Punjabi Machine Transliteration

M. G. Abbas Malik
Department of Linguistics
Denis Diderot, University of Paris 7
Paris, France
abbas.malik@gmail.com

Abstract

Machine Transliteration is to transcribe a word written in a script with approximate phonetic equivalence in another language. It is useful for machine translation, cross-lingual information retrieval, multilingual text and speech processing. Punjabi Machine Transliteration (PMT) is a special case of machine transliteration and is a process of converting a word from Shahmukhi (based on Arabic script) to Gurmukhi (derivation of Landa, Shardha and Takri, old scripts of Indian subcontinent), two scripts of Punjabi, irrespective of the type of word.

The Punjabi Machine Transliteration System uses transliteration rules (character mappings and dependency rules) for transliteration of Shahmukhi words into Gurmukhi. The PMT system can transliterate every word written in Shahmukhi.

1 Introduction

Punjabi is the mother tongue of more than 110 million people of Pakistan (66 million), India (44 million) and many millions in America, Canada and Europe. It has been written in two mutually incomprehensible scripts Shahmukhi and Gurmukhi for centuries. Punjabis from Pakistan are unable to comprehend Punjabi written in Gurmukhi and Punjabis from India are unable to comprehend Punjabi written in Shahmukhi. In contrast, they do not have any problem to understand the verbal expression of each other. Punjabi Machine Transliteration (PMT) system is an effort to bridge the written communication gap between the two scripts for the benefit of the millions of Punjabis around the globe.

Transliteration refers to phonetic translation across two languages with different writing systems (Knight & Graehl, 1998), such as Arabic to English (Nasreen & Leah, 2003). Most prior work has been done for Machine Translation (MT) (Knight & Leah, 97; Paola & Sanjeev, 2003; Knight & Stall, 1998) from English to other major languages of the world like Arabic, Chinese, etc. for cross-lingual information retrieval (Pirkola et al, 2003), for the development of multilingual resources (Yan et al, 2003; Kang & Kim, 2000) and for the development of cross-lingual applications.

PMT is a special kind of machine transliteration. It converts a Shahmukhi word into a Gurmukhi word irrespective of the type constraints of the word. It not only preserves the phonetics of the transliterated word but in contrast to usual transliteration, also preserves the meaning.

Two scripts are discussed and compared. Based on this comparison and analysis, character mappings between Shahmukhi and Gurmukhi are drawn and transliteration rules are discussed. Finally, architecture and process of the PMT system are discussed. When it is applied to Punjabi Unicode encoded text especially designed for testing, the results were complied and analyzed. PMT system will provide basis for *Cross-Scriptural Information Retrieval* (CSIR) and *Cross-Scriptural Application Development* (CSAD).

2 Punjabi Machine Transliteration

According to Paola (2003), "When writing a foreign name in one's native language, one tries to preserve the way it sounds, i.e. one uses an orthographic representation which, when read aloud by the native speaker of the language, sounds as it would when spoken by a speaker of the foreign language – a process referred to as Transliteration". Usually, transliteration is referred to phonetic translation of a word of some

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1137–1144,
Sydney, July 2006.

specific type (proper nouns, technical terms, etc) across languages with different writing systems. Native speakers may not understand the meaning of transliterated word.

PMT is a special type of Machine Transliteration in which a word is transliterated across two different writing systems used for the same language. It is independent of the type constraint of the word. It preserves both the phonetics as well as the meaning of transliterated word.

3 Scripts of Punjabi

3.1 Shahmukhi

Shahmukhi derives its character set form the Arabic alphabet. It is a right-to-left script and the shape assumed by a character in a word is context sensitive, i.e. the shape of a character is different depending whether the position of the character is at the beginning, in the middle or at the end of the word. Normally, it is written in Nastalique, a highly complex writing system that is cursive and context-sensitive. A sentence illustrating Shahmukhi is given below:

پنجابی میری ماݨ جوگی ماں بولی اے۔

It has 49 consonants, 16 diacritical marks and 16 vowels, etc. (Malik 2005)

3.2 Gurmukhi

Gurmukhi derives its character set from old scripts of the Indian Sub-continent i.e. Landa (script of North West), Sharda (script of Kashmir) and Takri (script of western Himalaya). It is a left-to-right syllabic script. A sentence illustrating Gurmukhi is given below:

ਪੰਜਾਬੀ ਮੇਰੀ ਮਾਣ ਜੋਗੀ ਮਾਂ ਬੋਲੀ ਏ.

