In this study, we present a Kurdish corpus and spelling and grammatical error correction. Such as automatic language and dialect identification of Zaza-Gorani languages in various NLP tasks pave the way for further developments in the process—culture and art. We believe that this resource can be used in several topics in Zaza, Gorani, and Shabaki. Zazaki, also known as Dimli, 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 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.

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 Dimli, 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 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 these 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

METHOD

Our approach is as follows:

- We selected websites based on the number of the available articles, availability of metadata in pages’ source and the diversity of the covered topics
- 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
- 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/zen ‘language’ for Zazaki and ziman for Kurmanji Kurdish
- 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.

<table>
<thead>
<tr>
<th></th>
<th>Zazaki</th>
<th>Gorani</th>
</tr>
</thead>
<tbody>
<tr>
<td># of articles</td>
<td>4,855</td>
<td>1,428</td>
</tr>
<tr>
<td># of word tokens</td>
<td>1,633,770</td>
<td>194,563</td>
</tr>
<tr>
<td># of word types</td>
<td>102,665</td>
<td>11,454</td>
</tr>
<tr>
<td># of characters</td>
<td>10,802,266</td>
<td>2,246,429</td>
</tr>
<tr>
<td>Average word length</td>
<td>4.84</td>
<td>5.50</td>
</tr>
</tbody>
</table>

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

REFERENCES


Download the corpus

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