Identification of Issues and Challenges in Romanized Sindhi Text

Irum Naz Sodhar[1]
Post Graduate Student, Department of Information Technology, Quaid-e-Awam University of Engineering Science and Technology, Nawabshah, Sindh, Pakistan

Muhammad Ibrahim Channa[3]
Professor, Department of Information Technology Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan

Akhtar Hussain Jalbani[2]
Associate Professor, Department of Information Technology, Quaid-e-Awam University of Engineering Science and Technology, Nawabshah, Sindh, Pakistan

Dil Nawaz Hakro[4]
Associate Professor, Institute of Information and Communication Technology, (IICT) University of Sindh, Jamshoro, Sindh, Pakistan.

***Abstract*—Now-a-days Sindhi language is widely used in internet for the various purposes such as: newspapers, Sindhi literature, books, educational/official websites and social networks communications, teaching and learning processes. Having developed technology of computer system, users face difficulties and problems in writing Sindhi script. In this study, various issues and challenges come in the Romanized Sindhi text by using Roman transliteration (Sindhi text (ST) forms of Romanized Sindhi text) are identified. These acknowledged issues are known as noise, written script of Romanized and its style, space issues in Romanized script, some characters not suitable in Romanized Sindhi, as a paragraph, rows, character issues, punctuation, row break and font style. However, this study provides the summary of issues and challenges of Romanized Sindhi text. This research work provides detailed information of issues and challenges faced by people during chatting in Romanized Sindhi text.**

Keywords*—Romanized Sindhi Text (RST); Sindhi language; issues and challenges; transliterator; social networks communication

I. Introduction

Sindhi Language is a historical language of the world, the majority of Sindhi language speakers are inhabited in Sindh province of Pakistan. Around 12% peoples of Pakistan have mother tongue is Sindhi and an official language of the Sindh [1]. Sindhi language is also spoken in different part of the world with different ratio. Sindhi language has its own script and written format. In Sindhi Language 52 alphabetical letters (Fig. 1) were used for written as well as in speaking purposes [2-4]. Since, Sindh language contains more alphabetical letters than other languages, which causes difficulties for the new learners. Sindhi script writing is a right handed script, same as Arabic and Urdu Script. Urdu is morphological prosperous, having different type of characters in Urdu script. Sindhi script follows the rules as like Arabic Script and Perso-Arabic script [5].

In these days Sindhi language is considered as extensively used in internet for the various purposes such as: daily newspapers, Sindhi literature, books, educational/official websites, social network communication (What's App, Text messages, and social network), Teaching and learning processes. In this regard, the use of the keyboard (Sindhi) is being increased day by day and on the other hand people are still facing the problem of unavailability of Sindhi keyboards. However, the communication system of local users is carried on by android based mobile phones services; these mobile phones are unable to provide facilities to write Sindhi language containing 52 letters. Therefore, to overcome these problems, Romanized Sindhi text is one of the best options [6].

Romanized Sindhi text is when used in different plate forms may face many issues and problems in writing of Romanized Sindhi text or when use of different translators for Sindhi of Romanized text. Also the use of translators and other sources for normal users are very difficult and they need an easy way for the solution of the problem.

New issues and challenges of Sindhi language has been found, when communicating with each other in Romanized Sindhi text because Sindhi language has 52 letters of alphabets having different shapes, different symbols and different orientation of dots. So, it is very difficult to communicate using Sindhi language on different social media. Therefore it is very important to have such platforms where people of different Sindhi community can communicate easily and properly using Romanized format.

Fig. 1. Sindhi Language Alphabet.

II. Related Work

In Sindhi data set construction, issues contain corpus acquisition; pre-processing and tokenization is discussed in this paper. The results of those issues based on observation which contains unigram, bigram and trigram frequencies; author explores the orthography and Sindhi script data construction [7]. The word corpus was used by German Scholar at first time. The plural of corpus was corpora, which was used for a huge number of text data consists of either millions or billions of data. Processing was challenging because scarcity of resources for computational linguistics and research, different text data have been developed in different languages of different countries [6].

A model for transliteration was provided by Leghari and Arain, this model provided two scripts of Sindhi language one was Perso-Arabic Script and other was Devanagari Script. Analyzing of both scripts, authors suggested that data on Roman Script also used for Sindhi Language and they proposed an algorithm for transliteration between two scripts [4].