It has 38 consonants, 10 vowels characters, 9 vowel symbols, 2 symbols for nasal sounds and 1 symbol that duplicates the sound of a consonant. (Bhatia 2003, Malik 2005)

4 Analysis and PMT Rules

Punjabi is written in two completely different scripts. One script is right-to-left and the other is left-to-right. One is Arabic based cursive and the other is syllabic. But both of them represent the phonetic repository of Punjabi. These phonetic sounds are used to determine the relation between the characters of two scripts. On the basis of this idea, character mappings are determined.

For the analysis and comparison, both scripts are subdivided into different group on the basis of types of characters e.g. consonants, vowels, diacritical marks, etc.

4.1 Consonant Mapping

Consonants can be further subdivided into two groups:

Aspirated Consonants: There are sixteen aspirated consonants in Punjabi (Malik, 2005). Ten of these aspirated consonants (بھ[b^h], پھ[p^h], تھ[t̪^{h}], ٹھ[ʈ^h], جھ[ʤ^h], چھ[ʧ^h], دھ[d̪^{h}], ڈھ[ɖ^h], کھ[k^h], گھ[g^h]) are very frequently used in Punjabi as compared to the remaining six aspirates (رھ[r^h], ڑھ[ɽ^h], لھ[l^h], مھ[m^h], نھ[n^h], وھ[v^h]). In Shahmukhi, aspirated consonants are represented by the combination of a consonant (to be aspirated) and HEH-DOACHASHMEE (ھ). For example ب [b] + ھ [h] = بھ [b^h] and ج [ʤ] + ھ [h] = جھ [ʤ^h].

In Gurmukhi, each frequently used aspirated-consonant is represented by a unique character. But, less frequent aspirated consonants are represented by the combination of a consonant (to be aspirated) and sub-joined PAIREEN HAAHAA e.g. ਲ [l] + ੍ + ਹ [h] = ਲ੍ਹ (لھ) [l^h] and ਵ [v] + ੍ + ਹ [h] = ਵ੍ਹ (وھ) [v^h], where ੍ is the sub-joiner. The sub-joiner character (੍) tells that the following ਹ [h] is going to change the shape of PAIREEN HAAHHA.

The mapping of ten frequently used aspirated consonants is given in Table 1.

Sr.	Shahmukhi	Gurmukhi	Sr.	Shahmukhi	Gurmukhi
1	بھ [b^h]	ਭ	6	چھ [ʧ^h]	ਛ
2	پھ [p^h]	ਫ	7	دھ [d̪^{h}]	ਧ
3	تھ [t̪^{h}]	ਥ	8	ڈھ [ɖ^h]	ਢ
4	ٹھ [ʈ^h]	ਠ	9	کھ [k^h]	ਖ
5	جھ [ʤ^h]	ਝ	10	گھ [g^h]	ਘ

Table 1: Aspirated Consonants Mapping

The mapping for the remaining six aspirates is covered under non-aspirated consonants.

Non-Aspirated Consonants: In case of non-aspirated consonants, Shahmukhi has more consonants than Gurmukhi, which follows the one symbol for one sound principle. On the other hand there are more then one characters for a single sound in Shahmukhi. For example, Seh

(ث), Seen (س) and Sad (ص) represent [s] and [s] has one equivalent in Gurmukhi i.e. Sassaa (ਸ). Similarly other characters like ਅ [a], ਤ [t̪], ਹ [h] and ਜ਼ [z] have multiple equivalents in Shahmukhi. Non-aspirated consonants mapping is given in Table 2.

