Building a Corpus for the Zaza–Gorani Language Family

Sina Ahmadi
Insight Centre for Data Analytics
National University of Ireland Galway, Ireland
ahmadi.sina@outlook.com

Abstract

Thanks to the growth of local communities and various news websites along with the increasing accessibility of the Web, some of the endangered and less-resourced languages have a chance to revive in the information era. Therefore, the Web is considered a huge resource that can be used to extract language corpora which enable researchers to carry out various studies in linguistics and language technology. The Zaza–Gorani language family is a linguistic subgroup of the Northwestern Iranian languages for which there is no significant corpus available. Motivated to create one, in this paper we present our endeavour to collect a corpus in Zazaki and Gorani languages containing over 1.6M and 194k word tokens, respectively. This corpus is publicly available[1].

1 Introduction

A language corpus refers to a collection of data in a specific language or languages which can be utilized as a sample of the language for linguistic purposes. With a significant number of tokens and sentences, a corpus contains various word forms and therefore, is beneficial in the linguistic analysis of a language, for instance in morphology and syntax. Moreover, the recent advances in applying statistical and neural methods in natural language processing (NLP) have proved the importance of language resources, including large corpora, in improving various tasks, particularly using language models. However, language resources are not evenly available for all languages; given the number of the human languages around the globe, most of the languages are still considered less-resourced, i.e. languages for which there are only general grammar and few electronic texts available.

Zaza-Gorani languages, a subgroup of the Northwestern Iranian languages, are not only less-resourced but are also deemed as endangered languages (Aryadoust et al., 2008; Arslan, 2016; Arslan, 2017). Zazaki and Gorani are two of the main and most known languages belonging to this family. Zazaki, also known as Dimlî, is spoken by an estimated number of 2 million speakers in various regions in Turkey (Paul, 1998; Extra and Gorter, 2001, p 418). On the other hand, Gorani[2], also written as Gurani, is the language of ∼300,000 speakers in the parts of the Iranian of Iraqi Kurdistan (Paul, 2007). Historically, Gorani was the high literary language within the Sorani Kurdish speaking regions in such a way that it played a great role in the formation of modern Sorani Kurdish and literature (Edmonds, 2013).

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 (Sultan, 2011). The corpus is built on the news articles from various sources in several topics such as science, politics, culture and art, and contains 1,633,770 tokens in Zazaki and 194,563 tokens in Gorani. 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 (Hassani and Medjedovic, 2016) and spelling and grammatical error correction. Given the similarities between these languages and Kurdish, this corpus can also be beneficial to take use of available resources and tools of Kurdish, such as named-entity recognition (Littell et al., 2016).

[1] https://github.com/sinaahmadi/ZazaGoraniCorpus

[2] Not to be confused with the Gorani people in the Balkans

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 70–78
Barcelona, Spain (Online), December 13, 2020

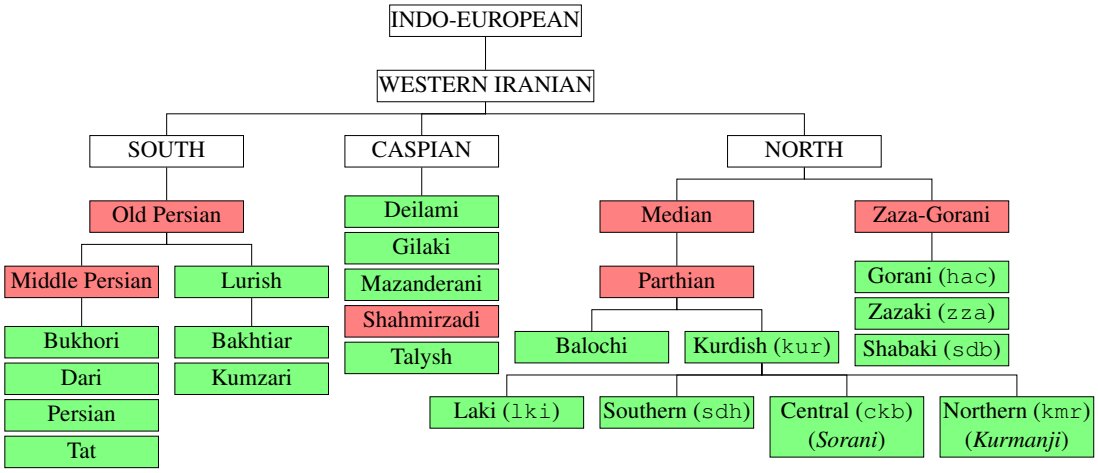

Figure 1: The place of the Kurdish and Zaza-Gorani languages in the Western Iranian language family. Dead and alive languages are respectively specified in red and green and, ISO 639-3 language codes are provided in parentheses

