Sorani Kurdish versus Kurmanji Kurdish: An Empirical Comparison

Kyumars Sheykh Esmaili
Nanyang Technological University
N4-B2a-02
Singapore
kyumarss@ntu.edu.sg

Shahin Salavati
University of Kurdistan
Sanandaj
Iran
shahin.salavati@ieee.org

Abstract

Resource scarcity along with diversity–both in dialect and script–are the two primary challenges in Kurdish language processing. In this paper we aim at addressing these two problems by (i) building a text corpus for Sorani and Kurmanji, the two main dialects of Kurdish, and (ii) highlighting some of the orthographic, phonological, and morphological differences between these two dialects from statistical and rule-based perspectives.

1 Introduction

Despite having 20 to 30 millions of native speakers (Haig and Matras, 2002; Hassanpour et al., 2012; Thackston, 2006b; Thackston, 2006a), Kurdish is among the less-resourced languages for which the only linguistic resource available on the Web is raw text (Walther and Sagot, 2010).

Apart from the resource-scarcity problem, its diversity –in both dialect and writing systems– is another primary challenge in Kurdish language processing (Gautier, 1998; Gautier, 1996; Esmaili, 2012). In fact, Kurdish is considered a *bi-standard* language (Gautier, 1998; Hassanpour et al., 2012): the Sorani dialect written in an Arabic-based alphabet and the Kurmanji dialect written in a Latin-based alphabet. The features distinguishing these two dialects are phonological, lexical, and morphological.

In this paper we report on the first outcomes of a project[1] at *University of Kurdistan* (*UoK*) that aims at addressing these two challenges of the Kurdish language processing. More specifically, in this paper:

1. we report on the construction of the first relatively-large and publicly-available text corpus for the Kurdish language,
2. we present some insights into the orthographic, phonological, and morphological differences between Sorani Kurdish and Kurmanji Kurdish.

The rest of this paper is organized as follows. In Section 2, we first briefly introduce the Kurdish language and its two main dialects then underline their differences from a rule-based (a.k.a. corpus-independent) perspective. Next, after presenting the Pewan text corpus in Section 3, we use it to conduct a statistical comparison of the two dialects in Section 4. The paper is concluded in Section 5.

2 The Kurdish Language and Dialects

Kurdish belongs to the Indo-Iranian family of Indo-European languages. Its closest better-known relative is Persian. Kurdish is spoken in Kurdistan, a large geographical area spanning the intersections of Turkey, Iran, Iraq, and Syria. It is one of the two official languages of Iraq and has a regional status in Iran.

Kurdish is a dialect-rich language, sometimes referred to as a dialect continuum (Matras and Akin, 2012; Shahsavari, 2010). In this paper, however, we focus on Sorani and Kurmanji which are the two closely-related and widely-spoken dialects of the Kurdish language. Together, they account for more than 75% of native Kurdish speakers (Walther and Sagot, 2010).

As summarized below, these two dialects differ not only in some linguistics aspects, but also in their writing systems.

2.1 Morphological Differences

The important morphological differences are (MacKenzie, 1961; Haig and Matras, 2002; Samvelian, 2007):

1. Kurmanji is more conservative in retaining both gender (feminine:masculine) and case opposition (absolute:oblique) for nouns and

[1] http://eng.uok.ac.ir/esmaili/research/klpp/en/main.htm

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 300–305,
Sofia, Bulgaria, August 4-9 2013.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Arabic-based	ا	ب	ج	چ	د	ێ	ف	گ	ژ	ک	ل	م	ن	ۆ	پ	ق	ر	س	ش	ت	وو	ڤ	خ	ز
Latin-based	A	B	C	Ç	D	Ê	F	G	J	K	L	M	N	O	P	Q	R	S	Ş	T	Û	V	X	Z

(a) One-to-One Mappings

	25	26	27	28
Arabic-based	ئ / ی	و	ی	ه
Latin-based	I	U / W	Y / Î	E / H

(b) One-to-Two Mappings

	29	30	31	32	33
Arabic-based	ڕ	ڵ	ع	غ	ح
Latin-based	(RR)	-	(E)	(X)	(H)

