Mehran University Research Journal of Engineering and Technology
Vol. 40, No. 2, 298 - 304, April 2021
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.2102.04

Romanized Sindhi Rules for Text Communication

Irum Naz Sodhar[1], Akhtar Hussain Jalbani[2a], Muhammad Ibrahim Channa[2b], Dil Nawaz Hakro[3]

RECEIVED ON 23.07.2019, ACCEPTED ON 03.03.2020

ABSTRACT

Sindhi is one of the historical languages which is widely used in all over the world, but especially in the province of Sindh Pakistan. Sindhi language has its own script and written by the right-handed. Nowadays the use of different Sindhi platforms is increasing especially for communication. The majority of the people of Sindh province read, write and speak very well, but they face the problem in text communication while using different communication platforms. However, the users of computer and mobile phone feel trouble/difficulty during the use of the Sindhi script in typing of text messages, tweets and comments while using different platforms in computer and mobile phone. Natural Language Processing (NLP) is one of the better options for the solution of these problems of text communication on different platforms. For the proper solution of text communication issues, Romanized Sindhi text is used instead of Sindhi text. Romanized text writing is easier than the Sindhi text writing because Sindhi text writing needs the special type of keyboard while writing of Romanized text does not need any special type of keyboard. For the writing of Romanized Sindhi text, rules are defined in this paper which provide easiness during writing and understanding of the text. Romanized Sindhi Rules (RSR) are simple and easy to understand the meaning of the text and provide fast communication (text). This study is also helpful for further research in the Romanized Sindhi text by using different approaches and provides easiness in communication.

Keywords: Sindhi Language, Natural Language Processing, Romanized Sindhi Rules, Text Communication.

1. INTRODUCTION

Sindhi is an Indo-Aryan language of the historical Sindh region in the northern part of the Indian sub-continent. Sindhi is one of the oldest languages in the world, mostly spoken in the Sindh - Province of Pakistan. According to a survey published by Pakistan Government, 47.893 Million people living in Sindh province read, write and speak Sindhi language and it is also an official language of the province [1]. Majority population of this province use Sindhi language in text communication like applications, letters as well as use text messages in today's mobile applications [2, 3]. Mobile phones, computers and laptops have become part of everyone's life. Communication through computers and mobile phones such as Short Messaging Service (SMS), Twitter, WhatsApp, Facebook and other such applications has enormously increased. English is commonly used in these types of services, however, in many countries, people prefer to communicate in their own mother tongue rather than English. Mother tongue is naturally a powerful source of communication. Therefore, in Pakistan and India researchers pay more attention on the issues related to the local languages (e.g. Urdu, Sindhi, Punjabi, Hindi).

[1] Department of Information Technology, Shaheed Benazir Bhutto University, Shaheed Benazirabad, Sindh, Pakistan. Email: irumnaz@sbbusba.edu.pk (Corresponding Author)
[2] Department of Information Technology, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan. Email: [a]jalbaniakhtar@quest.edu.pk, [b]Ibrahim.channa@quest.edu.pk
[3] Institute of Information and Communication Technology, University of Sindh, Jamshoro, Sindh, Pakistan. Email: dilnawaz@usindh.edu.pk

The Information Retrieval (IR) and Data Mining (DM) is responsible for the association between words in the sentences or investigation. Subject to classification, result analysis and sentiment classification are involved in NLP knowledge base. NLP features are described in different ways, such as stem, lemma, token, Parts of Speech Tagging (POS), Stop words, shallow-parsing and Named Entity Recognition (NER) cover major values in all NLP systems [4]. A lot of work has been done in English, Urdu and Arabic languages, but still there is a huge vacuum for research and investigation for Sindhi language. For survival of language and efficient communication in our own mother tongue, it is need of the hour to create applications in our own language. This paper therefore focusses Sindhi language.