2 Kurdish vs. Zaza-Gorani

The question of dialects and languages in the Kurdish inhabitant regions has been a matter of discussion both in academia and among people. Although some have suggested that Northern and Central Kurdish, also widely known by their endonyms, respectively, Kurmanji[3] and Sorani, are two distinct languages given the structural differences between them, they are more uncontroversially accepted as two dialects of Kurdish (Haig and Matras, 2002). Despite the common belief that Zazaki and Gorani are two dialects of Kurdish (Hassanpour, 1998), studies indicate a consensus among linguists that those two are two distinct languages on their own (MacKenzie, 1966; Minorsky, 1943). That said, 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 (Yavuz, 1998; Schmidinger, 2013). However, the formation of new identities among Gorani and Zazas to distinguish themselves from Kurds has been also studied more recently (Kane, 2003; Hassanpour et al., 2012; Sheyholislami, 2017).

Kurdish, Zazaki and Gorani languages are all in the Northwestern branch of the Iranian languages within the Indo-European language family. Regarding the Kurdish dialects, Kurmanji is spoken in all the regions of Kurdistan in Iraq, Iran, Syria and Turkey, with a predominant population in the two latter regions. Sorani and Southern Kurdish are both spoken by the Kurdish populations of Iraq and Iran. While the majority of the Southern Kurdish speakers are located in the southern parts of the Iranian Kurdistan, particularly in Kermanshah and Ilam provinces, Sorani is the most widely spoken dialect in both Kurdish regions of Iran and Iraq. On the other hand, Zazaki is spoken only in the Kurdish region of Turkey, mostly in Tunceli, Bingöl, Urfa, Elazığ and north of Diyarbakır in various dialects of *Dimli*, *Kirdki*, *Kirmanjki* and *Kirmanji* (White, 1995). Finally, Gorani is mostly spoken in the Iranian Kurdistan and smaller parts of the Iraqi Kurdistan in various dialects, including *Bajelani*, *Sarli*, *Gawrajuyi* and *Hawrami* (also known as *Awromani* or *Awramani*), among which the latter is the most popular and known dialect (MacKenzie, 2002; Mahmoudveysi et al., 2012).

These languages and dialects have linguistically influenced each other in various ways, including phonetics, vocabulary and morphology. More specifically, the mutual influence is observed between Kurmanji Kurdish and Zazaki (Haig and Matras, 2002; Karacan, 2020), and also, between Sorani Kurdish and Gorani (Leezenberg, 1993; Chaman Ara and Amiri, 2018). In order to better compare these languages, we discuss some of the major common features of Kurdish, Zazaki and Gorani from a comparative perspective. Additionally, we present the alphabets used for writing. We regret that Southern Kurdish, Laki and Shabaki could not be included due to scarcity of resources and grammar books (Fattah, 2000).

[3] *Badini* is also used to refer to the Kurmanji spoken in the Iraqi Kurdistan

2.1 Phonetics and Alphabets

IPA		b	t͡ʃ	d͡ʒ	d	f	g	h	ʒ	k	l	ɬ	m	n	p	q	ɾ	r	ɹˠ	s	sˤ	ʃ	tˤ	t	v	w	x	j	z	ʕ	ħ	ɣ	ʔ
Zazaki		b	ç	c	d	f	g	h	j	k	l	‘l	m	n	p	q	r	rr		s	‘s	ş	‘t	t	v	w	x	y	z	‘	‘h	ğ	’
Kurdish	Latin	b	ç	c	d	f	g	h	j	k	l	ł/ll	m	n	p	q	r	ř/rr		s		ş		t	v	w	x	y	z	ë/‘	ẗ/‘h	ẍ	’
	Arabic	ب	چ	ج	د	ف	گ	ه	ژ	ک	ل	ڵ	م	ن	پ	ق	ر	ڕ		س		ش		ت	ڤ	و	خ	ی	ز	ع	ح	غ	ئ
Gorani		ب	چ	ج	د	ف	گ	ه	ژ	ک	ل	ڵ	م	ن	پ	ق	ر	ڕ	ڎ	س		ش		ت	ڤ/ۆ	و	خ	ی	ز	ع	ح	غ	ئ

