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

Sina Ahmadi[1] Zahra Azin[2] Sara Belelli[3] Antonios Anastasopoulos[1]

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

{sahmad46}@gmu.edu

Highlights

https://github.com/sinaahmadi/KurdishLID

Email us for any questions!

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

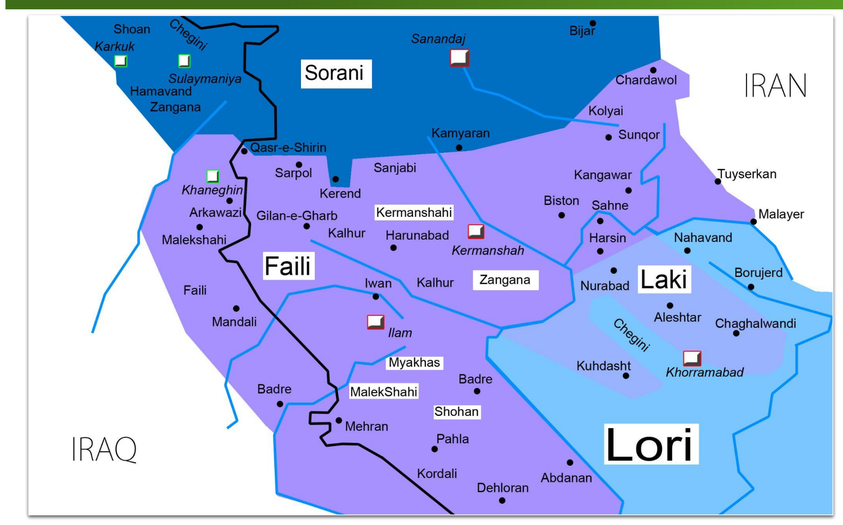
Our Contributions

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

Takeaways

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

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

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

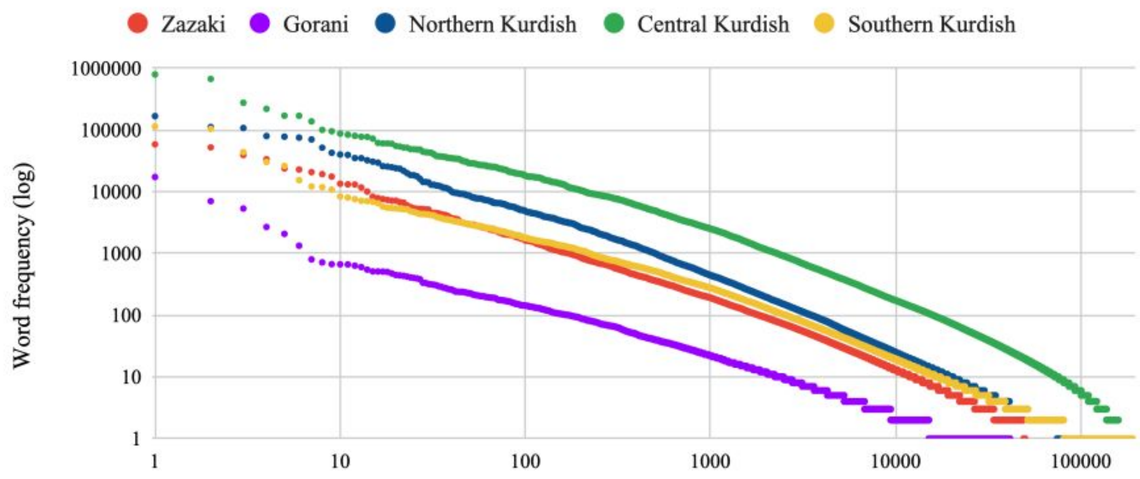
- Crawling a news website that publishes articles in Feyli
- 15,985 articles were collected in HTML and converted to text
- Preprocessing the raw text
  - unifying character encoding
  - removing private information
  - categorizing articles by topic
- Integrating metadata including source, topic, title,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

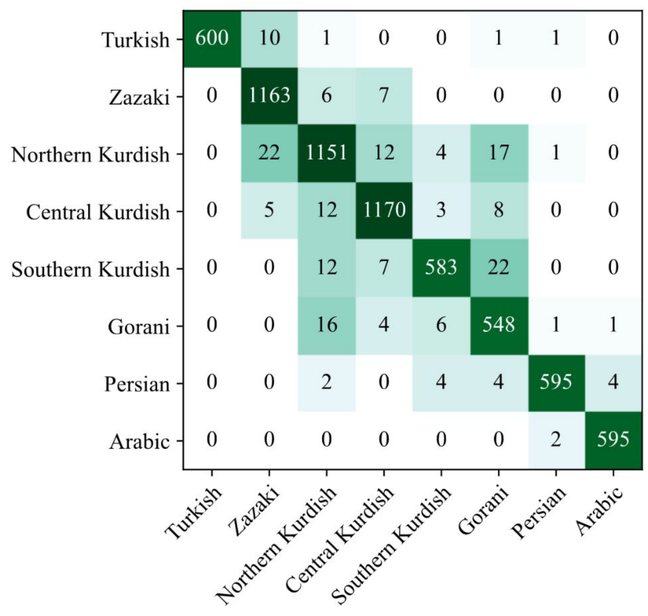
Analysis

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

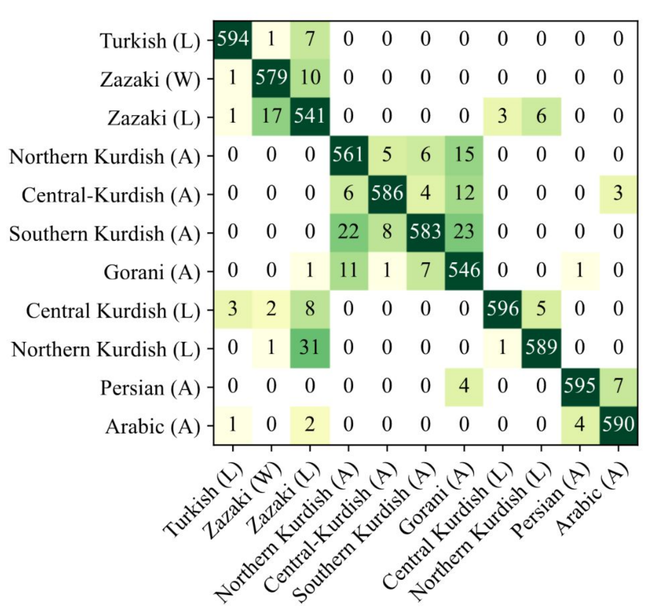


Zipfian distribution of Pewan corpora of Northern and Central Kurdish, a corpus of Zaza-Gorani and our corpus of Southern Kurdish

A Downstream Task: Language Identification



(a) Classification with language codes



(b) Classification with language and script codes