2. RELATED WORK

Bhatti *et al.* [5-6] proposed algorithms to build a Sindhi spell checker to check spellings in the Sindhi text. They gave valuable suggestions for the misspelt Sindhi text. Phonetic-based Sindhi language rules and patterns are required for the correctness and effectiveness for the execution of a spell checking system. This system builds for the phonetic support the soundEx algorithm and shapeEx algorithm for pattern matching which produces an accurate text as well as give suggestions about Sindhi text. The authors reported in their research how to generate the suggestions for correct expression. The authors developed the software for checking of spellings of the text. The important task of system application was to correct the spelling and suggestions about the misspelt text document. This system is buildup for the misspelt checker in text document and also checks the similar words in the text. In the initial stage, it is necessitated that how to work and respond. Initial architecture had been executed for basic architecture implementation. This research depends on three most important algorithms; those algorithms are used by the authors of reference [6] for checking in a misspelt Sindhi text document. The first is for distance (Distance edit), second for phonetic named (SoundEx) and finally used for finding out the patterns (ShapeEx).

The computer is the source which is used to read and write different languages [7]. Computers need instruction for these languages, but various analysis methods are available such as OCR [8]. From these methods, Optical Character Recognition (OCR) is one of them. A research conducted by the Hakro *et al.* [8] reported that lot of work is available on OCR in Japanese, Chinese and Arabic scripts. In this research, the authors worked on the OCR system for Sindhi and Arabic scripts. The Sindhi OCR needs more effort to build an OCR system. Also in other studies, OCR is a technology which is used for the handwritten text or images to understand and write by the machine.

Hakro *et al.* [11] worked on the OCR system and reported the issues and challenges in Sindhi script by using OCR. It was very difficult to identify the printed text of the Sindhi script in OCR system. It was a challenging task because the Sindhi language had 52 Characters as compared to 39 characters in Urdu, 32 characters in Persian and 28 characters in Arabic, 26 characters in English. Writing style and forms of writing Sindhi language were as same as those of Arabic language. Sindhi was found to be more complex as compared to other languages because dots (single dot, double, triple and four) are used in Sindhi script. Multiple placements of dots were observed and these were sometimes below, above, inside and in between the characters.

Dootio and Wagan [12, 13] worked on the NLP and reported in their research that Sindhi script had many classes and characteristics of Sindhi corpus. A lot of work is in English script and NLP tools are offered in English scripts which perform all tasks of English script, but in the Sindhi language, no powerful application is available for the feature extraction and corpus. Sindhi is the right-handed written script that is as the Arabic and Urdu [9]. But the usage of Sindhi script is increasing at every platform, especially in social media, text communication is also used in various sources (online magazines, newspapers, poetry, learning websites of Sindhi) *etc*. It means huge amount of data is available at different domains of websites. At this stage, NLP tools are not available online to perform different tasks like tokenization of documents in Paragraphs, sentences, words and characters. In this research authors found out Sindhi text in parts of speech and most important task was sentiment analysis of Sindhi text in different domains.

Nowadays use of Roman script is increased as the use of English in the field of computer science/Information technology has also increased. Majority of peoples use it on social media for their communication purposes because Roman script is easier than other scripts like Sindhi, Urdu, Arabic *etc*. in writing. Researchers have developed new applications for the use of this script. In the Arabic language, authors faced a more complex situation when they compared Roman script with Arabic script. The features of the Arabic language are also similar to the Sindhi language in writing scripts [8].

Kosurru *et al.* [14] proposed a system named RoLI to identify the Romanized text for a variety of Indian languages. They developed rules to find out the Romanized text in detail for the Indian Languages. Their proposed method can be applied to any language using less number of resources. Use of RoLI gave high accuracy of 98.3% in experiments conducted over five Indian language web pages containing a mix of those languages. In this paper, the authors describe variety of methods to Romanize the text.

Ali and Ijaz [15] worked on the NLP and used large corpus for the classification task. The documents contained 19.3 million-words, classified into six names; First is finance, second-culture, third-sports, fourth-news, fifth-personal and finally consumer information. They pre-processed the contents of six domains by performing tokenization, diacritics elimination, normalization, stop word removal, stemming and also used some statistical techniques.