(a) Consonants

IPA		aː	æ	e	eː	ɪ	iː	oː	uː	ʊ	ɔ
Zazaki		a	e		ê	i/ı	î/i	o	û	u	
Kurdish	Latin	a	e		ê	i	î	o	û	u	
	Arabic	ا	ه		ێ		ی	ۆ	وو	و	
Gorani		ا	ه	ێ/ئ	ێ		ی	ۆ	وو	و	ۆ/وَ

(b) Vowels

Table 1: A comparison of the Kurdish, Zazaki and Gorani alphabets

Historically, many scripts have been used for writing Kurdish, Zazaki and Gorani, namely, Cyrillic, Armenian, Arabic and Latin among which the latter two are still widely used (Ahmadi et al., 2019). Although the standardization of alphabets and orthographies has been widely discussed among scholars, to date it is considered an unsolved problem (Tavadze, 2019). The choice of scripts seems to be influenced by the administration where the language is spoken. For instance, the Kurmanji speakers of Iraqi Kurdistan still use the Arabic-based script while the majority of the Kurmanji speakers, who are in Turkey and Syria, use the Latin-based alphabet. Similarly, Zazaki uses the Latin-based alphabet (Werner, 2012). Regarding the Gorani language, the Arabic-based alphabet of Sorani Kurdish is used along with new graphemes for phonemes unique to Gorani. It should be noted that all these alphabets are used with phonemic orthographies, i.e. each phoneme is associated with a grapheme.

Table 1 presents the alphabets used in each language with their corresponding phonemes in the International Phonetic Alphabet (IPA). One can say that all the common vowels and consonants are identical in all languages, with a few subtle but audible differences (Haig, 2018; Odden, 2005, p 131). Regarding the number of consonants and vowels, Zaza-Gorani languages outnumber Kurdish. This is particularly because of the presence of pharyngalized consonants in Zazaki ([sˤ] and [tˤ]), the more diverse vowels ([e] and [ɔ]) and the approximant interdental plosive [ɹˠ] in Gorani (Naghshbandi, 2020; Karacan, 2020). The latter is unique to Gorani language and is also known as the Zagros [d] (Haig and Khan, 2018, p 386). Although [ɪ] exists in all languages, there is no grapheme for it in the Arabic-based script. It is worth noting that as we did not find any formal description of the Gorani alphabet, the alphabet described in the Wişename dictionary (Habiballah (Bedar), 2010) and (Kord Zafaranlu Kambuziya and Sajjadi, 2013) are used as reference. Moreover, in cases where more than a variation is found for a grapheme, they are specified with ”/”.

2.2 Grammar

All the languages have a system of tense-aspect-modality along with person marking (Haig and Matras, 2002) with subject-object-verb (SOV) positioning. In addition, similar to some of the other Western Iranian languages, there is a common feature in morphosyntactic alignment of all those languages and that is ergativity (Scheucher, 2019). Ergativity refers to the morphosyntactic property where the subject of a transitive verb is marked by an agentive, i.e. oblique case, which is distinct from the nominative case. In all Kurdish, Zazaki and Gorani languages the ergative-absolutive alignment appears only in the past tenses of transitive verbs.

Regarding passive voice, Kurmanji expresses passive forms periphrastically, using the auxiliary verb *hatin* ‘to come’ while Sorani Kurdish, Zazaki and Gorani apply morphological changes to the verb stem.

Language	Passive	Gender	Case	Alignment
Kurmanji Kurdish	periphrastic with *hatin* (to come) (Thackston, 2006a)	feminine, masculine (Thackston, 2006a)	nominative, oblique, Izafa, vocative (Thackston, 2006a)	nominative–accusative, only in past transitive ergative–absolutive (Matras, 1997)
Sorani Kurdish	morphological (Thackston, 2006b)	no gender (Thackston, 2006b)	nominative, oblique, locative, vocative (McCarus, 2007)	nominative–accusative, only in past transitive ergative–absolutive (Karimi, 2014)
Gorani	morphological (Aryadoust et al., 2008)	feminine, masculine (Sadjadi, 2019)	nominative, oblique, Izafa (MacKenzie, 1966)	nominative–accusative, only in past transitive ergative–absolutive (Rasekh Mahand and Naghshbandi, 2014)
Zazaki	morphological (Selcan, 1998; Todd, 2003)	feminine, masculine (Todd, 2003)	nominative, oblique, oblique of kinship terms, locative, vocative, double Izafe (Todd, 2003; Larson and Yamakido, 2006)	nominative–accusative, only in past transitive ergative–absolutive (Todd, 2003)

