Towards Transliteration between Sindhi Scripts by using Roman Script

Mehwish Leghari and Mutee U Rahman
Department of Computer Science, Isra University, Hyderabad Sindh 71000, Pakistan
legharimehwish@hotmail.com , muteeurahman@gmail.com

Abstract

In this research a model for transliteration is presented for two scripts of Sindhi language that is Perso-Arabic script and Devanagari script, based on an intermediate Roman script. After analyzing both Perso-Arabic and Devanagari scripts, a set of Roman script for Sindhi language is also suggested. Different issues, complexities and problems of Sindhi transliteration are discussed in detail. An algorithm to transliterate between two scripts of Sindhi language is also proposed.

1. Introduction

Transliteration is transformation of text from one script to another, usually based on phonetic equivalencies [1]. Popularity and simplicity of roman script is a major reason behind the motivation of transliteration of language scripts. People and software are getting benefited from these transliteration aids. Transliteration from native scripts to Roman script has been achieved for many Asian languages including Arabic, Bengali, Persian, Hindi, Punjabi and Urdu. Transliteration between Punjabi scripts [2] and Hindi to Urdu transliteration [3] are key examples of South Asian language transliterations. Different transliteration applications like Google Transliteration IME facilitate the users of different languages to transliterate from their native scripts to Roman script. It is currently available for 19 different languages including Arabic, Bengali, Farsi (Persian), Gujarati, Hindi, Punjabi, Sanskrit and Urdu. Transliteration of two scripts of Sindhi has not been achieved yet and in fact not even initiated.

Sindhi computational linguistics have not received much encouragement in either Pakistan or India [4] and works are limited to font design and word processing. But Sindhi computing should not only encircle font design and word processing but extensive research is needed in the areas of artificial intelligence, computational linguistics, natural language processing, corpus linguistics and script processing (including transliteration) [5].

Following sections discuss a brief history of Sindhi language, Perso-Arabic and Devanagari scripts of Sindhi language, composition of both scripts, a set of suggested Roman script for Sindhi language and an algorithm for transliteration between two scripts of Sindhi language.

2. Sindhi Language

Sindhi is an Indo-Aryan language with its roots in ancient history. Sindhi is being spoken by approximately 40 million [6] people in Sindh province of Pakistan as well as in several states of India. In Pakistan Sindhi is written in Perso-Arabic script while in India Sindhi is written in both Devanagari and Perso-Arabic scripts. Both regions heavily share the same vocabulary.

2.1. Sindhi Scripts

Sindhi scripts and their writing systems are briefly described below.

2.1.1. Perso-Arabic Script. The Perso-Arabic script of Sindhi language consists of 52 letters; most of those are taken mainly from Arabic alphabet, some letters from Persian and few modified letters. In Perso-Arabic script each letter has one to four forms according to its position beginning, middle, final and standalone. Letters in Perso-Arabic script are divided in different types on the basis of phonemes. These different types are discussed below with reference to their phonemes, writing style and position in a word.

First type is aspirated consonants. In Perso-Arabic script some of the aspirated sounds are written by combining two letters. For example aspirated form of گ can be written by combining گ (g) with ھ (h) as گھ (gh) similarly there are some other aspirated consonants. In Roman script "h" is combined with the letter of nearby sound. For example "g" + "h" = "gh" is used to represent گھ (gh). On the other hand there are some aspirated consonants in Sindhi that are represented using a single letter like "ڇ" (chh) and "ٿ" (th).

The non-aspirated consonants of Perso-Arabic script are transliterated according to their phonemes. At some places we find multiple non aspirated consonants for single phoneme and these all have single counterparts in Devanagari script. These are further discussed in section 3.

There are three main vowels in Perso-Arabic script those are ا و and ي these vowels when come at the beginning of a word are simply treated as non-aspirated consonants; and at the end of the word these are treated as vowels. But in the middle of a word they need to be tackled in context of nearby letters whether those are vowels or consonants.

The diacritical marks are essential for correct accent and if missing not only create ambiguity in transliteration but also cause misinterpretation of the words. Thus diacritical marks are equally important for avoiding ambiguities in transliteration as those are important for natural language processing and speech synthesis [7]. Some of the examples of diacritical marks include Zabar, Zer and Pesh. Examples of pairs of words that differ in meaning only because of difference in diacritical marks are shown in table 1.

Table 1. Effects of diacritical marks on meaning.