In another research paper authors addressed the issues of Sindhi word Segmentation and provide different techniques to implement on different algorithms [8]. Current research on multi linguistic writing has been carried out in transliteration. Authors in [9] explored English related forms, writing with Romanized Greek characters. Authors in [5] fond out the challenges in Urdu text and tokenized the Urdu text and also detected the sentence boundary. This was very difficult task for comparison of tokenization and detected the sentence boundary.

OCR is an Optical Character Recognition used in written text or (as well as) in printed documents. Authors in [4] found out the issues and challenges in Sindhi OCR which contains many character dots, different placement and direction of dots. The authors also provided the summary for issues and challenges related to the development of Sindhi OCR.

A research work was done on the sentiment analysis of Arabic Language facing an issue that was unstructured and non-grammatical text. Results were analyzed by using various parameters: accuracy, precision, recall and F-score [10]. Authors in [11, 12] also worked on the sentiment analyses of Arabic language by using support vector machine technique. This technique was more accurate for classification for Sentiment summarization and analysis of the Sindhi text by using machine learning techniques DTM and TF-IDF. DTM and TF-IDF analysis was used by n-gram model. The supervised machine learning model was mostly used for Sindhi text Sentiment analysis [13].

Authors in [14] worked on the sentiment analysis of Urdu text by using sentiment classification model. This system extracts Senti Units and the target expressions through the shallow parsing based chunking. It is observed that dependency parsing algorithm created associations between these extracted expressions and measure the results either positive or negative.

Dootio and Wagan did research on the Development of Sindhi alphabet and reported that Sindhi language is widely used in all over the world. They added that mostly its literature is used in printed forms such as in books, in newspaper, in online learning websites and in different web pages on internet to construct Sindhi data set. The authors also used NLP techniques and developed the Sindhi text corpora for the use of Sindhi script [1].

Form literature, it is observed that there is still a huge space available for the research in Sindhi language to improve its written format. It is concluded from the literature that Sindhi script is widely used, but many issues and challenges come in writing forms different sources. But Romanized Sindhi text is also one of the ways to reduce the issues and challenges in Sindhi Text.

III. Materials and Methods

A. Data Set of Sindhi Script

In this study Sindhi script was used in the Romanized Sindhi Text. This Sindhi script was transliterated by online tools which are easily available. After the translation of Sindhi script into Romanized Sindhi text was checked for correct transliteration text and for errors in transliteration as shown in Fig. 2. In this relation, Sindhi data were selected from different sources and are easily available in online sources such as: newspapers, Poetry websites, Sindhi Facebook pages, Text messages, etc. The sources used in this study were then verified from the various sources and updated on a daily basis as described in Table I.

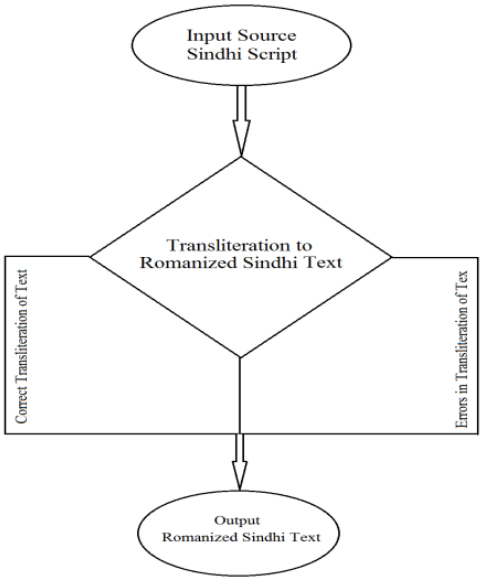

Fig. 2. Research Methodology Diagram.

TABLE. I. Sindhi Data Resources

Resources	Websites
Awami Awaz	https://awamiawaz.pk
Jhoongar	http://dailyjhoongar.com
Sindhi Poetry in roman	http://romansindhi.blogspot.com/2017/
Sindhi Adabi Board	http://www.sindhiadabiboard.org/Catalogue/Poetry/Book92/Book_page19.html
Sindhi Learning	http://sindhila.edu.pk

B. Dots in Sindhi Script

As Sindhi script has fifty two (52) alphabetical Letters and all letters Sindhi words along with Romanized words are given in Table II. In Table II, four letters have no words to start with these letters (گ,چ, ٽ, ڙ) those letters were used in middle or end of words (used for complete the words) as shown in Fig. 3.