Table 2: A comparison of the Sorani and Kurmanji dialects of Kurdish with Zazaki and Gorani languages

In regards to the grammatical cases, all languages have four major noun cases, namely nominative, oblique, locative and vocative. Additionally, like some other languages in the Iranian language family, Kurdish, Zazaki and Gorani have a linker morpheme called *Izafa* (also known as *Ezafe*) which appears between a head and its dependents in a noun phrase and is usually recognized as a specific grammatical case known as construct. Izafa is widely used for creating attributive adjectives and possessive constructions. In the latter, it can be translated as 'of' in English. Similar to Kurmanji Kurdish, Izafa in Gorani and Zazaki has several realizations. In Gorani, [-i], [-æ], [-e], [-ʊ] are used based on the modifier and the presence of those elements which require a grammatical agreement, such as definiteness and number (Holmberg and Odden, 2008). Likewise, Zazaki has various morphological forms for Izafa depending on the relationship between the noun and its dependents, namely *-o*, *-a*, *-ê* (Werner, 2012; Ludovico et al., 2015, p 322). Moreover, Zazaki has a special type of Izafa, called doubled-Izafa, which happens when an Izafa construction is used within another Izafa phrase (Larson and Yamakido, 2006). For this purpose, morphemes *-da* and *-de* are used depending on gender and number. In comparison to the aforementioned languages, Sorani dialect of Kurdish has a simpler Izafa construction where only *-î* and its allomorph *-e* are used (Salehi, 2018, p 53).

Language		Noun						Verb				Adjective	
		DEF			INDEF			INF	PROG	SBJV	NEG	COMP	SUPL
		M	F	PL	M	F	PL						
Kurdish	Sorani	*-eke*		*-ekan*	*-êk*		*-an, -gel*	-in	*de-, e-*	*bi-*	*ne-, na-, me-*	*-tir*	*-tirîn*
	Kurmanji	-	-	-	*-ek*	*-ek*	*-in*	*-in*	*di-*	*bi-*	*ne-, na-*	*-tir*	*-tirîn*
Gorani		-[(æ)kæ]	-[(æ)ke]	-[(æ)ke], -[(æ)kan]	-[ew], -[ewæ]	-[ewæ], -[evæ]	-[e], -[a:], -[a:n]	-[æj]	[mæ]-	[bɪ]-	[næ]-	-[tær]	-[tæri:n]
Zazaki		-	-	-	*-ê*	*-ê*	*-ê*	*-iş, -ene*	-	*bi-*	*nê-, ni-, me-, çi-*	*-êrî*	-

Table 3: Some inflectional morphemes in Zazaki, Gorani and Kurdish. Nominal morphemes are provided in nominative and morphophonological alternations are excluded. Sorani Kurdish does not have gender.

Tables 2 and 3 provide some of the major morphological and syntactic characteristics of Kurdish (Sorani and Kurmanji dialects), Zazaki and Gorani languages. Regarding nouns, definiteness is not specified with markers in Kurmanji Kurdish and Zazaki while Gorani and Sorani Kurdish use markers. The fully-marked article system is a distinct feature of Gorani and Sorani (Jugel, 2014). As Sorani does not have any grammatical gender, it has a simpler combination of noun markers in comparison to Kurmanji, Zazaki and Gorani. Regarding verbs and adjectives, Sorani and Kurmanji use identical prefixes, except in a few cases. However, such a similarity is less observed in Zazaki and Gorani. In Zazaki, *da, ra* and *-êrî* are used with adjectives among which only the latter appears as a suffix. Moreover, superlative adjectives are implicit without any specific morpheme (Todd, 2003). Zazaki also has a different oblique case for kinship terms, such as *cenî* 'wife'. Affixes in Kurmanji and Zazaki are compared in more detail in Malmîsanîj and Mosa (2017).

It is worth mentioning that the description of the grammar in this section may vary depending on the dialects. For instance, in some dialects of Sorani Kurdish, namely Ardalani and Babani, oblique case does not exist while in some other dialects, such as Mukryani, nouns are specified with oblique markers.

3 Methodology

Similar to the methodology proposed by Esmaili and Salavati (2013) to create the first standard test collection for the Kurdish language, 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. Among the available websites, we selected Zazaki.net[4] for Zazaki and Firat News Agency[5] for both languages. Our selection criteria were the number of the available articles, availability of metadata in pages' source and the diversity of the covered topics. Regarding the topics, the first website focuses on cultural issues and is composed of analytical articles in humanities and provides interviews in such fields. On the other hand, the latter addresses a wider range of news in topics such as women, politics, world, Kurdistan, science, culture and art.