Word	Meaning	Word	Meaning
اَٹ	Eight	اَٹ	Camel
چَپُ	Lip	چُپُ	Silent
کِلَ	Laughter	کَلَ	Skin
مَلْڻُ	To rub	مِلْڻُ	To meet
ڪَرَڻُ	To do	ڪِرَڻُ	To fall

In Sindhi, there are four implosive stops. Using Perso-Arabic script those are represented by adding extra “dot(s)” to the letter of nearby matching sound. For example implosive version of گ (g) is گ̣ (gg) similarly there are three more implosive stops or implosive sounds. In Roman script these letters are represented by doubling the letters of nearby sounds like: “bb”, “jj”, “dd”, “gg” as this convention is commonly used in Sindhi-Roman script.

2.1.2. Devanagari Script. Devanagari script adopted from Sanskrit system of writing in which each character represents a syllable. Devanagari script is written from left to right. Many of the letters having same phonemes or sounds in Perso-Arabic script are equivalent of a single letter in Devanagari script. It is also worth mention here that two letters of Perso-Arabic script have no equivalent in Devanagari script those are ء (‘a) and ع (A).

Devanagari script also have aspirated and non-aspirated consonants but unlike Perso-Arabic script the Devanagari script do not use composite letters for aspirated consonants. Hence all aspirated consonants are denoted by a single letter.

In Devanagari script two types of vowels are independent and dependent vowels. Independent form of a vowel is used at the beginning of a word and dependent form is used at the end of a word. While in the middle, usually dependent form is used but there are some exceptions.

Unlike Perso-Arabic script the diacritical marks are not optional in Devanagari script. As shown in Example 1.

Example 1.

तूं सुहिणी आहीं.
tooN suhiNree aaheeN.
You beautiful are
تُون سُهڻِي آهِين.
You are beautiful.

Table 2. Diacritical marks in Devanagari.

Devanagari	Roman	Perso-Arabic
त+ू+ं = तूं	tooN	تُون
स+ु+ह+ि+ ण+ी = सुहिणी	suhiNree	سُهڻِي
आ+ह+ी+ं = आहीं	AaheeN	آهِين

As shown in table 2 diacritical marks are most widely used and are integral part of words written in Devanagari script. Thus transliteration accuracy is assured while going from Devanagari to Perso-Arabic script.

3. Perso-Arabic, Devanagari and Roman scripts

Roman script is based on the alphabet developed by the ancient Romans, and used by most of the languages of Europe, including English, French, and German [8]. To achieve Sindhi transliteration an intermediate Roman script is used. As writing Sindhi and other languages in Roman script (English) is very common nowadays, so in this model, all possible steps are taken to preserve most common Roman style of writing. Therefore one do not feel any difficulty in adopting this method and can transliterate in any direction from Sindhi to Sindhi (Perso-Arabic, Roman and Devanagari or vice versa). Most of the consonants are transliterated to their matching sounds in Roman script.

Table 3. Simple consonants and independent vowels in Devanagari, Roman and Perso-Arabic scripts.

Devanagari	Roman	Perso-Arabic
आ	Aa	آ
अ	a	ا
ब	b	ب
भ	bh	ڀ
थ	th	ٿ
ट	T	ٽ
ठ	Thh	ٺ
प	p	پ
ज	j	ج
झ	jh	ڄ
ञ	nn	ڃ
च	ch	چ
छ	chh	ڇ
ख़	khh	خ
द	d	د
ध	dh	ڌ
ड	D	ڊ
ढ	Dh	ڍ
र	r	ر
ड़	R	ڑ
श	sh	ش
ग़	G	غ
फ़	f	ف
फ	ph	ڦ
क़	q	ق
क	k	ڪ
ख	kh	ک
ग	g	گ
घ	gh	گھ
ङ	Ng	ڱ
ल	l	ل
म	m	م
न	n	ن
ं	N	ن
ण	Nr	ڻ
व	v	و
य	y	ي

Perso-Arabic, Devanagari and equivalent Roman script mapping is shown in table 3. Table 3 contains all consonants except those having same phonemes with others and implosive stops. Four unique implosive stops in Sindhi shown in table 4.

Table 4. Implosive stops.

Devanagari	Roman	Perso-Arabic
ब॒	bb	ٻ
ज॒	jj	ڃ
ड॒	dd	ڈ
ग॒	gg	ڳ

Besides the letters listed in table 3 and table 4 there are multiple letters in Sindhi, using Perso-Arabic script with same equivalent in Devanagari script, as shown in table 5. This is because of similar sounds, though we have suggested separate equivalents in Roman script for these letters.