Sr.	Shahmukhi	Gurmukhi	Sr.	Shahmukhi	Gurmukhi
1	ب [b]	ਬ	21	ط [t̪]	ਤ
2	پ [p]	ਪ	22	ظ [z]	ਜ਼
3	ت [t̪]	ਤ	23	ع [ʔ]	ਅ
4	ٹ [ʈ]	ਟ	24	غ [ɣ]	ਗ਼
5	ث [s]	ਸ	25	ف [f]	ਫ਼
6	ج [ʤ]	ਜ	26	ق [q]	ਕ਼
7	چ [ʧ]	ਚ	27	ک [k]	ਕ
8	ح [h]	ਹ	28	گ [g]	ਗ
9	خ [x]	ਖ਼	29	ل [l]	ਲ
10	د [d̪]	ਦ	30	ڷ [ɭ]	ਲ਼
11	ڈ [ɖ]	ਡ	31	م [m]	ਮ
12	ذ [z]	ਜ਼	32	ن [n]	ਨ
13	ر [r]	ਰ	33	ݨ [ɳ]	ਣ
14	ڑ [ɽ]	ੜ	35	ں [ŋ]	ਂ
15	ز [z]	ਜ਼	35	و [v]	ਵ
16	ژ [ʒ]	ਜ਼	36	ہ [h]	ਹ
17	س [s]	ਸ	37	ھ [h]	੍ਹ
18	ش [ʃ]	ਸ਼	38	ی [j]	ਯ
19	ص [s]	ਸ	39	ے [j]	ਯ
20	ض [z]	ਜ਼			

Table 2: Non-Aspirated Consonants Mapping

4.2 Vowel Mapping

Punjabi contains ten vowels. In Shahmukhi, these vowels are represented with help of four long vowels (Alef Madda (آ), Alef (ا), Vav (و) and Choti Yeh (ی)) and three short vowels (Arabic Fatha – Zabar (◌َ), Arabic Damma – Pesh (◌ُ) and Arabic Kasra – Zer (◌ِ)). Note that the last two long vowels are also used as consonants. Hamza (ء) is a special character and always comes between two vowel sounds as a place holder. For example, in آسائش [ɑsɑɪʃ] (comfort), Hamza (ء) is separating two vowel sounds Alef (ا) and Zer (◌ِ), in آؤ [ɑo] (come), Hamza (ء) is separating two vowel sounds Alef Madda (آ) [ɑ] and Vav (و) [o], etc. In the first example آسائش [ɑsɑɪʃ] (comfort), Hamza (ء) is separating two vowel sounds Alef (ا) and Zer (◌ِ), but normally Zer (◌ِ) is dropped by common people. So Hamza (ء) is mapped on ਇ [ɪ] when it is followed by a consonant.

In Gurmukhi, vowels are represented by ten independent vowel characters (ਅ, ਆ, ਇ, ਈ, ਉ, ਊ, ਏ, ਐ, ਓ, ਔ) and nine dependent vowel signs (ਾ, ਿ, ੀ, ੁ, ੂ, ੇ, ੈ, ੋ, ੌ). When a vowel sound comes at the start of a word or is independent of some consonant in the middle or end of a word, independent vowels are used; otherwise dependent vowel signs are used. The analysis of vowels is shown in Table 4 and the vowel mapping is given in Table 3.

Sr.	Shahmukhi	Gurmukhi	Sr.	Shahmukhi	Gurmukhi
1	اَ[ə]	ਅ	11	ا[ə]	ਅ,ਾ
2	آ[ɑ]	ਆ	12	◌ِ [ɪ]	ਿ
3	اِ[ɪ]	ਇ	13	◌ِی [i]	ੀ
4	اِی [i]	ਈ	14	◌ُ [ʊ]	ੁ
5	اُ[ʊ]	ਉ	15	◌ُو [u]	ੂ
6	اُو[u]	ਊ	16	ے [e]	ੇ
7	اے [e]	ਏ	17	◌َے [æ]	ੈ
8	اَے [æ]	ਐ	18	و [o]	ੋ
9	او [o]	ਓ	19	َو [ɔ]	ੌ
10	اَو [ɔ]	ਔ	20	ء[ɪ]	ਇ

