Shahmukhi to Gurmukhi Transliteration System

Tejinder Singh Saini
ACTDPL, Punjabi University,
Patiala 147 002, India
tej@pbi.ac.in

Gurpreet Singh Lehal
DCS, Punjabi University,
Patiala 147 002, India
gslehal@yahoo.com

Virinder S Kalra
Sociology, SOSS
University of Manchester
kalra@manchester.uk

Abstract

The existence of two scripts for Punjabi language has created a script barrier between the Punjabi literature written in India and Pakistan. This research has developed a new system for the first time of its kind for Shahmukhi text without diacritical marks. The purposed system for Shahmukhi to Gurmukhi transliteration has been implemented with various research techniques based on language corpus. The corpus analysis of both scripts is performed for generating statistical data of different types like character and word frequencies and bi-gram frequencies. This statistical analysis is used in different phases of transliteration. Potentially, all members of the substantial Punjabi community will benefit vastly from this transliteration system.

1 Introduction

One of the great challenges before Information Technology is to overcome language barriers across the whole humanity so that everyone can communicate with everyone else on the planet in real time. South Asia is one of those unique parts of the world where a single language is written in different scripts. This is the case, for example, with Punjabi language, spoken by tens of millions of people, but written in Indian East Punjab (20 million) in Gurmukhi script (a Left to Right script based on Devanagari) and in Pakistani West Punjab (80 million), written in Shahmukhi script (a Right to Left script based on Arabic), and by growing number of Punjabis (2 million) in the EU and the US in the Roman script. Whilst in speech Punjabi spoken in the Eastern and the Western parts is mutually comprehensible, in the written form it is not so. The existence of two scripts for Punjabi has created a script barrier between the Punjabi literature written in India and Pakistan. More than 60 per cent of Punjabi literature of medieval period (500-1450 AD) is available in Shahmukhi script only, while most of the modern Punjabi writings are in Gurmukhi. Potentially, all members of the substantial Punjabi community will benefit vastly from this transliteration system.

1.1 Gurmukhi Script

The Gurmukhi script, derived from the Sharada script and standardised by Guru Angad Dev in the 16th century, was designed to write the Punjabi language. The meaning of "Gurmukhi" is literally “from the mouth of the Guru". The Gurmukhi script has forty one letters, including thirty eight consonants and three basic vowel sign bearers. There are five nasal consonants (ਙ, ਞ, ਣ, ਨ, ਮ) and two additional nasalization signs, bindi ਂ [ɲ] and tippi ਁ [ɲ]. In addition to this, there are nine dependent vowel signs (ੁ[ʊ], ੂ [u], ੋ[o], ਾ[ə], ਿ[ɪ], ੀ[i], ੇ[e], ੈ[æ], ੌ[ɔ]) used to create ten independent vowels with three bearer characters: Ura ੳ[ʊ], Aira ਅ [ə] and Iri ੲ[ɪ].

1.2 Shahmukhi Script

The meaning of "Shahmukhi" is literally “from the King's mouth". Shahmukhi is a local variant of the Urdu script used to record the Punjabi language. It is based on right to left Nastalique style of the Persian and Arabic script. It has thirty seven simple consonants, eleven frequently used aspirated consonants, four long vowels and three short vowel symbols (Malik 2006).

Coling 2008: Companion volume – Posters and Demonstrations, pages 177–180
Manchester, August 2008

2 Comparison with the Existing System

In actual practice, Shahmukhi script is written without short vowels and other diacritical marks. The PMT system discussed by Malik A. (2006) claims 98% accuracy only when the input text

