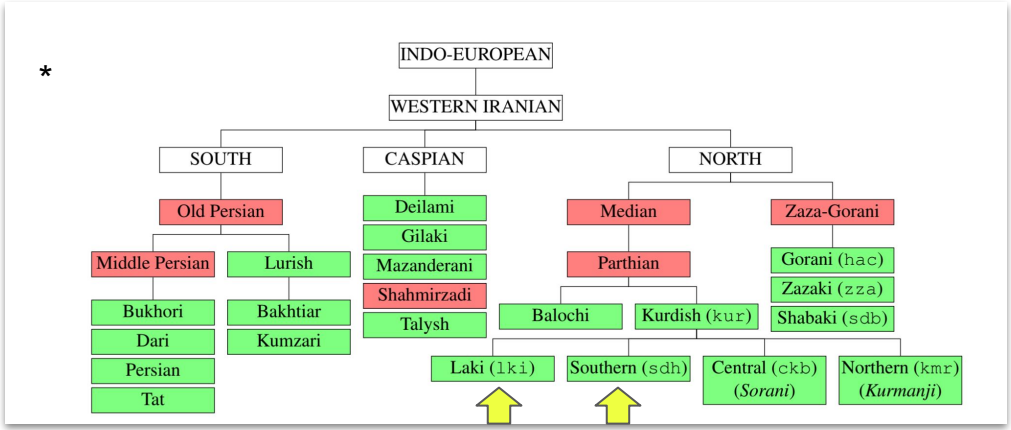
Sina Ahmadi
George Mason University

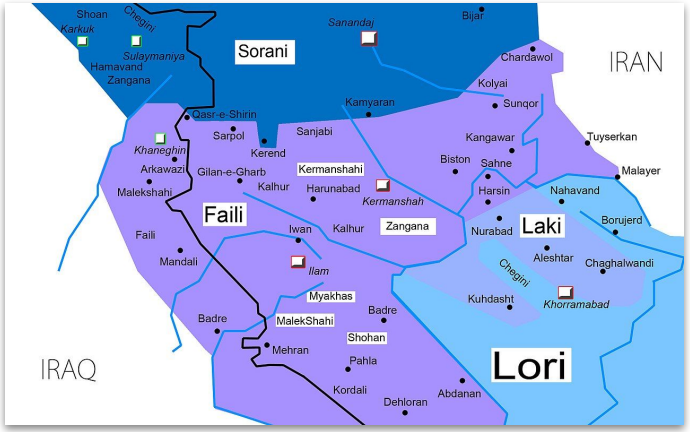
Zahra Azin
Carleton University

Sara Belelli
University of Tuscia

Antonios Anastasopoulos
George Mason University

Southern Kurdish and Laki





- Southern Kurdish is known as one of the main branches of Kurdish
- Spoken in western Iran, border regions of Iraq, and its capital, Baghdad
- The classification of Laki is still debated
- Faced various discriminatory language policies leading to pernicious sociolinguistic effects
- the lack of children's proficiency in Southern Kurdish and limited usage of the language in writing

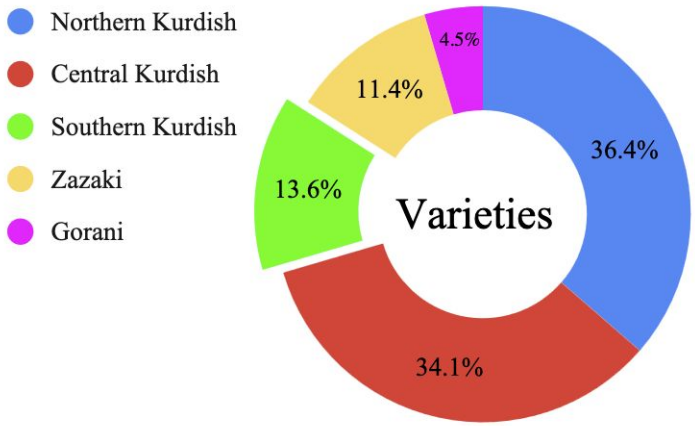
* Source: Ahmadi, Sina. "Building a corpus for the Zaza–Gorani language family." In *Proceedings of the 7th workshop on NLP for similar languages, varieties and dialects*, pp. 70-78. 2020.