Table 3: Vowels Mapping

Vowel	Shahmukhi	Gurmukhi	Example
ɑ	Represented by Alef Madda (آ) in the beginning of a word and by Alef (ا) in the middle or at the end of a word.	Represented by ਆ and ਾ	آدمی → ਆਦਮੀ [ɑdmi] (man) جاوݨا → ਜਾਵਣਾ [ʤɑvɳɑ] (go)
ə	Represented by Alef (ا) in the beginning of a word and with Zabar (َ) elsewhere.	Represented by ਅ in the beginning.	اجّ → ਅੱਜ [ɑʤʤ] (today)
e	Represented by the combinations of Alef (ا) and Choti Yeh (ی) in the beginning; a consonant and Choti Yeh (ی) in the middle and a consonant and Baree Yeh (ے) at the end of a word.	Represented by ਏ and ੇ	ایدھر → ਏਧਰ [ed̪ʰər] (here), میرا → ਮੇਰਾ [merɑ] (mine), سارے → ਸਾਰੇ [sɑre] (all)
æ	Represented by the combination of Alef (ا), Zabar (َ) and Choti Yeh (ی) in the beginning; a consonant, Zabar (َ) and Choti Yeh (ی) in the middle and a consonant, Zabar (َ) and Baree Yeh (ے) at the end of a word.	Represented by ਐ and ੈ	اَیہہ → ਐਹ [æh] (this), مَیل → ਮੈਲ [mæl] (dirt), ہَے → ਹੈ [hæ] (is)
ɪ	Represented by the combination of Alef (ا) and Zer (ِ) in the beginning and a consonant and Zer (ِ) in the middle of a word. It never appears at the end of a word.	Represented by ਇ and ਿ	اِکّو → ਇੱਕੋ [ɪkko] (one), بارِش → ਬਾਰਿਸ਼ [bɑrɪsh] (rain)
i	Represented by the combination of Alef (ا), Zer (ِ) and Choti Yeh (ی) in the beginning; a consonant, Zer (ِ) and Choti Yeh (ی) in the middle and a consonant and Choti Yeh (ی) at the end of a word	Represented by ਈ and ੀ	اِیتر → ਈਤਰ [it̪ər] (mean) اَمِیری → ਅਮੀਰੀ [ɑmiri] (richness), پنجابی → ਪੰਜਾਬੀ [pənʤɑbi] (Punjabi)
ʊ	Represented by the combination of Alef (ا) and Pesh (ُ) in the beginning; a consonant and Pesh (ُ) in the middle of a word. It never appears at the end of a word.	Represented by ਉ and ੁ	اُدّھر → ਉੱਧਰ [ʊd̪d̪hr] (there) مُلّ → ਮੁੱਲ [mʊll] (price)
u	Represented by the combination of Alef (ا), Pesh (ُ) and Vav (و) in the beginning, a consonant, Pesh (ُ) and Vav (و) in the middle and at the end of a word.	Represented by ਊ and ੂ	اُردُو → ਉਰਦੂ [ʊrd̪u] صُورت → ਸੂਰਤ [surt̪] (face)
o	Represented by the combination of Alef (ا) and Vav (و) in the beginning; a consonant and Vav (و) in the middle and at the end of a word.	Represented by ਓ and ੋ	اوچھاڑ → ਓਛਾੜ [oʧhɑɽ] (cover), پڑھولا → ਪੜ੍ਹੋਲਾ [pɽholɑ] (a big pot in which wheat is stored)
ɔ	Represented by the combination of Alef (ا), Zabar (َ) and Vav (و) in the beginning; a consonant, Zabar (َ) and Vav (و) in the middle and at the end of a word.	Represented by ਔ and ੌ	اَوڑا → ਔੜਾ [ɔɽɑ] (hindrance), مَوت → ਮੌਤ [mɔt̪] (death)
Note: Where → means 'its equivalent in Gurmukhi is'.			

Table 4: Vowels Analysis of Punjabi for PMT

4.3 Sub-Joins (PAIREEN) of Gurmukhi

There are three PAIREEN (sub-joins) in Gurmukhi, “Haahaa”, “Vaavaa” and “Raaraa” shown in Table 5. For PMT, if HEH-DOACHASHMEE (ھ) does come after the less frequently used aspirated consonants then it is transliterated into PAIREEN Haahaa. Other PAIREENS are very rare in their usage and are used only in Sanskrit loan words. In present day writings, PAIREEN Vaavaa and Raaraa are being replaced by normal Vaavaa (ਵ) and Raaraa (ਰ) respectively.