Input text (right to left)
اس گل وچ جدوں اسیں بہتے پنجابیاں نوں ویکھدے ہاں تاں پرنسپل تیجا سنگھ دے لیکھ وچ بیانیاں گئیاں کوڑیاں سچائیاں ہور وی شدت نال محسوس ہندیاں ہین۔ اسیں دیس نوں پیار کرن دا دعویٰ کردے ہاں پر اپنے صوبے نوں وساری بیٹھے ہاں۔ اس دا سبہ توں وڈا ثبوت ایہہ ہے کہ بھارت دے لگ بھگ بہتے صوبے اپنے اپنے ستھاپنا دوس بڑے اتشاہ تے جذبے نال مناؤندے ہین۔ اپنی زبان، اپنے سبھیاچار، اپنے پچھوکڑ تے اپنے ورثے تے مان کردے ہین۔ اپنی قومی پچھان تے مان کردے ہین۔ پر ساڈا بابا آدم ہی نرالا ہے۔ سرکاراں توں لے کے عام لوکاں تک پنجابی صوبے دے بنن دن بارے پوری طرحاں اویسلے ہی رہندے ہین۔
Output-A of PMT system (left to right)
ਅਸ ਗਲ ਵਚ ਜਦੋਂ ***ਅਸੈਂ ਬਹਤੇ ਪਨਜਾਬੇਆਂ ਨੈਂ*** ਵੇਖਦੇ ਹਾਂ ਤਾਂ ***ਪਰਨਸਪਲ*** ਤੇਜਾ ***ਸਨਘ*** ਦੇ ਲੇਖ ***ਵਚ ਬੇਅਨੇਆਂ*** ਗਈਆਂ ***ਕੌੜੇਆਂ*** ਸਚਾਈਆਂ ਹੋਰ ਵੀ ***ਸ਼ਦਤ*** ਨਾਲ ***ਮਹਸੈਸ ਹਨਦੇਆਂ ਹੇਨ।*** ***ਅਸੈਂ*** ਦੇਸ ***ਨੈਂ ਪੇਆਰ*** ਕਰਨ ਦਾ ਦਾਵਾ ਕਰਦੇ ਹਾਂ ਪਰ ***ਅਪਨੇ ਸੋਬੇ ਨੈਂ*** ਵਸਾਰੀ ***ਬੇਠੇ*** ਹਾਂ। ***ਅਸ*** ਦਾ ਸਭ ਤੋਂ ***ਵਡਾ ਸਬੋਤ ਇਹਾ ਹੇ ਕਹ*** ਭਾਰਤ ਦੇ ਲਗ ਭਗ ***ਬਹਤੇ ਸੋਬੇ ਅਪਨੇ ਅਪਨੇ*** ਸਥਾਪਨਾ ***ਦੋਸ*** ਬੜੇ ***ਅਤਸ਼ਾਹ*** ਤੇ ਜਜ਼ਬੇ ਨਾਲ ***ਮਨਾਈਵਨਦੇ ਹੇਨ।*** ***ਅਪਨੀ*** ਜ਼ਬਾਨ, ***ਅਪਨੇ ਸਭੇਅਚਾਰ, ਅਪਨੇ ਪਛੋਕੜ*** ਤੇ ***ਅਪਨੇ ਵਰਸੇ*** ਤੇ ***ਮਾਨ*** ਕਰਦੇ ***ਹੇਨ।*** ***ਅਪਨੀ ਕੋਮੀ ਪਛਾਨ*** ਤੇ ***ਮਾਨ*** ਕਰਦੇ ***ਹੇਨ।*** ਪਰ ਸਾਡਾ ਬਾਬਾ ਆਦਮ ਹੀ ***ਨਰਾਲਾ ਹੇ।*** ਸਰਕਾਰਾਂ ਤੋਂ ***ਲੇ*** ਕੇ ਆਮ ਲੋਕਾਂ ***ਤਕ ਪਨਜਾਬੀ ਸੋਬੇ*** ਦੇ ***ਬਨਨ ਦਨ*** ਬਾਰੇ ***ਪੋਰੀ ਤਰਹਾਂ ਉਓਂਸਲੇ*** ਹੀ ***ਰਹਨਦੇ ਹੇਨ।***
Output-B of proposed system (left to right)
ਇਸ ਗੱਲ ਵਿਚ ਜਦੋਂ ਅਸੀਂ ਬਹੁਤੇ ਪੰਜਾਬੀਆਂ ਨੂੰ ਵੇਖਦੇ ਹਾਂ ਤਾਂ ਪ੍ਰਿੰਸੀਪਲ ਤੇਜਾ ਸਿੰਘ ਦੇ ਲੇਖ ਵਿਚ ਬਿਆਨੀਆਂ ਗਈਆਂ ਕੌੜੀਆਂ ਸਚਾਈਆਂ ਹੋਰ ਵੀ ਸ਼ਿੱਦਤ ਨਾਲ ਮਹਿਸੂਸ ਹੁੰਦੀਆਂ ਹੈਨ। ਅਸੀਂ ਦੇਸ ਨੂੰ ਪਿਆਰ ਕਰਨ ਦਾ ਦਾਵਾ ਕਰਦੇ ਹਾਂ ਪਰ ਆਪਣੇ ਸੂਬੇ ਨੂੰ ***ਵਸਾਰੀ*** ਬੈਠੇ ਹਾਂ। ਇਸ ਦਾ ਸਭ ਤੋਂ ਵੱਡਾ ਸਬੂਤ ਇਹ ਹੈ ਕਿ ਭਾਰਤ ਦੇ ਲਗ ਭਗ ਬਹੁਤੇ ਸੂਬੇ ਆਪਣੇ ਆਪਣੇ ਸਥਾਪਨਾ ***ਦੋਸ*** ਬੜੇ ਉਤਸ਼ਾਹ ਤੇ ਜਜ਼ਬੇ ਨਾਲ ਮਨਾਉਂਦੇ ਹੈਨ। ਆਪਣੀ ਜ਼ਬਾਨ, ਆਪਣੇ ਸਭਿਆਚਾਰ, ਆਪਣੇ ਪਿਛੋਕੜ ਤੇ ਆਪਣੇ ਵਿਰਸੇ ਤੇ ਮਾਣ ਕਰਦੇ ਹੈਨ। ਆਪਣੀ ਕੌਮੀ ਪਛਾਣ ਤੇ ਮਾਣ ਕਰਦੇ ਹੈਨ। ਪਰ ਸਾਡਾ ਬਾਬਾ ਆਦਮ ਹੀ ਨਿਰਾਲਾ ਹੈ। ਸਰਕਾਰਾਂ ਤੋਂ ਲੈ ਕੇ ਆਮ ਲੋਕਾਂ ਤੱਕ ਪੰਜਾਬੀ ਸੂਬੇ ਦੇ ਬਣਨ ਦਿਨ ਬਾਰੇ ਪੂਰੀ ਤਰ੍ਹਾਂ ਅਵੈਸਲੇ ਹੀ ਰਹਿੰਦੇ ਹੈਨ।