The current dataset has no availability of Urdu Language Processing (ULP) at the Centre for Research in ULP (CRULP) and Computing Research Laboratory (CRL) as reported by Mukund *et al.* [16]. So the authors took data from different resources to build a tool for research. Urdu and Hindi languages which have same speaking and writing forms are different from each other. Urdu is also the national language of Pakistan; Urdu language is a right-sided language. Pre-processing of any kind of language is an important step before applying any technique of NLP, sub-task of periods, worlds-removals, Stem words, *etc*. Ayesha *et al.* [17] also worked with Roman Urdu and found out the polarity of the text.

Sodhar *et al.* [18] worked on the issues and challenges in Romanized Sindhi Text and reported in their research that nowadays computer technology has become very advanced and increased number of applications have been developed for the users for communication. They faced the problem of typing of Sindhi text for communication purposes and they felt difficulty of punctuations, symbols, dots, noise issues, row break and font style.

Bhatti *et al.* [19] worked on the academic informatics portal for the Sindhi community. The authors developed software for sharing ideas, pictures and other related materials. This software is based on the PHP, JavaScript and MySQL for compatible operating systems. Bhatti *et al.* [20] also worked on the word segmentation of Sindhi language text. The authors used various techniques of NLP for the solution of the segmentation problem and they selected tokenization of Sindhi text.

3. SINDHI DATASET

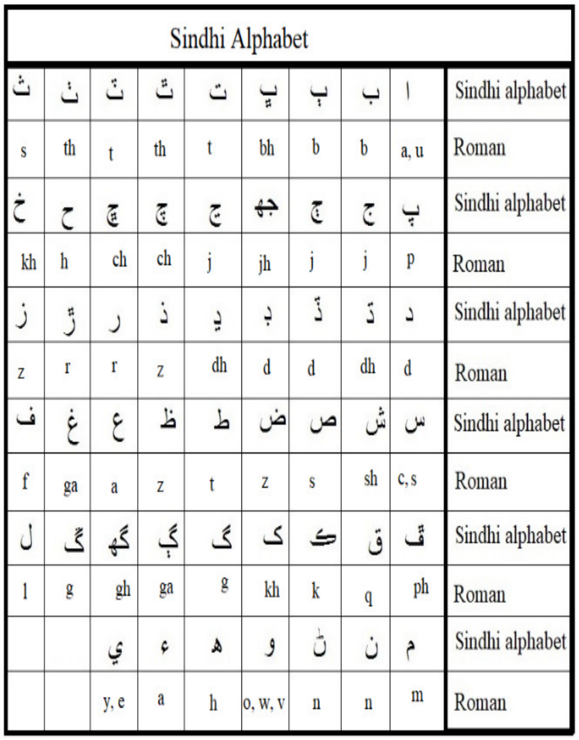
A key component to start research is availability of an appropriate dataset. NLP techniques also implement on the dataset and analyze the results completely based on the dataset. Sindhi training datasets are yet not available for the research. Huge gaps in Sindhi language are determined and availability of the resources is minimum.

3.1 Sindhi Alphabet

Sindhi language is more difficult than the other languages in speaking as well as writing because Sindhi alphabet has fifty-two (52) letters, as shown in Fig. 1. Usually in mobile phones and in computer applications we have English type keyboard for usage, but in Sindhi alphabet, all 52 letters are written in different ways and styles and they have different sounds. Due to this reason it is very difficult to write and read in the Sindhi script [8,10].

Natural selection of datasets has two ways for the construction of datasets for the Arabic-script as (a) Unicode-Character set (b) Extensible Markup Language (XML) file. Sindhi has been just like Arabic based languages and uses Unicode for the storage.

Multiple file formats are used for Sindhi text, such as ".txt" or ".doc", ". xcls", ". spss" *etc*.