After crawling the websites, 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. In most cases, the page's address schema enabled us to identify the language. However, in none of the websites specific tags were found to explicitly identify the language or the dialect in which the article is written. As such, we use a simple classifier to exclude English, Turkish and Kurdish articles from the Zazaki and Gorani ones. For this purpose, we manually selecte a list of the most frequent and unique words in each language as features. For instance, *ziwan/zan/zon* 'language', *kerd* 'did' and *zî* 'too, also' are unique to Zazaki while *ziman*, *kir* and *jî* are respectively used in Kurmanji Kurdish. In the case of Sorani Kurdish and Gorani, in addition to the frequent words, we use unique characters as features as well. This step is followed by a manual verification of the documents.

#	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

Table 4: Basic statistics of the Zaza-Gorani corpus

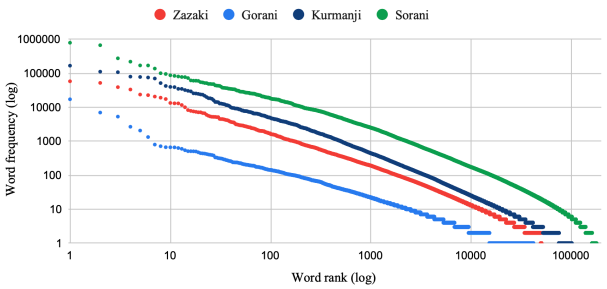

Figure 2: Zipfian distribution of the Zaza-Gorani and Kurdish corpora

[4]http://www.zazaki.net/
[5]https://anfsorani.com, https://anfkirmancki.com

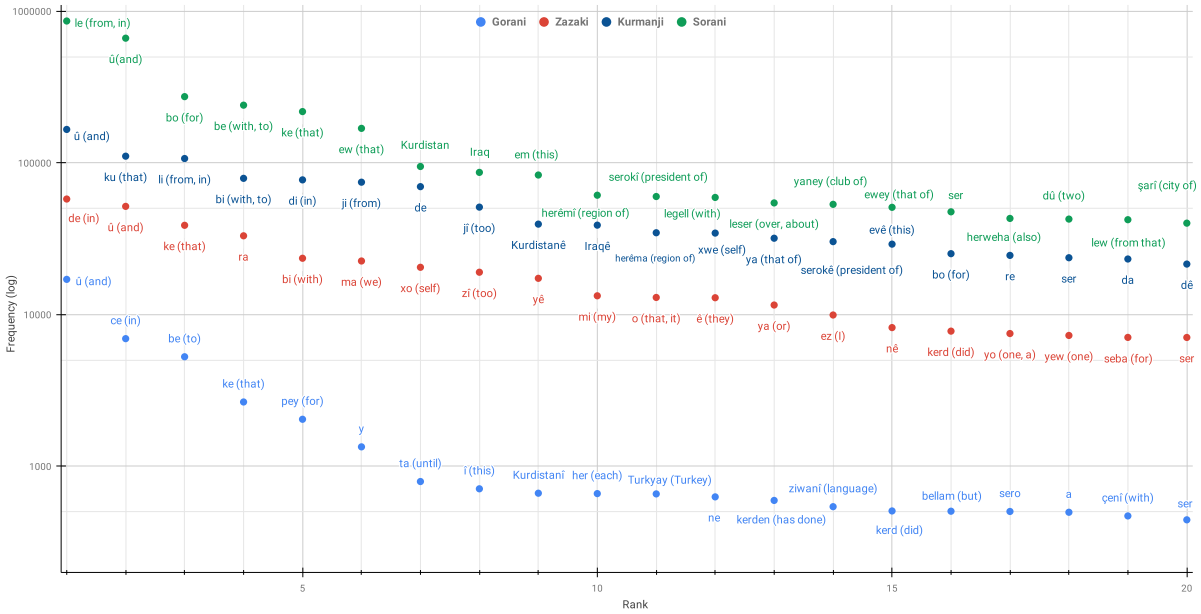

Figure 3: The 20 most frequent words in the Zaza-Gorani and Kurdish corpora

In order to keep the document-level information, we collected the cleaned documents in two directories based on the language. We provide further metadata such as topic, original source of the article and date of publication in a separate JSON file that can be associated to each document. Regarding the date of publication, the Zazaki articles dated between 2009 and 2020 while the Gorani ones are more recent (2018-2020).