Arabic script has 28 characters, Urdu has 39 characters while Sindhi script has 52 characters. Sindhi script is the right handed language as like Arabic and Urdu. The use of Dots is very important in these scripts, otherwise the letters had no any sense and meaning and very difficult to pronounce these letters. Arabic script is mostly used up to three dots, but in Sindhi script four different types of dots (one-1), (Two-2), (Three-3) and (Fourth-4) are used. These dots are placed at different position in letters such as above, below and inside the letter and dots are placed in letters at horizontal or vertical direction. In Sindhi script a few words have many dots in a single word just like (fifar in the Romanized Sindhi text and Lung in English), 12 dots are used in this (fifar) single word of Sindhi script.

Thirty four letters are used with 1 to 4 dots in different axis and positions in Sindhi alphabet. Table III shows details of the Characters of Sindhi script with dots, one letter used small symbol and without dots are presented in this table.

Single dot letters in Sindhi script is shown in Table IV, single dot is placed in letters below, above and inside is used. A total of nine (9) letters having single dot is shown in below Table IV. Double dots or two dots are used in horizontal or vertical direction and placed below, above and inside of the letter. Total eight (8) letters having double dots are used as shown in below Table V.

There are seven (7) letters, which take three dots were used in Sindhi script and placed in three positions below, above and inside of the letter shown in Table VI. Below placement of dots in letter is only (one-1), dots in the horizontal direction three (3) letters are placed, dots in the above position in letters are only (two). The characters having dots in the direction vertical and inside placement is only (1) one. In Sindhi script four dots are also used in characters as shown in Table VII. Total five (5) letters having four dots were used in above (3), below (1) and inside (1) the letter.

ٽ → وڻ → Wanu → Tree

Fig. 3. Sindhi Letter Wanu.

TABLE. II. SINDHI CHARACTERS WITH SINDHI WORDS AND ROMANIZED SINDHI WORDS

S. No.	Sindhi Letters	Word Start with Letter	Romanized Words
1	ا	انب	Unab
2	ب	بدڪ	Badk
3	ٻ	ٻڪ	buk
4	پ	پگ	pag
5	ڀ	ڀٽ	bhiti
6	ت	تارو	taro
7	ٿ	ٿلهو	thulho
8	ٽ	ٽبي	ttbi
9	ٺ	ٺونٺ	thonth
19	ٹ	ٹواب	Sawab
11	ج	جبل	jabal
12	جھ	جھرڪي	jharki
13	ڄ	ڄبھ	jibh
14	ڃ	ڃڃ	janj
15	چ	چپ	chupu
16	ڇ	ڇڇ	chahj
17	ح	حجم	hajjm
18	خ	خچر	khchr
19	د	در	daru
20	ڏ	ڏوبي	dhobi
21	ڊ	ڊيل	Del
22	ڈ	ر ڈيڈ	dedr
23	ڊ	ڊور	dhor
24	ذ	ذرو	zaro
25	ر	رازو	Razo
26	ڑ	رڑ	Radh
27	ز	زمين	Zamen
28	س	ساز	Saaz
29	ش	شخص	Shakis
30	ص	صوف	Soof
31	ض	ضعيف	Zaif
32	ط	طوطو	Toto
33	ظ	ظلم	Zulum
34	ع	عينڪ	Ayeenak
35	غ	غلام	Ghulam
36	ف	فروش	Farosh
37	ڦ	ڦوهارو	Foharo
38	ق	قلم	Kalam
39	ڪ	ڪعبو	Kabo
40	ک	کٹ	Khat
41	گ	گانو	Gano
42	ڳ	ڳئون	Gaon
43	گھ	گھوڑو	Ghoro
44	ڱ	سڱ	Singhan
45	ل	لکير	Lakeer
46	م	مفهوم	Mafhoom
47	ن	نالو	Nalo
48	ڻ	وڻ	Wandh
49	و	وڏيڪ	Wadheek
50	ھ	ھڪڑو	Hikro
51	ي	يڳانو	Yagano
52	ء	No start word	Hamzo