Sindhi Alphabet									
ث	ٺ	ٽ	ٿ	ت	ڀ	ٻ	ب	ا	Sindhi alphabet
s	th	t	th	t	bh	b	b	a, u	Roman
خ	ح	ڇ	چ	ڃ	جھ	ڄ	ج	پ	Sindhi alphabet
kh	h	ch	ch	j	jh	j	j	p	Roman
ز	ڙ	ر	ذ	ڍ	ڊ	ڏ	ڌ	د	Sindhi alphabet
z	r	r	z	dh	d	d	dh	d	Roman
ف	غ	ع	ظ	ط	ض	ص	ش	س	Sindhi alphabet
f	ga	a	z	t	z	s	sh	c, s	Roman
ل	ڱ	گھ	ڳ	گ	ک	ڪ	ق	ڦ	Sindhi alphabet
l	g	gh	ga	g	kh	k	q	ph	Roman
		ي	ء	ه	و	ڻ	ن	م	Sindhi alphabet
		y, e	a	h	o, w, v	n	n	m	Roman

Fig. 1: Sindhi-Roman Alphabet

3.2 Romanized Sindhi Text

Romanized Sindhi text is mostly used by the Sindhi people of different regions of Sindh as well as different parts of the world for communicating with each other by using different ways. The use of Romanized Sindhi text in text messages on social networks is one of them. Resources for use of Romanized Sindhi Text includes Cell phones (Text messages and Social networks just like WhatsApp Chat, Facebook Communications, Online Websites, Translators and so on).

In Sindhi scripts many symbols are used but in Romanized Sindhi Text there is no use of symbols as described in rule No. 1. Rule 2-15 may be used to write Romanized Sindhi Text instead of Sindhi script. Vowels are the same as of the English language when write Romanized Sindhi Text as described in rule No. 16. Use of single letter (as shown in rule No. 17) of Sindhi script in words has different meanings but when these letters are used separately in sentences creates different meanings and connect the paragraphs. Many words of Sindhi script use double letters but when in Romanized Sindhi, Text may be written in different ways.

Rules For Romanized Sindhi Text		
Rule-1	No use of symbols in RST, because its time taking process when two people are communicating (in written) at different resources.	
	Example:	Paishu, zer, zabar and so on.
Rule-2	ب and ٻ used in Romanized letters are same when write. ب (ba) and ٻ (ba) most of the time written same letters in Romanized Sindhi text.	
	Examples:	Lamp →Batee→بتي
		Baru→ٻار
Rule-3	ٿ and ٺ have same sounds in Romanized Sindhi text.ٿ and ٺ Romanized form (Th and Th).	
	Examples:	Bag →Thelho→ٿيلهو
		Elbow →Thonth→ٺونٺ
Rule-4	ڇ and چ used in Romanized letters are same when write. ڇ (ch) and چ (ch) in Sindhi form and Romanized form.	
	Examples:	What → Cha →ڇا
		Umbrella →Chattee→ڇٽي
Rule-5	ح , ھ usage and sound same.	
	Examples:	Right→Haq→حق
		Hand→Hath→هٿ
Rule-6	خ and ک have sound and written form of Romanized text is same.	
	Examples:	Cot→ Khat → کٽ
		Letter→ Khaat→ خط
Rule-7	ذ, ز and ض .Mostly same sound of ذ, ز when in write and speak. In Romanized sindhi text ذ, ز and ض are same. ذ, ز and ض in Romanized form (Z).	
	Examples:	Zakhero→ذخيرو
		Zainab→Zainab→ز ينب
		Zaeef→Zaeef→ضعيف
Rule-8	ر and ڙ same pronunciation in Romanized.	
	Exampes:	Train→Rail→ ريل
Rule-9	L (لَ) is used for connectivity.	
	Example:	Abdul Hafeez→ عبدالحفيظ
Rule-10	ء (hamzah) used for supporting.(mostly used in written communication in Urdu and Sindhi languages). Hamzah not a particular character, only used for the support of particular word to start or end the word.	
	Examples:	Mother →Maau→ماء
		Father →Peeu→پيء
Rule-11	N is nasal sound so at the word is end in speaking using N sound.	
	Examples:	Me → Maa'n → مان

		Come → Achaa'n→اچان
		Go → Vancha'n → وچان
Rule-12	C, Q, and W are mostly not used in Romanized Sindhi text.	
	Example:	Car → Karu → ڪار Quilt → Ralee → رلي Watch → Gharee → گهڙي There is no use of X in Roman
Rule-13	In Romanized form Q used as K and K uses as Q.	
	Example:	Pen → Qalamu or Kalamu → قلم Shirt → Qamees or Kamees → قميص
Rule-14	In Romanized form W used as V and V uses as W.	
	Example:	Carpenter → Wadho or Vadho → وايو
Rule-15	There is no any word start with X but in a few cases it will use.	
	Example:	Zohaib → زوهيب → Xohaib