4 Results

We carry out an intrinsic evaluation based on the statistics and the most frequent words in our corpus with comparison to Esmaili and Salavati's PEWAN corpus (2013). This corpus contains 18M tokens in Sorani and 4M tokens in Kurmanji Kurdish.

4.1 Frequency

Table 4 provides the basic statistics of the corpus in Zazaki and Gorani. 6.28% of the Zazaki words, i.e. 102,665 words, and 21.3% of the Gorani words, i.e. 41,454 words, are distinct in the corpus. Such ratios of word types with respect to word tokens demonstrate the richness of the corpora and the diversity of the words if the corpora are of the same size. Regarding the average word length, Esmaili and Salavati (2013) report 4.8 and 5.6 for Kurmanji and Sorani, respectively, which are almost identical to Zazaki and Gorani averages. This can be explained by the usage of the Arabic-based alphabets in both Gorani and Sorani where words, particularly one-letter morphemes such as present copula, are often concatenated with other words.

Figure 3 provides the 20 most frequent words in our corpus in comparison with the Kurdish corpus. Conjunctions ('and', 'that'), demonstratives ('this', 'that') and prepositions ('in', 'from', 'until') appear in all the languages. Postpositions *de*, *re* and *da*, which appear mostly as a part of circumpositions, are among the most frequent words in the Kurmanji and Zazaki corpora as well. The same cannot be observed in Gorani and Sorani as postpositions are mostly attached to the previous word in the Arabic-based script. For the same reason, the counterparts of the emphatic clitics *jî* in Kurmanji and *zî* in Zazaki, namely *îş* in Sorani and *îç* in Gorani, do not appear among the most frequent words. These clitics are usually translated as 'also', 'too' and 'even' in English. In Zazaki and Gorani, the past form of the verb 'to do' (*kerd*) is of high frequency. Finally, the occurrence of non-function words such as *Kurdistan* and *Iraq*, indicates the tendency of the news content towards Kurdish-related issues.

4.2 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, both in logarithmic scales, that follows a power law function. Using data of 50 languages, Yu et al. (2018) demonstrate that the patterns of such a correlation, i.e. Zipf's law, in all their studied languages share a three-segment structural pattern: upper segment where the most frequent words appear, middle segment where the curve gets smoother and finally, the lower segment where the rest of the words with low frequency appear. Figure 2 illustrates the rank-size distribution in the Zaza-Gorani and Kurdish corpora where a similar three-segment pattern is observed. The first 10 most frequent words in all the languages appear in the upper segment. While Zazaki and Kurmanji closely follow the same pattern, there is a sharp drop between the upper and the middle segments in Sorani and Gorani. Zipf's law is beneficial to understand the significance of words in a language with various applications in information retrieval and computational psycholinguistics (Powers, 1998).

5 Conclusion and Future Work

In this paper, we presented our efforts in creating a language corpus for two endangered languages of the Zaza-Gorani language family. Zazaki, Gorani and Shabaki are the three languages belonging to this language family and are popularly believed to belong to Kurdish. We briefly discuss how these languages are different from Kurdish, Sorani and Kurmanji dialects, in terms of phonetics, morphology and syntax. We also report our efforts in collecting documents in various topics from news websites and create the first corpus for Zazaki and Gorani. We believe that this corpus can pave the way for further developments in linguistics and computer science, particularly in information retrieval and NLP where language modeling is beneficial to various applications such as grammatical and spell checking.

As a future work, we suggest a better documentation of the Kurdish and Zaza-Gorani languages, particularly Shabaki, Southern Kurdish and Laki, by promoting the usage of those languages within local communities, websites and social media platforms. In the same vein, we invite researchers and native speakers to pay further attention to these languages, both in linguistics and NLP, by providing more analytical grammars, particularly in Gorani and Southern Kurdish, and developing basic language processing tools, such as tokenization, stemming and lemmatization, and resources, such as WordNet (Aliabadi et al., 2014) and parallel corpora.

Acknowledgements

The author would like to thank the constructive comments of Dr. Ilyas Arslan and Mesut Keskin regarding Zazaki and the invaluable insights of Dr. Parvin Mahmoudveysi regarding Gorani. Likewise, the comments of the anynomous reviewers are very much appreciated.