(c) One-to-Zero Mappings

Figure 1: The Two Standard Kurdish Alphabets

pronouns[2]. Sorani has largely abandoned this system and uses the pronominal suffixes to take over the functions of the cases,

2. in the past-tense transitive verbs, Kurmanji has the full ergative alignment[3] but Sorani, having lost the oblique pronouns, resorts to pronominal enclitics,

3. in Sorani, passive and causative are created via verb morphology, in Kurmanji they can also be formed with the helper verbs `hatin` ("to come") and `dan` ("to give") respectively, and

4. the definite marker `-aka` appears only in Sorani.

2.2 Scriptural Differences

Due to geopolitical reasons (Matras and Reershemius, 1991), each of the two dialects has been using its own writing system: while Sorani uses an Arabic-based alphabet, Kurmanji is written in a Latin-based one.

Figure 1 shows the two standard alphabets and the mappings between them which we have categorized into three classes:

- one-to-one mappings (Figure 1a), which cover a large subset of the characters,

- one-to-two mappings (Figure 1b); they reflect the inherent ambiguities between the two writing systems (Barkhoda et al., 2009). While transliterating between these two alphabets, the contextual information can provide hints in choosing the right counterpart.

- one-to-zero mappings (Figure 1c); they can be further split into two distinct subcategories: (i) the strong L and strong R characters ({ڵ} and {ڕ}) are used only in Sorani Kurdish[4] and demonstrate some of the inherent phonological differences between Sorani and Kurmanji, and (ii) the remaining three characters are primarily used in the Arabic loanwords in Sorani (in Kurmanji they are approximated with other characters).

It should be noted that both of these writing systems are phonetic (Gautier, 1998); that is, vowels are explicitly represented and their use is mandatory.

3 The Pewan Corpus

Text corpora are essential to Computational Linguistics and Natural Language Processing. In spite the few attempts to build corpus (Gautier, 1998) and lexicon (Walther and Sagot, 2010), Kurdish still does not have any large-scale and reliable general or domain-specific corpus.

At *UoK*, we followed TREC (TREC, 2013)'s common practice and used news articles to build a text corpus for the Kurdish language. After surveying a range of options we chose two online news agencies: (i) *Peyamner* (Peyamner, 2013), a popular multi-lingual news agency based in Iraqi Kurdistan, and (ii) the Sorani (VOA, 2013b) and the Kurmanji (VOA, 2013a) websites of *Voice Of America*. Our main selection criteria were: (i) number of articles, (ii) subject diversity, and (iii) crawl-friendliness.

For each agency, we developed a crawler to fetch the articles and extract their textual content. In case of Peyamner, since articles have no language label, we additionally implemented a simple classifier that decides each page's language

[2] Although there is evidence of gender distinctions weakening in some varieties of Kurmanji (Haig and Matras, 2002).

[3] Recent research suggests that ergativity in Kurmanji is weakening due to either internally-induced change or contact with Turkish (Dixon, 1994; Dorleijn, 1996; Mahalingappa, 2010), perhaps moving towards a full nominative-accusative system.

[4] Although there are a handful of words with the latter in Kurmanji too.

Property		Sorani Corpus	Kurmanji Corpus
No. of Articles	from VOA	18,420	5,699
	from Peyamner	96,920	19,873
	total	115,340	25,572
No. of distinct words		501,054	127,272
Total no. of words		18,110,723	4,120,027
Total no. of characters		101,564,650	20,138,939
Average word length		5.6	4.8

Table 1: The Pewan Corpus's Basic Statistics

based on the occurrence of language-specific characters.

Overall, 115,340 Sorani articles and 25,572 Kurmanji articles were collected[5] . The articles are dated between 2003 and 2012 and their sizes range from 1KB to 154KB (on average 2.6KB). Table 1 summarizes the important properties of our corpus which we named *Pewan* –a Kurdish word meaning "measurement."

