Building a Corpus for the Zaza–Gorani Language Family

Sina Ahmadi

Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

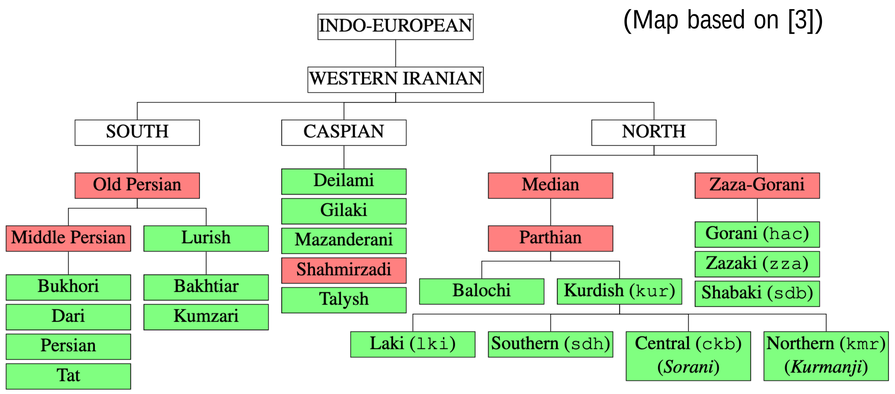
Objectives

- Provide a description of two endangered languages in the **Zaza-Gorani language family**: **Zazaki** and **Gorani**
- Describe some of the linguistic features of these two languages in comparison to **Kurdish**
- **Create a language corpus for Zazaki and Gorani**
- Analyze the curated corpus and compare it with a Kurdish corpus

Introduction

Zazaki and **Gorani** are two of the main and most known languages belonging to the **Zaza-Gorani language family**. Zaza-Gorani languages are not only **less-resourced** but also deemed **endangered languages**. Zazaki, also known as *Dimlî*, is spoken by an estimated number of 2 million speakers. On the other hand, Gorani, also written as *Gurani*, is the language of 300,000 speakers.

In this study, we present **a corpus for Zazaki and Gorani**. **Shabaki**, as the last language in this language family could not be included due to it being extremely under-documented and least known. The corpus is built on the **news articles** from various sources in **several topics** such as science, politics, culture and art. We believe that this resource can pave the way for further developments in the processing of Zaza-Gorani languages in various NLP tasks such as automatic language and dialect identification and spelling and grammatical error correction.



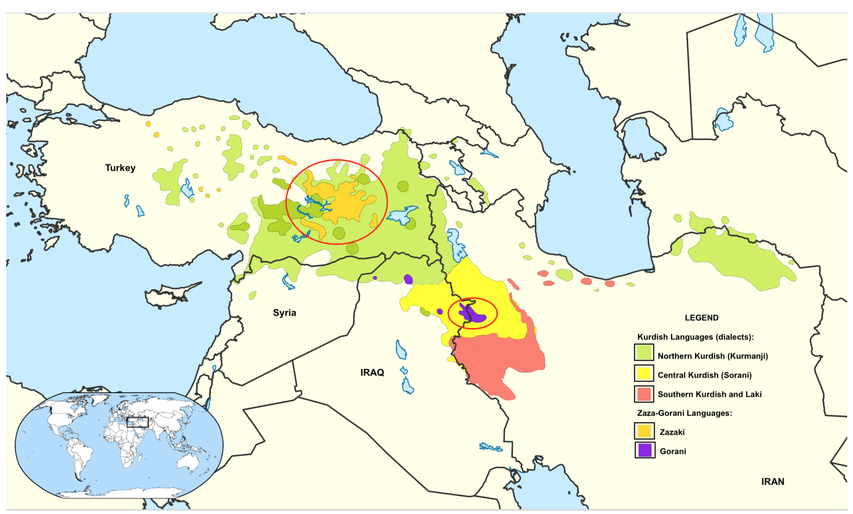
Kurdish vs. Zaza-Gorani

- Despite the common belief that Zazaki and Gorani are two dialects of Kurdish, studies indicate a consensus among linguists that those two are two distinct languages on their own [1]
- Kurdish, Zazaki and Gorani languages are **all in the Northwestern branch of the Iranian languages** within the **Indo-European** language family
- These languages and dialects have **linguistically influenced each other** in various ways, including phonetics, vocabulary and morphology
- **Mutual influence** is particularly observed between Kurmanji Kurdish and Zazaki and Sorani Kurdish and Gorani
- There is generally **a close feeling** among all the three ethnic groups, Kurds, Goranis and Zazas, with respect to the **Kurdish identity and culture** with many centuries living together

Approach

We used the **material published on news websites** in Zazaki and Gorani languages to build the first corpus for those two languages. In comparison to the Sorani and Kurmanji dialects of Kurdish for which many websites are available, there are a very limited number of websites for Zaza-Gorani languages.

Figure 1. Zazaki and Gorani are spoken in the red encircled areas
(Map based on [3])



Our approach is as follows:

1. We **selected websites** based the number of the available articles, availability of metadata in pages' source and the diversity of the covered topics
2. We **extract the content** of the HTML pages and further clean them by removing non-relevant information such as URLs, hashtags, contact details and cited sentences in languages other than our target ones, e.g. Koranic verses in Arabic
3. **Identify the language** or the dialect in which the article is written using a simple classifier using the most frequent and unique words in each language as features, e.g. *ziwan/zan/zon* 'language' for Zazaki and *ziman* for Kurmanji Kurdish
4. **Manually verify** the selected articles

Results

Basic statistics

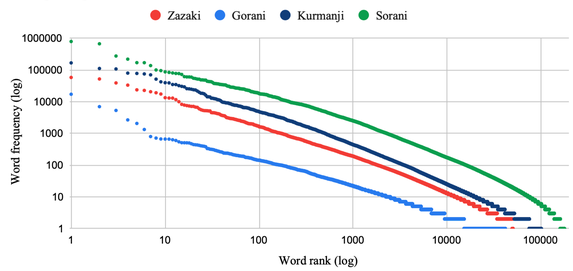
Among the 20 most frequent words in our corpus and the Kurdish corpus of [2], conjunctions 'and', 'that', demonstratives 'this', 'that' and prepositions 'in', 'from', 'until' appear in all the languages.

#	Zazaki	Gorani
articles	4,855	428
word tokens	1,633,770	194,563
word types	102,665	41,454
characters	10,802,266	2,246,425
average word length	4.84	5.50

Zipf's law

Zipf's Law, also known as the rank-size distribution, states that in a reasonably huge data set, including language corpus, there is a correlation between word frequencies and word ranks that follows a power law function. It is beneficial to understand the **significance of words** in a language.

The following Figure illustrates the rank-size distribution in the Zaza-Gorani and Kurdish corpora where a three-segment pattern is observed equally between the languages.



References

[1] David Neil MacKenzie.
The Dialect of Awroman (Hawraman-i Luhon): Grammatical Sketch, Texts, and Vocabulary.
E. Munksgaard., 1966.

[2] Kyumars Sheykh Esmaili and Shahin Salavati.
Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013.

[3] Philippe Rekacewicz.
The languages of Kurdistan (*Les langues du Kurdistan*) [in French].
(Date accessed: 23.06.2020), `https://www.monde-diplomatique.fr/cartes/langueskurdes` edition, 2008.

Download the corpus

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

Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS
Data Science Institute Institiúid Eolaíochta Sonraí
OÉ Gaillimh NUI Galway