Sr.	PAIREEN	Shahmukhi	Gurmukhi	English
1	◌੍ਹ	بُلّھ	ਬੁੱਲ੍ਹ	Lips
2	◌੍ਰ	چندرما	ਚੰਦ੍ਰਮਾ	Moon
3	◌੍ਵ	سویمان	ਸ੍ਵੈਮਾਨ	Self-respect

Table 5: Sub-joins (PAIREEN) of Gurmukhi

4.4 Diacritical Marks

Both in Shahmukhi and Gurmukhi, diacritical marks (dependent vowel signs in Gurmukhi) are the back bone of the vowel system and are very important for the correct pronunciation and understanding the meaning of a word. There are sixteen diacritical marks in Shahmukhi and nine dependent vowel sings in Gurmukhi (Malik, 2005). The mapping of diacritical marks is given in Table 6.

Sr.	Shahmukhi	Gurmukhi	Sr.	Shahmukhi	Gurmukhi
1	◌َ [ə]	---	9	◌ٍ [ɪn]	ਿ◌ਨ
2	◌ِ [ɪ]	ਿ◌	10	◌ٗ	◌ੱ
3	◌ُ [ʊ]	◌ੁ	11	◌ٓ	---
4	◌ؕ	---	12	◌ٔ	---
5	◌ً [ən]	ਨ	13	◌ٕ	---
6	◌ٌ [ʊn]	◌ੁਨ	14	◌ٖ	---
7	◌ْ	---	15	◌ٛ	---
8	◌ّ	---	16	◌ٰ [ɑ]	◌ਾ

Table 6: Diacritical Mapping

Diacritical marks in Shahmukhi are very important for the correct pronunciation and understanding the meaning of a word. But they are sparingly used in writing by common people. In the normal text of Shahmukhi books, newspapers, and magazines etc. one will not find the diacritical marks. The pronunciation of a word and its meaning would be comprehended with the help of the context in which it is used.

For example,

اَیہہ سڑک بڑی چوڑی اے۔

میری چوڑی لال اے۔

In the first sentence, the word چوڑی is pronounced as [ʧɔɽi] and it conveys the meaning of ‘wide’. In the second sentence, the word چوڑی is pronounced as [ʧuɽi] and it conveys the meaning of ‘bangle’. There should be Zabar (◌َ) after Cheh (چ) and Pesh (◌ُ) after Cheh (چ) in the first and second words respectively, to remove the ambiguities.

It is clear from the above example that diacritical marks are essential for removing ambiguities, natural language processing and speech synthesis.

4.5 Other Symbols

Punctuation marks in Gurmukhi are the same as in English, except the full stop. DANDA (।) and double DANDA (॥) of Devanagri script are used for the full stop instead. In case of Shahmukhi, these are same as in Arabic. The mapping of digits and punctuation marks is given in Table 7.

Sr.	Shahmukhi	Gurmukhi	Sr.	Shahmukhi	Gurmukhi
1	۰	੦	8	۷	੭
2	۱	੧	9	۸	੮
3	۲	੨	10	۹	੯
4	۳	੩	11	،	,
5	۴	੪	12	؟	?
6	۵	੫	13	؛	;
7	۶	੬	14	۔	।

Table 7: Other Symbols Mapping

4.6 Dependency Rules

Character mappings alone are not sufficient for PMT. They require certain dependency or contextual rules for producing correct transliteration. The basic idea behind these rules is the same as that of the character mappings. These rules include rules for aspirated consonants, non-aspirated consonants, Alef (ا), Alef Madda (آ), Vav (و), Choti Yeh (ی) etc. Only some of these rules are discussed here due to space limitations.

Rules for Consonants: Shahmukhi consonants are transliterated into their equivalent

Gurmukhi consonants e.g. س → ਸ [s]. Any diacritical mark except Shadda (ّ) is ignored at this point and is treated in rules for vowels or in rules for diacritical marks. In Shahmukhi, Shadda (ّ) is placed after the consonant but in Gurmukhi, its equivalent Addak (ੱ) is placed before the consonant e.g. پ + ّ → ੱਪ [pp]. Both Shadda (ّ) and Addak (ੱ) double the sound a consonant after or before which they are placed.