References

Sina Ahmadi, Hossein Hassani, and John P. McCrae. 2019. Towards electronic lexicography for the Kurdish language. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*, pages 881–906, Sintra, Portugal, 10.

Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards building Kurdnet, the Kurdish wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 1–6.

Ilyas Arslan. 2016. *Verbfunktionalität und Ergativität in der Zaza-Sprache*. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf.

Zeynep Arslan. 2017. Zazaki–yesterday, today and tomorrow. *Survival and standardization of a threatened language. Dieter Halwachs: Grazer Plurlingualismus Studien (GPS 04). Graz: GLM*.

Seyed Vahid Aryadoust, Narges Marandi, and Masoud Aryadoust. 2008. A contrastive analysis of modern Hawrami Kurdish and Persian verbs and tenses. *National Institute of Education, Singapore*.

Behrooz Chaman Ara and Cyrus Amiri. 2018. Gurani: practical language or Kurdish literary idiom? *British Journal of Middle Eastern Studies*, 45(4):627–643.

Alexander Johannes Edmonds. 2013. The Dialects of Kurdish. *Ruprecht-Karls-Universität Heidelberg*.

Kyumars Sheykh Esmaili and Shahin Salavati. 2013. Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 300–305.

Guus Extra and Durk Gorter. 2001. *The other languages of Europe: Demographic, sociolinguistic, and educational perspectives*, volume 118. Multilingual Matters.

Ismaïl Kamandâr Fattah. 2000. *Les dialectes kurdes méridionaux: étude linguistique et dialectologique*. Acta Iranica : Encyclopédie permanente des études iraniennes. Peeters.

Jamal Habiballah (Bedar). 2010. *Wişename (Hawrami-Sorani Kurdish dictionary)*. Aras Publishing and Printing House.

Geoffrey Haig and Geoffrey Khan. 2018. *The Languages and Linguistics of Western Asia: An Areal Perspective*. The World of Linguistics. De Gruyter.

Geoffrey Haig and Yaron Matras. 2002. Kurdish linguistics: a brief overview. *STUF-Language Typology and Universals*, 55(1):3–14.

Geoffrey Haig. 2018. The Iranian languages of Northern Iraq. *The Languages and Linguistics of Western Asia: An Areal Perspective*, 6:267.

Hossein Hassani and Dzejla Medjedovic. 2016. Automatic Kurdish dialects identification. *Computer Science & Information Technology*, 6(2):61–78.

Amir Hassanpour, Jaffer Sheyholislami, and Tove Skutnabb-Kangas. 2012. Introduction. Kurdish: Linguicide, resistance and hope. *De Gruyter Mouton*.

Amir Hassanpour. 1998. The identity of Hewrami speakers: Reflections on the theory and ideology of comparative philology. *Anthology of Gorani Kurdish poetry*, 35:49.

Anders Holmberg and David Odden. 2008. The noun phrase in Hawrami. *Aspects of Iranian linguistics*, 12952.

Thomas Jugel. 2014. On the linguistic history of Kurdish. *Kurdish Studies*, 2(2):123–142.

Andrew Kane. 2003. The Reality of Intra-Kurdish Rivalry Undermines the Notion of Pan-Kurdish Nationalism. *Geopolitics*, 8(1):48.

Hasan Karacan. 2020. Kurmanji and Zazaki Dialects: Comparative Study on their Phonetics. *International Journal of Kurdish Studies*, 6(1):35–51.

Yadgar Karimi. 2014. On the syntax of ergativity in Kurdish. *Poznan Studies in Contemporary Linguistics*, 50(3):231–271.

Aliye Kord Zafaranlu Kambuziya and Seyed Mehdi Sajjadi. 2013. Syllable Structure in Hawrami (Takht Dialect) [In Persian]. *The journal of Western Iranian Languages and Dialects*, 1(2):57–78.

Richard K Larson and Hiroko Yamakido. 2006. Zazaki "double Ezafe" as double case-marking. In *annual meeting of the Linguistic Society of America, Albuquerque, NM*.

Michiel Leezenberg. 1993. *Gorani influence on Central Kurdish: Substratum or prestige borrowing?* Universiteit van Amsterdam. Instituut voor Taal, Logica en Informatie (ITLI).

Patrick Littell, David R Mortensen, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Bridge-language capitalization inference in western Iranian: Sorani, Kurmanji, Zazaki, and Tajik. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3318–3324.

Franco Ludovico, M Rita Manzini, and Leonardo M Savoia. 2015. Linkers and agreement. *The Linguistic Review*, 32(2):277–332.

