Creating an Electronic Lexicon for the Under-resourced Southern Varieties of Kurdish Language

Zahra Azin[1] and Sina Ahmadi[2]

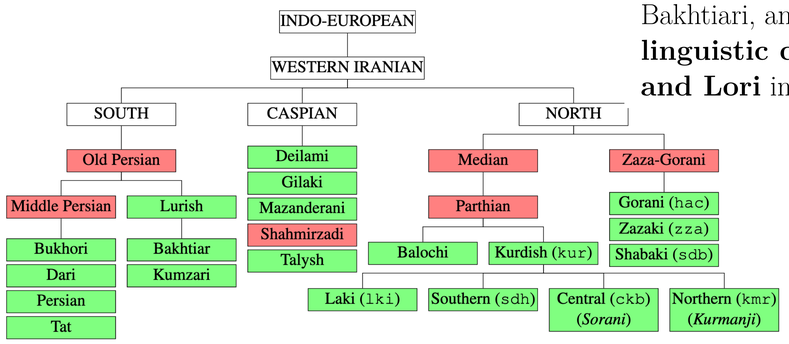
[1] Geomatics and Cartographic Research Center, Carleton University, Canada [2] Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

Objectives

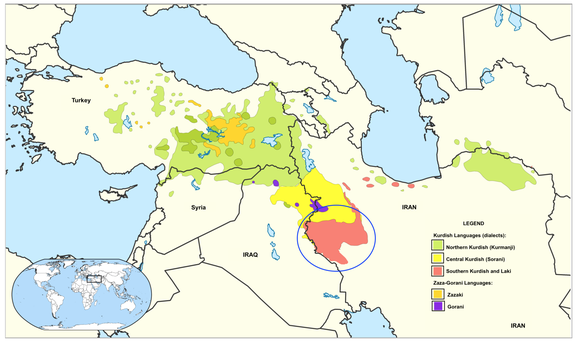
- Provide a description of the **under-represented** Southern variants of Kurdish
- Describe some of the linguistic features of these in comparison to other **Kurdish** variants
- **Create an electronic lexicon for Southern Kurdish**
- Align the lexicon with Sorani, Kurmanji and Gorani lexicons

Introduction

A **machine-readable dictionary** (MRD) not only provides **lexicographic** information in an **electronic** form, but is also a **database** which can be queried and therefore integrated in **natural language processing** tools.
In comparison to **other Kurdish variants** such as central Kurdish (also known as **Sorani**) and northern Kurdish (also known as **Kurmanji**), **southern Kurdish** has received trivial attention making it an under-documented and under-resourced language that is spoken primarily in the Kurdish regions of Iran, particularly Kermanshah and Ilam provinces.



Southern Kurdish



- Southern Kurdish is a variety of Kurdish language consisting of a group of vernaculars spoken by almost **three million** people across an extensive region of **western Iran** including Ilam, a large area of Kermanshah, and some parts of Lorestan and Kurdistan provinces.
- This variety is also spoken in **eastern Iraq** in Khanaqin and Mandali, very close to the borders with Iran.
- Due to the geography of areas where southern Kurdish varieties are spoken, the **population** of southern Kurdish speakers is quite **dispersed**.
- The presence of other languages such as Lori, mainly spoken in Lorestan, Chaharmahal and Bakhtiari, and parts of Ilam has resulted in a **linguistic continuum** between **Kurdish and Lori** in those areas.

Approach

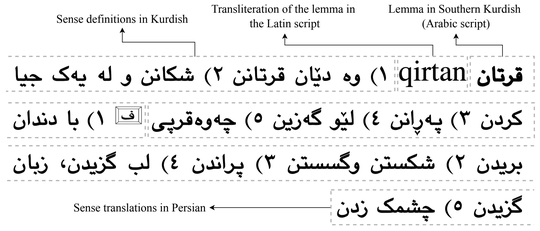
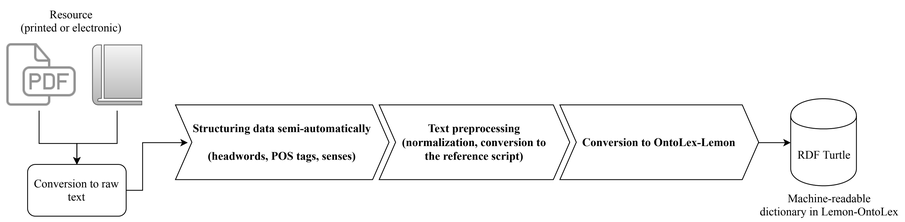


Figure 1:"*qirtan*" 'to cut', an entry in the printed version the dictionary

Our approach is illustrated in 2 as follows:

- We used "**Ferhengî Başur**" (literally meaning "south dictionary"), a southern Kurdish-central Kurdish-Persian dictionary edited with the purpose of codification of southern Kurdish.
- Following [1]'s approach, we use a **semi-automatic technique** to extract entries from the printed dictionary using regular expression.
- To **increase the interoperability and accessibility** of this resource, we provide the electronic dictionary in **Ontolex-Lemon** [2] according to **linguistic linked open data**
- Using pivot-based techniques, we **align senses** across Sorani and Kurmanji dictionary with the current resource

Figure 2:Our approach based on [1]



Limitations

- Despite the attempt to document the general and folkloric vocabulary of southern variants of Kurdish in this resource, there is a **lack of coverage of topics** due to the scarcity of **terminologies** for Kurdish in general, and for these variants in particular.
- Similar to the majority of Kurdish dictionaries, our resource lacks consistent **definition of entries** in such a way that for only a few lemmata sense glosses are provided.
- The same issue can be observed with respect to **idioms, examples and pronunciation**.

References

[1] Sina Ahmadi, Hossein Hassani, and John P McCrae. Towards electronic lexicography for the Kurdish language.
In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*. eLex 2019, 2019.

[2] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano.
The ontolex-lemon model: development and applications.
In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.

Download the corpus

This corpus is publicly available under a CC BY-SA 4.0 license at `https://github.com/sinaahmadi/SKurdishLexicon`.

Carleton UNIVERSITY — Data Science Institute / Institiúid Eolaíochta Sonraí — OÉ Gaillimh NUI Galway