This rule is applicable to all consonants in Table 1 and 2 except Ain (ع), Noon (ن), Noonghunna (ں), Vav (و), Heh Gol (ہ), Dochashmee Heh (ھ), Choti Yeh (ی) and Baree Yeh (ے). These characters are treated separately.

Rule for Hamza (ء): Hamza (ء) is a special character of Shahmukhi. Rules for Hamza (ء) are:

– If Hamza (ء) is followed by Choti Yeh (ی), then Hamza (ء) and Choti Yeh (ی) will be transliterated into ਈ [i].

– If Hamza (ء) is followed by Baree Yeh (ے), then Hamza (ء) and Baree Yeh (ے) will be transliterated into ਏ [e].

– If Hamza (ء) is followed by Zer (ِ), then Hamza (ء) and Zer (ِ) will be transliterated into ਇ [ɪ].

– If Hamza (ء) is followed by Pesh (ُ), then Hamza (ء) and Pesh (ُ) will be transliterated into ਉ [ʊ].

In all other cases, Hamza (ء) will be transliterated into ਇ [ɪ].

5 PMT System

5.1 System Architecture

The architecture of PMT system and its functionality are described in this section. The system architecture of *Punjabi Machine Transliteration System* is shown in figure 1.

Unicode encoded Shahmukhi text input is received by the `Input Text Parser` that parses it into Shahmukhi words by using simple parsing techniques. These words are called Shahmukhi Tokens. Then these tokens are given to the `Transliteration Component`. This component gives each token to the `PMT Token Converter` that converts a Shahmukhi Token into a Gurmukhi Token by using the `PMT Rules Manager`, which consists of character mappings and dependency rules. The `PMT Token Converter` then gives the Gurmukhi Token back to the `Transliteration Component`. When all Shahmukhi Tokens are converted into Gurmukhi Tokens, then all Gurmukhi Tokens are passed to the `Output Text Generator` that generates the output Unicode encoded Gurmukhi text. The main PMT process is done by the `PMT Token Converter` and the `PMT Rules Manager`.

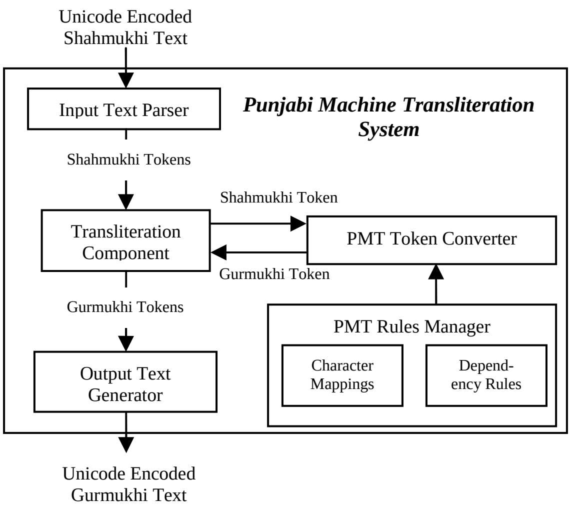

Figure 1: Architecture of PMT System

PMT system is a rule based transliteration system and is very robust. It is fast and accurate in its working. It can be used in domains involving Information Communication Technology (web, WAP, instant messaging, etc.).

5.2 PMT Process

The PMT Process is implemented in the `PMT Token Converter` and the `PMT Rules Manager`. For PMT, each Shahmukhi Token is parsed into its constituent characters and the character dependencies are determined on the basis of the occurrence and the contextual placement of the character in the token. In each Shahmukhi Token, there are some characters that bear dependencies and some characters are independent of such contextual dependencies for transliteration. If the character under consideration bears a dependency, then it is resolved and transliterated with the help of dependency rules.