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

David Neil MacKenzie. 2002. Gurāni. *Encyclopaedia Iranica*, XI:401–403.

Parvin Mahmoudveysi, Denise Bailey, Ludwig Paul, and Geoffrey Haig. 2012. The Gorani language of Gawrajū, a village of West Iran. *Wiesbaden: Reichert*.

Mehmet T Malmîsanîj and Abdulwahab X Mosa. 2017. Prefixes, suffixes and infixes in Kurmanji (Zazaki) (Comparative descriptive study) [In Kurdish]. *Humanities Journal of University of Zakho*, 5(2):508–526.

Yaron Matras. 1997. Clause combining, ergativity, and coreferent deletion in Kurmanji. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 21(3):613–653.

Ernst M McCarus. 2007. Kurdish morphology. *Morphologies of Asia and Africa*, 2:1021–1049.

Vladmir Minorsky. 1943. The Gūrān. *Bulletin of the School of Oriental and African Studies*, 11(1):75–103.

Shahram Naghshbandi. 2020. The Approximantization of Alveolar Plosives in Hawrami (Paveh Variety) [In Persian]. *Language Related Research, Tarbiat Modares University Press*, 11(1).

David Odden. 2005. *Introducing phonology*. Cambridge university press.

Ludwig Paul. 1998. The position of Zazaki among West Iranian languages. *Old and Middle Iranian Studies*, pages 163–176.

Ludwig Paul. 2007. Zur Lage der Gōrānī-Dialekte im Iran und ihrer Erforschung. *Iranian Languages and Texts from Iran and Turan–Ronald E. Emmerick Memorial*, pages 285–296.

David MW Powers. 1998. Applications and explanations of Zipf's law. In *New methods in language processing and computational natural language learning*.

Mohammad Rasekh Mahand and Zaniar Naghshbandi. 2014. The effect of discourse factors on case system in Hawrami. *Language Related Research*, 4(4):87–109.

Mahdi Sadjadi. 2019. Grammatical Gender in Arabic and Hawrami. *International Journal of Language and Linguistics*, 6(2).

Ali Salehi. 2018. Constraints on Izāfa in Sorani Kurdish. In *Theses and Dissertations–Linguistics. 31.*

Bernhard Scheucher. 2019. Ergativity in New West Iranian. *Essays on Typology of Iranian Languages*, 328:5.

Thomas Schmidinger. 2013. The Kurdish diaspora in Austria and its imagined Kurdistan. *The Kurdish Spring: Geopolitical Changes and the Kurds (308-338). Costa Mesa, California: Mazda Publishers*.

Zılfi Selcan. 1998. *Grammatik der Zaza-Sprache: Nord-Dialekt (Dersim-Dialekt)*. Wiss.-und-Technik-Verlag.

Jaffer Sheyholislami. 2017. Language Status and Party Politics in Kurdistan-Iraq: The case of Badini and Hawrami Varieties. *Zazaki–yesterday, today and tomorrow. Survival and standardization of a threatened language. Dieter Halwachs: Grazer Plurlingualismus Studien (GPS 04). Graz: GLM*.

Abbas Sultan. 2011. An account of light verb constructions in Shabaki. *Acta Linguistica Journal*, 5(2).

Givi Tavadze. 2019. Spreading of the Kurdish Language Dialects and Writing Systems Used in the Middle East. *Bull. Georg. Natl. Acad. Sci*, 13(1).

Wheeler M Thackston. 2006a. *Kurmanji Kurdish:A Reference Grammar with Selected Readings*. Harvard University.

Wheeler M Thackston. 2006b. *Sorani Kurdish: A Reference Grammar with Selected Readings*. Harvard University.

Terry Lynn Todd. 2003. *A grammar of Dimili: also known as Zaza*. Ph.D. thesis, UMI Ann Arbor.

Brigitte Werner. 2012. Morphological Sketch of Southern Zazaki. *SIL International*.

Paul J White. 1995. Ethnic Differentiation among the Kurds: Kurmanci, Kızılbaş and Zaza. *Journal of Arabic, Islamic & Middle Eastern Studies*, 1:67–90.

M Hakan Yavuz. 1998. A preamble to the Kurdish question: The politics of Kurdish identity. *Journal of Muslim Minority Affairs*, 18(1):9–18.

Shuiyuan Yu, Chunshan Xu, and Haitao Liu. 2018. Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. *arXiv preprint arXiv:1807.01855*.