Using Pewan and similar to the approach employed in (Savoy, 1999), we also built a list of Kurdish stopwords. To this end, we manually examined the top 300 frequent words of each dialect and removed the corpus-specific biases (e.g., "Iraq", "Kurdistan", "Regional", "Government", "Reported" and etc). The final Sorani and Kurmanji lists contain 157 and 152 words respectively, and as in other languages, they mainly consist of prepositions.

Pewan, as well as the stopword lists can be obtained from (Pewan, 2013). We hope that making these resources publicly available, will bolster further research on Kurdish language.

4 Empirical Study

In the first part of this section, we first look at the character and word frequencies and try to obtain some insights about the phonological and lexical correlations and discrepancies between Sorani and Kurmanji.

In the second part, we investigate two well-known linguistic laws –Heaps' and Zipf's. Although these laws have been observed in many of the Indo-European languages (Lü et al., 2013), the their coefficients depend on language (Gelbukh and Sidorov, 2001) and therefore they can be used a tool to measure similarity/dissimilarity of languages. It should also be noted that in practice, knowing the coefficients of these laws is important in, for example, full-text database design, since it allows predicting some properties of the index as a function of the size of the database.

[5]The relatively small size of the Kurmanji collection is part of a more general trend. In fact, despite having a larger number of speakers, Kurmanji has far fewer online sources with raw text readily available and even those sources do not strictly follow its writing standards. This is partly a result of decades of severe restrictions on use of Kurdish language in Turkey, where the majority of Kurmanji speakers live (Hassanpour et al., 2012).

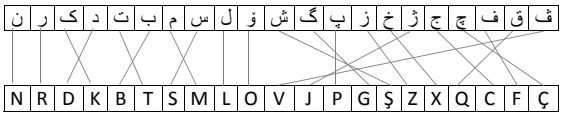


Figure 2: Relative Frequencies of Sorani and Kurmanji Characters in the Pewan Corpus

#	English Trans.	Freq.	Sorani Word
1	from	859694	لە
2	and	653876	و
3	with	358609	بە
4	for	270053	بۆ
5	which	241046	کە
6	that	170096	ئەو
7	this	83445	ئەم
8	of	74917	ی
9	together	58963	لەگەڵ
10	made/did	55138	کرد

Kurmanji Word	Freq.	English Trans.	#
û	166401	and	1
ku	112453	which	2
li	107259	from	3
de	82727	-	4
bi	79422	with	5
di	77690	at	6
ji	75064	from	7
jî	57655	too	8
xwe	35579	oneself	9
ya	31972	of	10

Figure 3: The Top 10 Most-Frequent Sorani and Kurmanji Words in Pewan

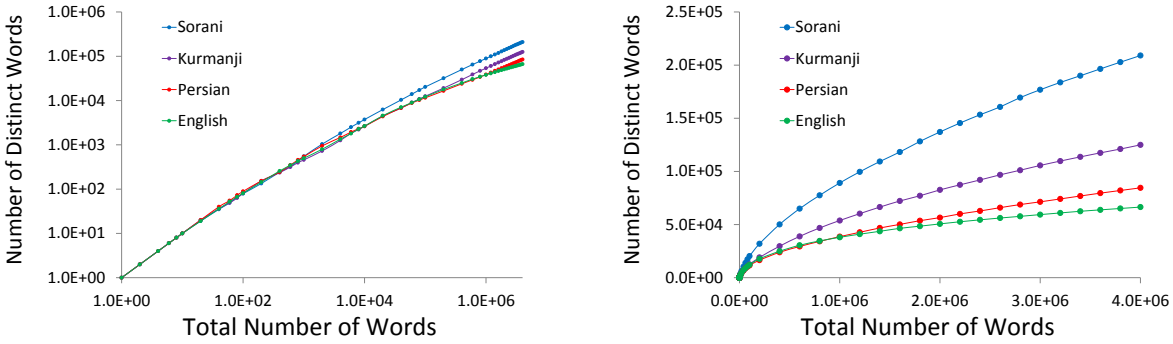
4.1 Character Frequencies

In this experiment we measure the character frequencies, as a phonological property of the language. Figure 2 shows the frequency-ranked lists (from left to right, in decreasing order) of characters of both dialects in the Pewan corpus. Note that for a fairer comparison, we have excluded characters with 1-to-0 and 1-to-2 mappings as well as three characters from the list of 1-to-1 mappings: A, Ê, and Û. The first two have a skewed frequency due to their role as Izafe construction[6] marker. The third one is mapped to a double-character (وو) in the Sorani alphabet.