If the character under consideration does not bear a dependency, then its transliteration is achieved by character mapping. This is done through mapping a character of the Shahmukhi token to its equivalent Gurmukhi character with the help of character mapping tables 1, 2, 3, 6 and 7, whichever is applicable. In this way, a Shahmukhi Token is transliterated into its equivalent Gurmukhi Token.

Consider some input Shahmukhi text S. First it is parsed into Shahmukhi Tokens (S_1, S_2… S_N). Suppose that S_i = “واليياں” [vɑlejɑ̃] is the i^{th} Shahmukhi Token. S_i is parsed into characters Vav (و) [v], Alef (ا) [ɑ], Lam (ل) [l], Choti Yeh (ی) [j], Alef (ا) [ɑ] and Noon Ghunna (ں) [ŋ]. Then PMT mappings and dependency rules are applied to transliterate the Shahmukhi Token into a Gurmukhi Token. The Gurmukhi Token G_i=“ਵਾਲਿਆਂ” is generated from S_i. The step by step process is clearly shown in Table 8.

Sr.	Character(s) Parsed	Gurmukhi Token	Mapping or Rule Applied
1	و → ਵ [v]	ਵ	Mapping Table 4
2	ا → ਾ [ɑ]	ਵਾ	Rule for ALEF
3	ل → ਲ [l]	ਵਾਲ	Mapping Table 4
4	یِا → ਿਆ [ɪɑ]	ਵਾਲਿਆ	Rule for YEH
5	ں → ਂ [ŋ]	ਵਾਲਿਆਂ	Rule for NOONGHUNNA
Note: → is read as ‘is transliterated into’.			

Table 8: Methodology of PMTS

In this way, all Shahmukhi Tokens are transliterated into Gurmukhi Tokens (G_1, G_2 … G_n). From these Gurmukhi Tokens, Gurmukhi text G is generated.

The important point to be noted here is that input Shahmukhi text must contain all necessary diacritical marks, which are necessary for the correct pronunciation and understanding the meaning of the transliterated word.

6 Evaluation Experiments

6.1 Input Selection

The first task for evaluation of the PMT system is the selection of input texts. To consider the historical aspects, two manuscripts, poetry by *Maqbal* (Maqbal) and *Heer* by *Waris Shah* (Waris, 1766) were selected. Geographically Punjab is divided into four parts eastern Punjab (Indian Punjab), central Punjab, southern Punjab and northern Punjab. All these geographical regions represent the major dialects of Punjabi. Hayms of Baba Nanak (eastern Punjab), Heer by Waris Shah (central Punjab), Hayms by Khawaja Farid (southern Punjab) and Saif-ul-Malooq by Mian Muhammad Bakhsh (northern Punjab) were selected for the evaluation of PMT system. All the above selected texts are categorized as classical literature of Punjabi. In modern literature, poetry and short stories of different poets and writers were selected from some issues of Puncham (monthly Punjabi magazine since 1985) and other published books. All of these selected texts were then compiled into Unicode encoded text as none of them were available in this form before.

The main task after the compilation of all the selected texts into Unicode encoded texts is to put all necessary diacritical marks in the text. This is done with help of dictionaries. The accuracy of the PMT system depends upon the necessary diacritical marks. Absence of the necessary diacritical marks affects the accuracy greatly.

6.2 Results

After the compilation of selected input texts, they are transliterated into Gurmukhi texts by using the PMT system. Then the transliterated Gurmukhi texts are tested for errors and accuracy. Testing is done manually with help of dictionaries of Shahmukhi and Gurmukhi by persons who know both scripts. The results are given in Table 9.

Source	Total Words	Accuracy
Manuscripts	1,007	98.21
Baba Nanak	3,918	98.47
Khawaja Farid	2,289	98.25
Waris Shah	14,225	98.95
Mian Muhammad Bakhsh	7,245	98.52
Modern lieratutre	16,736	99.39
Total	**45,420**	**98.95**

Table 9: Results of PMT System

If we look at the results, it is clear that the PMT system gives more than 98% accuracy on classical literature and more than 99% accuracy on the modern literature. So PMT system fulfills the requirement of transliteration across two scripts of Punjabi. The only constraint to achieve this accuracy is that input text must contain all necessary diacritical marks for removing ambiguities.