Rule-16: Vowels of Romanized Sindhi is same as like vowels used as in English.		
Roman Vowels	English Vowels	Sindhi Vowels
A	A	آ زبر,
E	E	اِي
I	I	زِير
O	O	اُو
U	U	پيش

Rule-17: Single letter in Sindhi-Romanized Text		
No.	Sindhi Words	Romanized
1.	۽	Aen
2.	□ه	Taa
3.	به	Bh
4.	نه	Naa
5.	۾	Maen
6.	ٻه	Baa

Examples: 1. Ta → □ه

2. ه → ت

3. Ali and Ahmed are close friends with each other → علي ۽ احمد پاڻ ۾ گهرا دوست آهن →Ali aen Ahmed pann maen gahera dosta aahn.

Rule-18: Double Letters used in Sindhi-Romanized Text		
No.	Sindhi Words	Romanized
1.	هڪ	Hik
2.	جو	Jow
3.	هن	Hen
4.	پڻ	Penn
5.	پٽ	Putt
6.	پر	Par
7.	ان	Una
8.	کي	Khe
9.	جي	Je
10.	ڏک	Dukh
11.	ڪم	Kamu
12.	ڪن	Kanu
13.	پڳ	Pagh
14.	چا	Chaa
15.	□ي	Te
16.	ٿي	The
17.	سر	Siru
18.	وٽ	Watt
19.	سَٻ	Sabh
20.	حق	Haq
21.	سي	Se
22.	فن	Fun
23.	ٿي	Ty
24.	جن	Jin
25.	هو	Huo
36.	هٿ	Hath
37.	حد	Had
28.	سو	So
29.	ڪپ	Kap
30.	ئي	Ee
31.	اٿ	Utha
32.	ٻن	Bin
33.	سٽ	Sita
34.	هم	Hum
35.	يا	Ya
36.	جڻ	Jaen

Examples:

1. Ali and Ahmad live in the same village → علي ۽ احمد هڪ ڳوٺ ۾ رهندا آهن→ Ali aen Ahmed hik gaoth maen rahan tha.
2. Ali was shocked at Ahmed's words → علي کي احمد جي ڳاله□ي ڏاڍو ڏک ٿيو هو → Ali khe Ahmed jy galhiyan ty dhadho dukh thiyo huo.
3. This year we will also be visiting Karachi → هن سال اسان پڻ ڪراچي گهمڻ وينداسون →hin sal asan penn Karachi ghuman wendason.

4. CONCLUSION

In this research work, Rules for Romanized Sindhi text were introduced for different sources like text messages on mobile phones, WhatsApp chat, Facebook comments, Twitter (Tweets) and so on. In the rigorous study of literature review, the Sindhi trained corpus was not available for the text classification of Romanized Sindhi text. This study provides the solution for the problem/issues faced in communication (text) and the data given in this paper is useful for computer/mobile users and users will use these rules for different purposes. This research work may be very helpful for easy communication between people of different regions of the world on social media and also very helpful for the sentiment analysis and summarization of Romanized Sindhi text in the future.

ACKNOWLEDGMENT

The authors thank the reviewers for their helpful comments. This work was conducted under the support of the Department of Information Technology, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan. The authors also gratefully acknowledge the financial support provided by the Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah.