TABLE. III. DOTS USED IN SINDHI ALPHABET SYMBOLS

Sindhi Letters	Number of letters	Dots	Symbols	Total letters
ب, ج, جھ, خ, ڊ, ذ, ز, ض, ظ, غ, ف, ن	12	One dot or Single dot used	Not used	12
چ, ڃ, ڈ, ڊ, ق, ڀ, ٽ, ڦ, گ, ي	11	Two dots or double dots used	Not used	11
پ, ٺ, ٹ, ڇ, ڏ, ش	6	Three dots or triple dots used	Not used	6
ٻ, ٽ, ڇ, ڑ, ڦ	5	Four dots used	Not used	5
ڻ	1	Not used	(ط)	1
س, ص, ط, ع, ڪ, ک, گ, ڳ, ل, م, و, ه, ء, ا, ح, د, ر,	17	Not used	Not used	17
Total Letters	52			

TABLE. IV. SINGLE DOT LETTER IN SINDHI ALPHABET

S. NO:	Placement	Number of Letter	Placement of Dots
1.	Above or Up	04	خ , ذ , ز , غ
2.	Below or Down	03	ب , ڄ , ڊ
3.	Insider or Within	02	ج , ن

TABLE. V. DOUBLE DOTS CHARACTER IN SINDHI ALPHABET

S. NO	Placement of in dots in letters	Direction	
		Horizontal	Vertical
1.	Above letters	ت , گ	ٺ
2.	Below letter	ي	ٻ , ٻ
3.	Inside letter	ڇ	ڇ

TABLE. VI. THREE DOTS CHARACTER IN SINDHI ALPHABET

S. NO:	Placement of in dots in letters	Direction	
		Horizontal	Vertical
1.	Above letters	ٽ,ٿ , ش	ٽ , ش
2.	Below letter	ڀ	NA
3.	Inside letter	ڃ	NA

TABLE. VII. FOUR DOTS CHARACTER IN SINDHI ALPHABET

S. NO.	Placement	Letters
1.	Above letters	ڦ ,ڙ , ٽ
2.	Below letter	ڀ
3.	Inside letter	ڃ

IV. ISSUES AND CHALLENGES IN ROMANIZED SINDHI TEXT

A. Noise Letters

Use of dots in Sindhi letters shows the appearance and pronunciation of the alphabet. Dots are also used in the meaning of the sentence. Sindhi script contains (Seventeen-17) letters are used without dots and (Thirty Five-35) letters are used with dots. But few letters have an important appearance in words for their correct noise and meaning as shown in Table VIII. When these letters used in Sindhi words, sentences, paragraphs create problem in converting into the Romanized text by using transliterate. Letter ڻ have no doubt used as in letters/words, but a small symbol used in the letters in Sindhi script as shown in Fig. 4. Due to this reason it is very difficult to recognize the use of this letter in transliteration in the Romanized Sindhi text. Above Table II shows there is no any word on Sindhi Script start with ڻ, but it was used in the middle of the words or end of the words.

B. Font of Letters

Font of the letter is imported part of written communication and beautification of letters according to the situation, but in transliteration of text has no any facility available of font family for the users need. The font is one of the main issues in transliteration of Sindhi text into the Romanized Sindhi Text as shown in Table VIII.

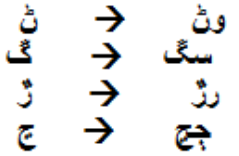

Fig. 4. Noise Letters.

TABLE. VIII. SUMMARY OF ISSUES AND CHALLENGES

S. NO	Issues and Challenges	Explanation
1	Written script with Romanized style as the English style	Left Handed Romanized Script and same as English Example: **Hinaa: aslaamu alekum** **حنا: السلام عليڪم**
2	Space Issues in Romanized Script	When no space in a row, so word completely not show in the same row, but the word must be tokenized last letter into next line that's why problem occur in reading text for readers. Example: **ننين مالي سال 20-2019ع جي سالياني بجيٽ پيش ڪندي چيو** naen maale saala 20-2019a je saalaeunae bjett peshu kande chayo
3	Some Characters are not suitable in Romanized Sindhi	Hamzo and Wao, NN, Dhe, Example: **Asmaa: khhudaa hafiz** **اسماء: خداحافظ**
4.	Paragraph	Not show properly according to paragraph
6	Row	Not Identify either row is complete or next line start.
7	Characters Issues	Sindhi characters have 52 and Roman English have 26 characters so feel difficult to pronunciation of complex word. Example: **توهان→ t~vhaa`n**
8	Punctuation	Comma, Question, Double Quotation and so on show revert in translating text that's why feel difficult to read.
9	Row Break	غريب→ Gwhen this word comes row breakdown and other words shows on the next line.
10	Font Style	No any facility available to change the text in different Font style.