Table 5. Perso-Arabic multiple consonants with same Devanagari equivalents.

Devanagari	Roman	Perso-Arabic
त	t	ت
त	Tt	ط
ह	H	ح
ह	h	ه
ज़	Z	ذ
ज़	z	ز
ज़	zz	ض
ज़	Zz	ظ
स	s	س
स	S	ص
स	c	ث

The letter ع came in Sindhi alphabet from Arabic. Native Sindhi speakers do not pronounce it properly (generating sound from inner throat) in normal conversations and there is no equivalent of ع (A) in Devanagari script [9]. Same is true for ء (‘a) of the Perso-Arabic script. These two letters can be transliterated easily from Perso-Arabic to Roman script and vice versa. In case of Devanagari transliteration these are either ignored or transliterated into अ (a) or े (e). Mostly these letters are ignored during transliteration of Perso-Arabic or Roman script to Devanagari script as shown in example 2 and table 6.

Example 2.

معاف ڪجو!

mAaaf kajo!

forgive do

माफ़ कजो !

Do forgive!

Table 6. Letters with no equivalents.

Devanagari	Roman	Perso-Arabic
म	m	م
-	A	ع
ा	aa	ا
फ़	f	ف

We can clearly illustrate from example 2, by further analyzing the word معاف (mAaaf) in table 6 that the use of ع (A) has been completely omitted in transliteration from Roman / Perso-Arabic to Devanagari. Those two letters are separately shown in table 7, with their equivalent Roman letters.

Table 7. Perso-Arabic letters with no Devanagari equivalent.

Devanagari	Roman	Perso-Arabic
-	A	ع
-	‘a	ء

There are two special words that are written in some special form by using single letter and two elongated quotation marks beneath the letter. These are shown in table 8 in all three scripts. Note that roman representation of these letters is capitalized to avoid ambiguity with other letters in the word or sentence.

Table 8. Special single letter words of Sindhi.

Devanagari	Roman	Perso-Arabic
ऐं	AEN	۽
मैं	MEN	۾

The dependent vowels and diacritical marks are shown in table 9 in all three scripts. These are shown in combination with letter “ज” (j) in Devanagari, “j” in Roman script and “ج” (j) in Perso-Arabic script to make it more understandable.

4. Sample Conversions and Problems

Figure 1 shows the transliteration model for Sindhi scripts.

Table 9. Dependent vowels and diacritical marks

Devanagari	Roman	Perso-Arabic
ज	ja	جَ
ज + ि= जि	ji	جِ
ज + ु =जु	ju	جُ
ज+ो=जो	jo	جو
ज + ू =जू	joo	جُو
ज + े =जे	je	جي
ज + ी =जी	jee	جِي
ज+ा=जा	jaa	جا

The dictionary lookup is used for transliteration of the words, in which one or more letters follow none of the rules. The conversions from one script of Sindhi to any other script can be achieved by implementing the rules given below:

i. Take a whole word as input.
ii. Dictionary lookup for especial words.
iii. If not in dictionary, start transliterating.
iv. Transliterate the first letter as a consonant or independent vowel.
v. From second to second last letter if any letter is consonant or vowel with a diacritical mark, transliterate it as consonant.
vi. If the letter is a vowel and it has no diacritical mark, transliterate it as dependent vowel.
vii. If last letter is vowel, transliterate it as dependent vowel.

The model and algorithm suggested in this research is designed on the basis of above rules. These rules get more complex while transliterating from Perso-Arabic to Devanagari when there are words without proper diacritical marks. In this situation transliteration is done by analyzing the letters that come before and after the letter that have no diacritical marks. If there is no match with any condition (for example in case of consecutive vowels) then finally transliteration can be achieved on probability basis. A sample transliteration is shown in example 3.

Example 3.

बा॒हिर डा॒ढी गरमी आहे.

bbaahir ddaaDhee garmee Aahe.

outside very hot is

ٻاهر ڏاڍي گرمي آهي.

It is very hot outside.