Overall, the relative positions of the equivalent characters in these two lists are comparable (Figure 2). However, there are two notable discrepancies which further exhibit the intrinsic phonological differences between Sorani and Kurmanji:

- use of the character J is far more common in Kurmanji (e.g., in prepositions such as ji "from" and jî "too"),
- same holds for the character V; this is, how-

[6]Izafe construction is a shared feature of several Western Iranian languages (Samvelian, 2006). It, approximately, corresponds to the English preposition "*of*" and is added between prepositions, nouns and adjectives in a phrase (Shamsfard, 2011).



(a) Standard Representation

(b) Non-logarithmic Representation

Figure 4: Heaps' Law for Sorani and Kurmanji Kurdish, Persian, and English.

ever, due to Sorani's phonological tendency to use the phoneme `W` instead of `V`.

4.2 Word Frequencies

Figure 3 shows the most frequent Sorani and Kurmanji words in the Pewan corpus. This figure also contains the links between the words that are transliteration-equivalent and again shows a high level of correlation between the two dialects. A thorough examination of the longer version of the frequent terms' lists, not only further confirms this correlation but also reveals some other notable patterns:

- the Sorani generic preposition لە ("from") has a very wide range of use; in fact, as shown in Figure 3, it is the semantic equivalent of three common Kurmanji prepositions (`li`, `ji`, and `di`),
- in Sorani, a number of the common prepositions (e.g., پش "too") as well as the verb بوون "to be" are used as suffix,
- in Kurmanji, some of the most common prepositions are paired with a postposition (mostly `da`, `de`, and `ve`) and form circumpositions,
- the Kurmanji's passive/accusative helper verbs (`hatin` and `dan`) are among its most frequently used words.

4.3 Heaps' Law

Heaps's law (Heaps, 1978) is about the growth of distinct words (a.k.a vocabulary size). More specifically, the number of distinct words in a text is roughly proportional to an exponent of its size:

$$\log n_i \approx D + h \log i \qquad (1)$$

Language	$\log n_i$	h
Sorani	$1.91 + 0.78 \log i$	0.78
Kurmanji	$2.15 + 0.74 \log i$	0.74
Persian	$2.66 + 0.70 \log i$	0.70
English	$2.68 + 0.69 \log i$	0.69

Table 2: Heaps' Linear Regression

where n_i is the number of distinct words occurring before the running word number i, h is the exponent coefficient (between 0 and 1), and D is a constant. In a logarithmic scale, it is a straight line with about $45°$ angle (Gelbukh and Sidorov, 2001).

We carried out an experiment to measure the growth rate of distinct words for both of the Kurdish dialects as well as the Persian and English languages. In this experiment, the Persian corpus was drawn from the standard Hamshahri Collection (AleAhmad et al., 2009) and The English corpus consisted of the Editorial articles of The Guardian newspaper[7] (Guardian, 2013).

As the curves in Figure 4 and the linear regression coefficients in Table 2 show, the growth rate of distinct words in both Sorani and Kurmanji Kurdish are higher than Persian and English. This result demonstrates the morphological complexity of the Kurdish language (Samvelian, 2007; Walther, 2011). One of the driving factors behind this complexity, is the wide use of suffixes, most notably as: (i) the Izafe construction marker, (ii) the plural noun marker, and (iii) the indefinite marker.

Another important observation from this experiment is that Sorani has a higher growth rate compared to Kurmanji ($h = 0.78$ vs. $h = 0.74$).

[7]Since they are written by native speakers, cover a wide spectrum of topics between 2006 and 2013, and have clean HTML sources.