Some of the Challenges of Southern Kurdish & Laki

- Lack of standardization: how to write?
- Lack of linguistic resources
- Limited access to technology by the speakers
- Lack of funding

Language Technology for Southern Kurdish & Laki

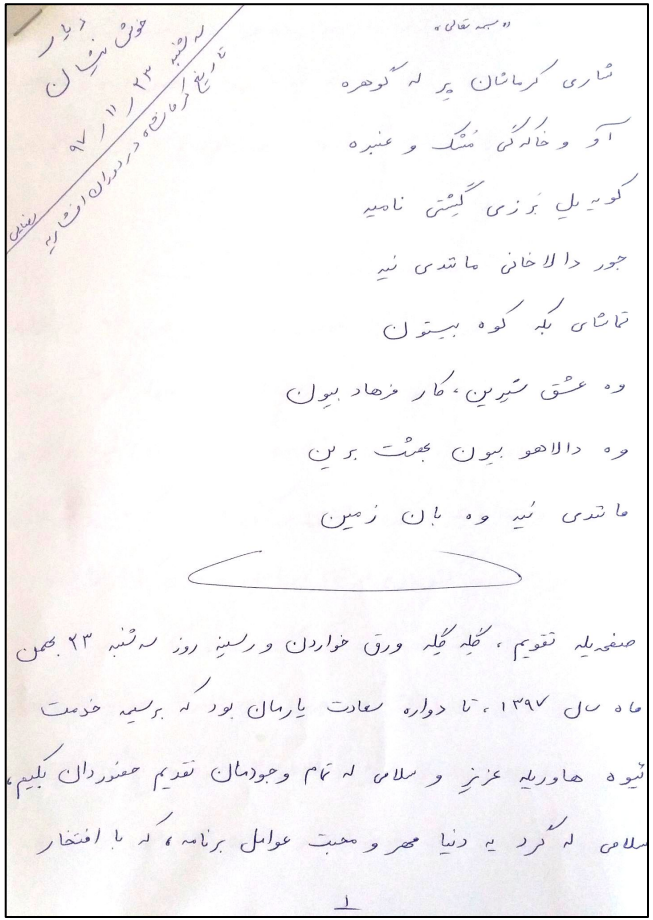
- Not received much attention in (computational) linguistics
- Few available digital resources available
- No tools for processing these varieties
- Lack of data



Percentage of the existing lexicographical resources for Kurdish

Approach 1: Radio shows

- Local radio broadcaster in Kermanshah province (Iran)
- Collect a set of handwritten scenarios of radio shows
- Scenarios cover educational, cultural and daily topics
- Digitize them by manually typing the content
    - Original scenarios in the Persian script
    - What script?
    - Central Kurdish Perso-Arabic script
- 18 scenarios are digitized for talk shows and short comedies, mostly in the form of dialogue



A sample of the original handwritten scenarios of the local radio broadcaster

Approach 2: 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

Approach 3: Fieldwork

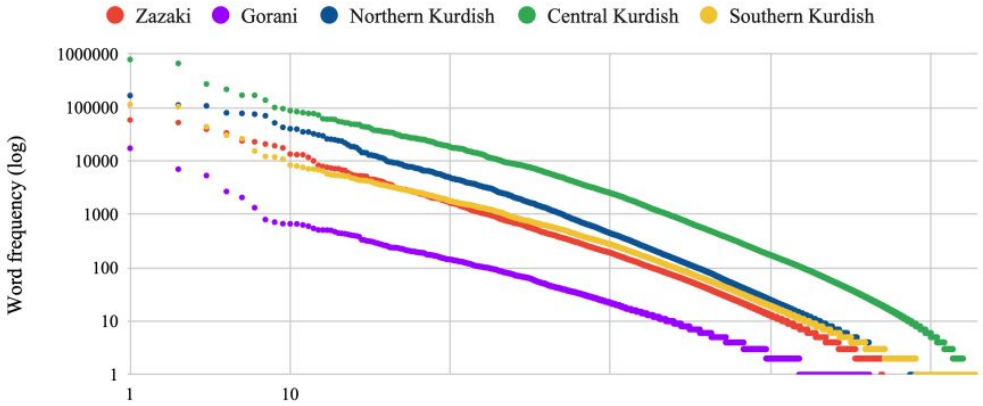
- Fieldwork conducted to document Laki language in Harsin city, western Iran
- Corpus includes 7 traditional narratives
  - 5 folktales
  - 2 anecdotes
  - in monologue form
  - recorded from 4 native speakers
- Texts manually transcribed using a conventional transcription system
- One text interlinearized with morpheme-by-morpheme glosses
- No standard writing system or orthography for Laki