C. Punctuation of Letters

Fourteen-14 punctuation is used frequently in text communication, but much punctuation is changed their positions as well as axis after transliteration Sindhi text into Romanized text. After the transliteration punctuations are not changed, but they are still in the same condition as before transliteration as shown in Table VIII. These issues come in punctuation because translators do not follow the punctuation rules when the Sindhi script is converted into the Romanized script.

In Table VIII, different issues and challenges are presented, these issues and challenges are almost occurs after transliteration of Sindhi text into Romanized Sindhi text by using online converters.

V. CONCLUSION

Sindhi Script is morphological rich in literature; it is written and spoken by worldwide. Now-a-days Romanized Sindhi text is used as communication purpose. The written style of Sindhi script is right handed as same as Arabic and Urdu written style. In writing of Romanized text there is no use of dots or small symbol, but still limited resources are available for Romanized Sindhi text online conversion. Many issues and challenges were found during Sindhi script translated into Romanized Sindhi Text which are: written script of Romanized and style, space issues with Romanized script, some characters not suitable in Romanized Sindhi, paragraph, rows, character issues, punctuation, row break and font style were observed. This research work may be very helpful for sentiment analysis and summarization of Romanized Sindhi text in the future.

REFERENCES

[1] M. A. Dootio, & A. I. Wagan, "Development of Sindhi text corpus," Journal of King Saud University-Computer and Information Sciences, 2019.

[2] A. Pirzado, "Sindhi language and literature (a brief account)," Hyderabad, Sindh: Sindhi language Authority (2009).

[3] M. Leghari, & M. U. Rahman, "Towards Transliteration between Sindhi Scripts by using Roman Script," In Conference on Language and Technology, 2010.

[4] D. N. Hakro, I. A. Ismaili, A. Z. Talib, Z. Bhatti, & G. N. Mojai, "Issues and challenges in Sindhi OCR," Sindh University Research Journal (Science Series), 46(2), 143-152, 2014.

[5] Z. Rehman, W, Anwar, & U. I. Bajwa, "Challenges in Urdu text tokenization and sentence boundary disambiguation," In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP), pp. 40-45, 2011.

[6] F. H. Khoso, M. A. Memon, H. Nawaz, & S. H. A. Musavi, (2019). To Build Corpus Of Sindhi.

[7] M. U. Rahman, "Towards Sindhi corpus construction," In Conference on Language and Technology, Lahore, Pakistan. 2010.

[8] Z. Bhatti, I. A. Ismaili, W. J. Soomro, & D. N. Hakro, "Word segmentation model for Sindhi text," American Journal of Computing Research Repository, 2(1), 1-7, 2014.

[9] T, Spilioti, "From transliteration to trans-scripting: Creativity and multilingual writing on the internet," Discourse, Context & Media, 29, 100294, 2019.

[10] A. Assiri, A. Emam, & H. Aldossari, "Arabic sentiment analysis: a survey," International Journal of Advanced Computer Science and Applications, 6(12), 75-85, 2015.

[11] A. Ziani, N. Azizi, D. Zenakhra, S. Cheriguene, & M. Aldwairi, "Combining RSS-SVM with genetic algorithm for Arabic opinions analysis," International Journal of Intelligent Systems Technologies and Applications, 18(1-2), 152-178, 2019.

[12] H. Al Suwaidi, T. R. Soomro, & K. Shaalan, "Sentiment analysis for emiriti dialects in twitter," Sindh University Research Journal-SURJ (Science Series), 48(4), 2016.

[13] M. Ali, & A. I. Wagan, "Sentiment summerization and analysis of Sindhi text," Int. J. Adv. Comput. Sci. Appl, 8(10), 296-300, 2017.

[14] A. Z. Syed, M. Aslam, & A. M. Martinez-Enriquez, "Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text," Artificial intelligence review, 41(4), 535-561, 2014.