The suggested set of rules transliterates majority of sentences correctly like in example 3. However there are some ambiguities for the letters that are not

properly present in Devanagari script. For Example the letter ء (‘a) of Perso-Arabic script is sometimes equivalent of अ (a) in Devanagari, while अ (a) is actually equivalent of ا (a) of Perso-Arabic script. Similarly same ء (‘a) is sometimes equivalent to ी (e) (equivalent for Perso-Arabic ي (e)) and sometimes the ء (‘a) is completely omitted to achieve the correct transliteration. As we can see in example 4 the letter अ (a) of first word is wrongly transliterated into letter ا (a) of the Perso-Arabic Script while its correct transliteration would be (‘a) ء.

Example 4.

हूअ डा॒ढी पियारी आहे.
hoo’a ddaaDhee piyaaree aahe.
she very lovely is
هُوا ڏاڍي پيارِي آهي. (incorrect)
هُوءَ ڏاڍي پيارِي آهي. (correct)
She is very lovely.

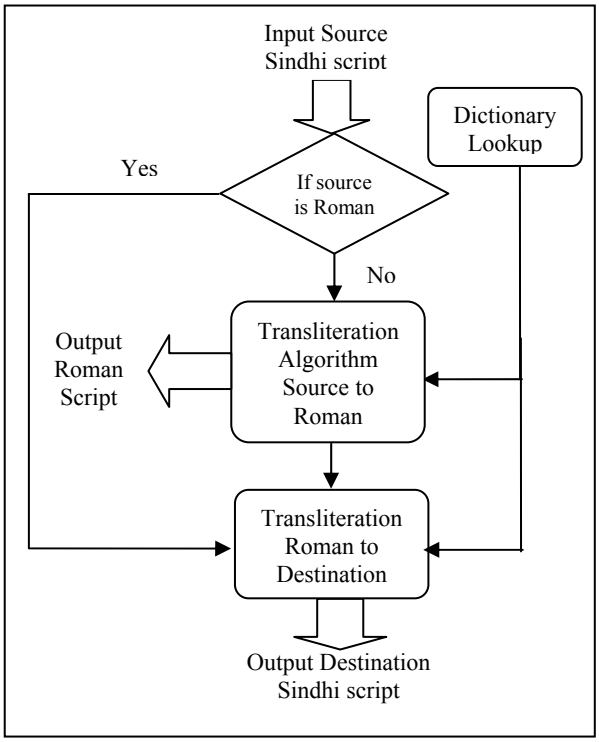

Figure1. Transliteration model for Sindhi scripts.

5. Conclusion and Future Work

After successful implementation of transliteration model discussed, one would be able to transliterate Sindhi from one script to another. People familiar with one script, will be able to understand the writings in other script. It will also be useful in implementing simple Roman to Sindhi (any of two scripts) transliteration. By implementation of successful transliteration design, the transliteration aids like Google Transliteration IME would be able to use Roman to Sindhi mapping, to make it possible to transliterate between Roman and native Sindhi scripts. Automatic transliteration will help to end up the discussions and dispute of Roman script adoption for Sindhi language. The proposed model needs to be checked on large scale by applying the algorithm on a reasonably large corpus.

6. References

[1] Unicode glossary from IBM website. http://www.ibm.com/developerworks/library/glossaries/unicode.html.(Accessed: 2010).

[2] Malik, M G Abbas.. “Punjabi Machine Transliteration”, in proceedings of the 21st *International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, Manchester, UK, July 17 - 21, 2008.

[3] Malik, M. G. Abbas. Boitet, Christian. Bhattcharyya, Pushpak, “Hindi Urdu Machine Transliteration using Finite-state Transducers”, proceedings of *COLING*, Manchester, UK 2008.

[4] Lachman. M. Khubchandani “Current Trends in Linguistics”, pp. 219.

[5] Rahman. M. “Computational Linguistics and Sindhi Language” published in “Sindhi Boli Research Journal” Sindhi Language Authority, Hyderabad 2009.

[6] Sindhi Language Authority Website: http://www.sindhila.org/SindhiLanguage.htm. (Accessed: 2010).

[7] Malik, M. G. Abbas. Boitet, Christian. Bhattcharyya, Pushpak “A Hybrid Model for Urdu Hindi Transliteration” In Proceedings of the *2009 Named Entities Workshop, ACL-IJCNLP 2009*, pp. 177–185, Suntec, Singapore, 2009.

[8] NRSI: Computers & Writing Systems. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&cat_id=Glossary. (Accessed: 2010).

[9] Malik, M. G. Abbas, *Hindi Urdu Machine Transliteration System*, MS thesis, Department of Linguitics, University of Paris 7, Denis Diderot, 2 Place Jussieu, Paris France. 2006.