Quantitative Analysis

- The collected data contains 16,003 documents written in varieties of Southern Kurdish and seven narratives in Laki-Kermanshahi.
- Differences in the average type length as an indicator of the morphological complexity of word forms can be due to:
    1. The orthography of the Southern Kurdish corpus
    2. Conventions in writing multiword expressions
    3. Excessive concatenation of words

Number (#)	Kermanshahi	Feyli	Laki
articles	18	15,985	7
tokens	10,127	2,182M	6,340
types	3,248	179,208	2,074
characters	21,359	1,591M	13,378
average length	6.57	8.8	6.45

Quantitative Analysis



Rank-size distribution for Pewan corpus of Northern and Central Kurdish (2013), Zaza-Gorani corpus (2020), and our Southern Kurdish data

Northern Kurdish	Central Kurdish	Southern Kurdish		Laki	Gorani	Zazaki
		Feyli	Kermanshahi			
û (and)	*le* (from, in)	*e* (is)	*û* (and)	*muše* (IND-SAY.PRS-3SG)	*û* (and)	*de* (in)
ku (that)	*û* (and)	*û* (and)	*we* (and)	*ī* (this, these)	*ce* (in)	*û* (and)
li (from, in)	*bo* (for)	*ki* (that)	*le* (in)	*ye* (a, an)	*be* (to, with)	*ke* (that)
bi (with, to)	*be* (with, to)	*we* (and)	*abadî* (village)	*ařā* (for)	*ke* (that)	*ra*
di (in)	*ke* (that)	*era* (for)	*naw* (in; name)	*va* (to)	*pey* (for)	*bi* (with)
ji (from)	*ew* (that)	*ew* (that)	*wegerd* (with)	*maču* (IND-GO.PRS-3SG)	*y*	*ma* (we)
de	*Kurdistan*	*kird* (ind-do.pst-3sg)	*ta* (until)	*ya* (this, this one)	*ta* (until)	*xo* (self)
jî (too)	*Iraq*	wit (IND-SAY.PST-3SG)	*ê* (this)	*a* (yes; that)	*î* (this)	*zî* (too)
Kurdistanê	*em* (this)	*herêm* (region, region of)	*î* (this)	*make* (IND-DO.PRS-3SG)	*Kurdistanî*	*yê*
Iraqê	*herêmî* (region of)	*Kurdistan*	*weşûn* (after)	*mi* (me, mine)	*her* (each)	*mi* (my)
herêma (region of)	*serokî* (president of)	*ta* (until)	*bün* (IND-BE.PST-3PL)	*nām* (name)	*Turkyay* (Turkey)	*o* (that, it)

Qualitative Analysis

- Differences in textual structure
- Lexical borrowings from other regional languages
- Article headlines for discourse analysis
- The narrative in the Laki data useful to analyze folkloric stories
- Cross-dialectal and cross-lingual analyses

A downstream task: Language Identification

- Given a text, identify the language/dialect
- Train a model on the following languages:
  - Kurdish (Northern, Central, Southern)
  - Zazaki & Gorani
  - Arabic, Persian & Turkish
- We did not include the Laki data due to script issues
- We evaluate the pre-trained language identifier of ***fastText*** (Bojanowski et al., 2017) as baseline system
- Our model was trained using *fastText*

Measure	`lid.176`	Our model		
		language code	language & script code	SDH-unconventional
Precision	0.0552	0.969	0.9638	0.25
Recall	0.0674	0.971	0.9636	0.126
F_1	0.06	**0.97**	0.9634	0.168

Conclusion

- Three approaches to data collection and corpus creation for Southern Kurdish and Laki:
  - Fieldwork
  - Web crawling
  - Radio scenarios
- Approaches can be adopted by other underrepresented languages with limited data
- A brief analysis from quantitative and qualitative perspectives + language identification

Future work:

- Annotating the corpora for other tasks:
  - part-of-speech tagging
  - named entity recognition
- Leverage data of other varieties to develop tools for Southern Kurdish and Laki & transfer learning


Thank you!
Check the repository of the project at
https://github.com/sinaahmadi/KurdishLID
© Photo by Sara Belelli