REFERENCES

1. Sindh (Province, Pakistan) - Population Statistics, Charts, Map and Location https://www.citypopulation.de/php/ Pakistan-admin.php?adm1id=8. [Retrieved on September 4, 2019].
2. Alejandro G., Beatriz A., "Sindhi", The Languages Gulper, http://www.languagesgulper.com/eng/Sindhi.html. Retrieved December 27, 2013.
3. Christopher, S., "Sindhi Language”, Encyclopedia Britannica, https://www.britannica.com/topic/Sindhi-language, [Retrieved December 29, 2018].
4. Riaz K., “Rule-Based Named Entity Recognition in Urdu”, *Proceedings of Named Entities Workshop, Association for Computational Linguistics*, pp. 126-135, Stroutdsburg, PA, USA, 2010.
5. Bhatti Z., Ahmad W., Ismaili I.A., Hakro D.N., Soomro W.J., “Phonetic Based SoundEx and ShapeEx Algorithm for Sindhi Spell Checker System”, *Advances in Environmental Biology*, Vol. 8, No. 4, pp. 1147-1155, 2014.
6. Bhatti Z., Ismaili I.A., Hakro D.N., Soomro W.J., “Phonetic-Based Sindhi Spell Checker System Using a Hybrid Model”, *Digital Scholarship in the Humanities*, Vol. 31, No. 2, pp. 264-282, 2015.
7. Memon N., Abbasi F., Zardari S., “Glyph Identification and Character Recognition for Sindhi OCR”, *Mehran University Research Journal of Engineering and Technology*, Vol. 36, No. 4, pp. 933-940, October, 2017.
8. Hakro D.N., Talib A.Z., Bhatti Z., Mojai G.N., “A Study of Sindhi Related and Arabic Script Adapted Languages Recognition”, *Sindh University Research Journal (Science Series)*, Vol. 46, No. 3, pp. 323-334, 2014.
9. Awan S.A., Abro Z.H., Jalbani A.H., Hameed M., “Handwritten Sindhi Character Recognition Using Neural Networks”, *Mehran University Research Journal of Engineering and Technology*, Vol. 37, No. 1, pp. 191-196, January, 2018.
10. Hakro D.N., Talib A.Z., “Printed Text Image Database for Sindhi OCR,” ACM *Transactions on Asian Low-Resource Language Information Processing*, Vol. 15, No. 4, May, 2016.
11. Hakro D.N., Ismaili I.A., Talib A.Z., Bhatti Z., Mojai G.N., “Issues and Challenges in Sindhi OCR”, *Sindh University Research Journal (Science Series)*, Vol. 46, No. 2, pp. 143-152, Jamshoro, Pakistan, 2014.
12. Dootio M.A., Wagan A.I., “Unicode-8 Based Linguistics Dataset of Annotated Sindhi Text”, *Data in Brief*, Vol. 19, pp. 1504-1514, 2018.
13. Dootio M.A., Wagan A.I., “An Analysis of Sindhi Annotated Corpus Using Supervised Machine Learning Methods”, *Mehran University Research Journal of Engineering and Technology*, Vol. 38, No. 1, pp. 185-196, January 2019.
14. Paven K., Tandon N., Varma V., “Addressing Challenges in Automatic Language Identification of Romanized Text”, *Proceedings of 8th International Conference on Natural Language Processing*, Kharagpur, India, 2010.
15. Ali A., Ijaz M., “Urdu Text Classification”, *Proceedings of 7th ACM International Conference on Frontiers of Information Technology*, Abbottabad, Pakistan, pp. 1-7, 16-18 December 2009.
16. Mukund S., Srihari R., Peterson E., “An Information-Extraction System for Urdu - A Resource-Poor Language”, *ACM Transactions on Asian Language Information Processing*, Vol. 9, No. 4, pp. 15, 2010.
17. Ayesha R, Malik, M.K., Bukhari, Z.N.F., and Jalbani, A.H., “Sentiment Analysis for Roman Urdu”, *Mehran University Research Journal of Engineering and Technology*, Vol. 38, No. 2, pp. 389-396, Jamshoro, Pakistan, 2019.
18. Sodhar I.N., Jalbani A.H., Channa M.I., Hakro D.N., “Identification of Issues and Challenges in Romanized Sindhi Text”, *International Journal of Advanced Computer Science and Application*, Vol. 10, No. 9, pp. 229-233, 2019.
19. Bhatti Z., Hakro D.N., Jarwar A.A., “Sindhi Academic Informatic Portal”, *American Journal*

of Information Systems, Vol. 1, No. 1, pp. 21-25, 2013.

20. Bhatti Z., Ismaili I.A., Soomro W.J., Hakro D.N., “Word Segmentation Model for Sindhi Text”, *American Journal of Computing Research Repository*, Vol. 2, No. 1, pp. 1-7, 2014.