7 Conclusion

Shahmukhi and Gurmukhi being the only two prevailing scripts for Punjabi expressions encompass a population of almost 110 million around the globe. PMT is an endeavor to bridge the ethnical, cultural and geographical divisions between the Punjabi speaking communities. By implementing this system of transliteration, new horizons for thought, idea and belief will be shared and the world will gain an impetus on the efforts harmonizing relationships between nations. The large repository of historical, literary and religious work done by generations will now be available for easy transformation and critique for all. The research has future milestone enabling PMT system for back machine transliteration from Gurmukhi to Shahmukhi.

Reference

Ari Pirkola, Jarmo Toivonen, Heikki Keskustalo, Kari Visala, and Kalervo Järvelin. 2003. *Fuzzy Translation of Cross-Lingual Spelling Variants.* In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp: 345 – 352

Baba Guru Nanak, arranged by Muhammad Asif Khan. 1998. آکھیا بابا نانک نے (*Sayings of Baba Nanak* in Punjabi Shahmukhi). Pakistan Punjabi Adbi Board, Lahore

Bhatia, Tej K. 2003. *The Gurmukhi Script and Other Writing Systems of Punjab: History, Structure and Identity.* International Symposium on Indic Script: Past and future organized by Research Institute for the Languages and Cultures of Asia and Africa and Tokyo University of Foreign Studies, December 17 – 19. pp: 181 – 213

In-Ho Kang and GilChang Kim. 2000. *English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks.* In Proceedings of the 17th conference on Computational Linguistics. 1: 418 – 424

Khawaja Farid (arranged by Muhammad Asif Khan). آکھیا خواجہ فرید نے (*Sayings of Khawaja Farid* in Punjabi Shahmukhi). Pakistan Punjabi Adbi Board, Lahore

Knight, K. and Stalls, B. G. 1998. *Translating Names and Technical Terms in Arabic Tex.* Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages

Knight, Kevin and Graehl, Jonathan. 1997. *Machine Transliteration.* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. pp. 128-135

Knight, Kevin; Morgan Kaufmann and Graehl, Jonathan. 1998. *Machine Transliteration.* In Computational Linguistics. 24(4): 599 – 612

Malik, M. G. Abbas. 2005. *Towards Unicode Compatible Punjabi Character Set.* In proceedings of 27th Internationalization and Unicode Conference, 6 – 8 April, Berlin, Germany

Maqbal. مدحِ غوث. Punjabi Manuscript in Oriental Section, Main Library University of the Punjab, Quaid-e-Azam Campus, Lahore Pakistan; 7 pages; Access # 8773

Mian Muhammad Bakhsh (Edited by Fareer Muhammad Faqeer). 2000. *Saif-ul-Malooq.* Al-Faisal Pub. Urdu Bazar, Lahore

Nasreen AbdulJaleel, Leah S. Larkey. 2003. *Statistical transliteration for English-Arabic cross language information retrieval.* In Proceedings of the 12th international conference on information and knowledge management. pp: 139 – 146

Paola Virga and Sanjeev Khudanpur. 2003. *Transliteration of proper names in cross-language applications.* In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. pp: 365 – 366

Rahman Tariq. 2004. *Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift.* Crossing the Digital Divide, SCALLA Conference on Computational Linguistics, 5 – 7 January 2004

Sung Young Jung, SungLim Hong and Eunok Peak. 2000. *An English to Korean transliteration model of extended markov window.* In Proceedings of the 17th conference on Computational Linguistics. 1:383 – 389

Tanveer Bukhari. 2000. لغت پنجابی اُردو. Urdu Science Board, 299 Uper Mall, Lahore

Waris Shah. 1766. ہیر رانجھا. Punjabi Manuscript in Oriental Section, Main Library University of the Punjab, Quaid-e-Azam Campus, Lahore Pakistan; 48 pages; Access # [Ui VI 135/]1443

Waris Shah (arranged by Naseem Ijaz). 1977. ہیر رانجھا. Lehran, Punjabi Journal, Lahore

Yan Qu, Gregory Grefenstette, David A. Evans. 2003. *Automatic transliteration for Japanese-to-English text retrieval.* In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. pp: 353 – 360