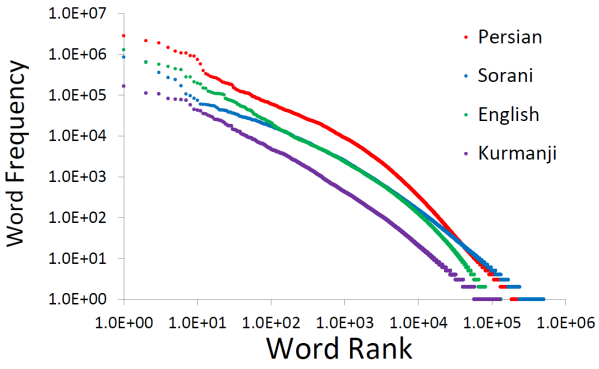


Figure 5: Zipf's Laws for Sorani and Kurmanji Kurdish, Persian, and English.

Language	$\log f_r$	z
Sorani	$7.69 - 1.33 \log r$	1.33
Kurmanji	$6.48 - 1.31 \log r$	1.31
Persian	$9.57 - 1.51 \log r$	1.51
English	$9.37 - 1.85 \log r$	1.85

Table 3: Zipf's Linear Regression

Two primary sources of these differences are: (i) the inherent linguistic differences between the two dialects as mentioned earlier (especially, Sorani's exclusive use of definite marker), (ii) the general tendency in Sorani to use prepositions and helper verbs as suffix.

4.4 Zipf's Law

The Zipf's law (Zipf, 1949) states that in any large-enough text, the frequency ranks of the words are inversely proportional to the corresponding frequencies:

$$\log f_r \approx C - z \log r \qquad (2)$$

where f_r is the frequency of the word having the rank r, z is the exponent coefficient, and C is a constant. In a logarithmic scale, it is a straight line with about $45°$ angle (Gelbukh and Sidorov, 2001).

The results of our experiment–plotted curves in Figure 5 and linear regression coefficients in Table 3– show that: (i) the distribution of the top most frequent words in Sorani is uniquely different; it first shows a sharper drop in the top 10 words and then a slower drop for the words ranked between 10 and 100, and (ii) in the remaining parts of the curves, both Kurmanji and Sorani behave similarly; this is also reflected in their values of coefficient z (1.33 and 1.31).

5 Conclusions and Future Work

In this paper we took the first steps towards addressing the two main challenges in Kurdish language processing, namely, resource scarcity and diversity. We presented Pewan, a text corpus for Sorani and Kurmanji, the two principal dialects of the Kurdish language. We also highlighted a range of differences between these two dialects and their writing systems.

The main findings of our analysis can be summarized as follows: (i) there are phonological differences between Sorani and Kurmanji; while some phonemes are non-existent in Kurmanji, some others are less-common in Sorani, (ii) they differ considerably in their vocabulary growth rates, (iii) Sorani has a peculiar frequency distribution w.r.t. its highly-common words. Some of the discrepancies are due to the existence of a generic preposition (ﻯ) in Sorani, as well as the general tendency in its writing system and style to use prepositions as suffix.

Our project at *UoK* is a work in progress. Recently, we have used the Pewan corpus to build a test collection to evaluate Kurdish Information Retrieval systems (Esmaili et al., 2013). In future, we plan to first develop stemming algorithms for both Sorani and Kurmanji and then leverage those algorithms to examine the lexical differences between the two dialects. Another avenue for future work is to build a transliteration/translation engine between Sorani and Kurmanji.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments that helped us improve the quality of the paper.

References

Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2009. Hamshahri: A standard Persian Text Collection. *Knowledge-Based Systems*, 22(5):382–387.

Wafa Barkhoda, Bahram ZahirAzami, Anvar Bahrampour, and Om-Kolsoom Shahryari. 2009. A Comparison between Allophone, Syllable, and Diphone based TTS Systems for Kurdish Language. In *Signal Processing and Information Technology (ISSPIT), 2009 IEEE International Symposium on*, pages 557–562.

Robert MW Dixon. 1994. *Ergativity*. Cambridge University Press.