Table 1. I/O of PMT and Proposed Systems

Output Type	Transliteration Tokens			Accuracy %
	Total	Wrong	Right	
A	116	64	52	44.8275
B	116	02	114	98.2758

Table 2. Comparison of Output-A & B

has all necessary diacritical marks for removing ambiguities. But this process of putting missing diacritical marks is not practically possible due to many reasons like large input size, manual intervention, person having knowledge of both the scripts and so on. We have manually evaluated PMT system against the following Shahmukhi input published on a web site and the output text is shown as output-A in Table 1.The output of proposed system on the same input is shown as output-B. The wrong transliteration of Gurmukhi tokens is shown in bold and italic and the comparison of both outputs is shown in Table 2.Clearly, our system is more practical in nature than PMT and we got good transliteration with different inputs having missing diacritical marks.

3 The Complexity

The Shahmukhi script has many complexities by its nature and the major two of them are:

3.1 Recognition of Shahmukhi Text without Diacritical Marks

Shahmukhi script is usually written without short vowels and other diacritical marks, often leading to potential ambiguity. Arabic orthography does not provide full vocalization of the text, and the reader is expected to infer short vowels from the context of the sentence. In the written Shahmukhi script it is not mandatory to put short vowels below or above the Shahmukhi character to clear its sound. These special signs are called "Aerab" in Urdu. It is a big challenge in the process of machine transliteration to recognize the right word from the written text.

3.2 Multiple Mappings

It is observed that there is multiple possible mapping in Gurmukhi script corresponding to a single character in the Shahmukhi script as shown in Table 3.

Name	Shahmukhi Character	Gurmukhi Mapping
Vav	و [v]	ਵ [v], ੋ [o], ੌ [ɔ], ੁ [ʊ], ੂ [u], ਓ [o]
Yeh	ی[j]	ਯ [j], ਿ [ɪ], ੇ [e], ੈ[æ], ੀ[i], ਈ [i]

Table 3. Multiple Mapping into Gurmukhi Script

4 Transliteration System

The transliteration system as shown in figure 1 is virtually divided into two phases. The first phase performs pre-processing on the input Shahmukhi token by performing dictionary lookup. If the dictionary lookup fails then the token will go for rule based transliteration and ultimately this phase will generate best possible Gurmukhi token(s). The second phase performs

the task of post-processing. Unicode Alignment component performs context analysis of input Gurmukhi token(s). All Forms generator (AFG) component will perform critical task of handling missing diacritical marks. This component will suggest similar possible forms of a Gurmukhi token which is not most frequent one. The queue manager of post-processing phase is designed to work on bi-gram language model. This will select the best possible unigram for final output by consulting bi-gram weights of the current token with its neighboring tokens