Margreet Dorleijn. 1996. The Decay of Ergativity in Kurdish.

Kyumars Sheykh Esmaili, Shahin Salavati, Somayeh Yosefi, Donya Eliassi, Purya Aliabadi, Shownm Hakimi, and Asrin Mohammadi. 2013. Building a Test Collection for Sorani Kurdish. In *(to appear) Proceedings of the 10th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA '13)*.

Kyumars Sheykh Esmaili. 2012. Challenges in Kurdish Text Processing. *CoRR*, abs/1212.0074.

Gérard Gautier. 1996. A Lexicographic Environment for Kurdish Language using 4th Dimension. In *Proceedings of ICEMCO*.

Gérard Gautier. 1998. Building a Kurdish Language Corpus: An Overview of the Technical Problems. In *Proceedings of ICEMCO*.

Alexander Gelbukh and Grigori Sidorov. 2001. Zipf and Heaps Laws' Coefficients Depend on Language. In *Computational Linguistics and Intelligent Text Processing*, pages 332–335. Springer.

Guardian. 2013. The Guardian. www.guardian.co.uk/.

Goeffrey Haig and Yaron Matras. 2002. Kurdish Linguistics: A Brief Overview. *Sprachtypologie und Universalienforschung / Language Typology and Universals*, 55(1).

Amir Hassanpour, Jaffer Sheyholislami, and Tove Skutnabb-Kangas. 2012. Introduction. Kurdish: Linguicide, Resistance and Hope. *International Journal of the Sociology of Language*, 2012(217):118.

Harold Stanley Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc. Orlando, FL, USA.

Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. 2013. Deviation of Zipf's and Heaps' Laws in Human Languages with Limited Dictionary Sizes. *Scientific reports*, 3.

David N. MacKenzie. 1961. *Kurdish Dialect Studies*. Oxford University Press.

Laura Mahalingappa. 2010. The Acquisition of Split-Ergativity in Kurmanji Kurdish. In *The Proceedings of the Workshop on the Acquisition of Ergativity*.

Yaron Matras and Salih Akin. 2012. A Survey of the Kurdish Dialect Continuum. In *Proceedings of the 2nd International Conference on Kurdish Studies*.

Yaron Matras and Gertrud Reershemius. 1991. Standardization Beyond the State: the Cases of Yiddish, Kurdish and Romani. *Von Gleich and Wolff*, 1991:103–123.

Pewan. 2013. Pewan's Download Link. https://dl.dropbox.com/u/10883132/Pewan.zip.

Peyamner. 2013. Peyamner News Agency. http://www.peyamner.com/.

Pollet Samvelian. 2006. When Morphology Does Better Than Syntax: The Ezafe Construction in Persian. *Ms., Université de Paris*.

Pollet Samvelian. 2007. A Lexical Account of Sorani Kurdish Prepositions. In *The Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 235–249, Stanford. CSLI Publications.

Jacques Savoy. 1999. A Stemming Procedure and Stopword List for General French Corpora. *JASIS*, 50(10):944–952.

Faramarz Shahsavari. 2010. Laki and Kurdish. *Iran and the Caucasus*, 14(1):79–82.

Mehrnoush Shamsfard. 2011. Challenges and Open Problems in Persian Text Processing. In *Proceedings of LTC'11*.

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.

TREC. 2013. Text REtrieval Conference. http://trec.nist.gov/.

VOA. 2013a. Voice of America - Kurdish (Kurmanji) . http://www.dengeamerika.com/.

VOA. 2013b. Voice of America - Kurdish (Sorani). http://www.dengiamerika.com/.

Géraldine Walther and Benoît Sagot. 2010. Developing a Large-scale Lexicon for a Less-Resourced Language. In *SaLTMiL's Workshop on Less-resourced Languages (LREC)*.

Géraldine Walther. 2011. Fitting into Morphological Structure: Accounting for Sorani Kurdish Endoclitics. In Stefan Müller, editor, *The Proceedings of the Eighth Mediterranean Morphology Meeting (MMM8)*, pages 299–322, Cagliari, Italy.

George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.