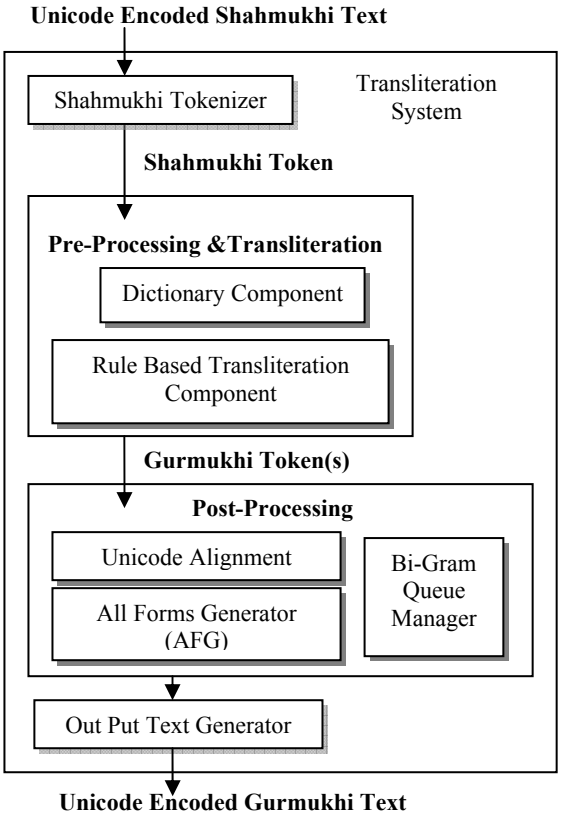

Figure 1. System Overview

5 Lexical Resources Used

Shahmukhi Corpus: 3.3 million words.
Gurmukhi Corpus: 7 million words.
Shahmukhi-Gurmukhi Dictionary
Unigram and Bi-gram Table
All Forms Generator (AFG)

6 Example

Here we show the internal working of the system through an example. Suppose we observe a Shahmukhi string as shown in figure 2. First, we pass this through the pre-processing and transliteration phase where the input string has been tokenized into eleven Shahmukhi tokens. Every input token has been searched in the dictionary for their existence. This status result is shown in table 4 where the tokens 1st, 2nd, 4th, 5th, 6th, 7th, 8th, 10th and 11th are found in dictionary and their intermediate Weighted Gurmukhi Forms (WGF) have been generated. These tokens directly jump to bi-gram queue manager for bi-gram analysis in post-processing phase.

Input 11 Shahmukhi tokens (Right to Left)

Transliterated 11 Gurmukhi tokens (Left to Right)

Figure 2. Shahmukhi Gurmukhi Tokens

Token	Shahmukhi Token	Found in Dictionary	WGF token{weight}
1	فیر	Yes	ਫੇਰ{4513}; ਫਿਰ{8714}
2	ہور	Yes	ਹੋਰ{14054}; ਹੋਰ{18}
3	حیرانی	No	ਹੈਰਾਨੀ{524}
4	ایس	Yes	ਇਸ{59998}; ਏਸ{1186}
5	گلوں	Yes	ਗੱਲੋਂ{107}
6	ہوندی	Yes	ਹੁੰਦੀ{7699}
7	اے	Yes	ਏ{7927}; ਐ{3600}
8	جے	Yes	ਜੇ{295}; ਜੇ{9791}
9	اجی	No	ਅਜਹੀ{4}
10	گل	Yes	ਗੱਲ{447};ਗਲ{47};ਗਿੱਲ{9} ਗੁਲ{5};ਗੁੱਲ{5}
11	کرن	Yes	ਕਰਨ{21582};ਕਰਣ{174}; ਕਿਰਨ{159}

Table 4. Pre-Processing Transliteration Status

On the other hand, the input tokens 3rd and 9th are not found in dictionary. Therefore, in this phase they will pass through transliteration component and then in post-processing phase they will pass through Unicode formatting. After that they will test for Most Frequent (MF) check by comparing their weights with a predefined threshold value[2]

[2] Threshold value is minimum probability of occurrence among most frequent tokens in target script corpus.

(100 in this case). As shown in table 5 the WGF of 3rd token ਹੈਰਾਨੀ{524} is most frequent one and will move to bi-gram queue whereas the WGF of 9th token ਅਜਹੀ{4} is not a most frequent token and will reach at bi-gram queue manager only after passing through all forms generator (AFG).

Token	MF	AFG Status	Bi-gram Found	Output
1	-	-	hold ਫੇਰ;ਫਿਰ	-
2	-	-	ਫੇਰ-ਹੋਰ,12; ਫਿਰ-ਹੋਰ, 20;	ਫਿਰ
3	Yes	-	hold ਹੋਰ	ਹੋਰ
4	-	-	ਹੈਰਾਨੀ-ਇਸ,10;	ਹੈਰਾਨੀ
5	-	-	hold ਇਸ	ਇਸ
6	-	-	ਇਸ-ਗੱਲੋਂ,45;	ਗੱਲੋਂ
7	-	-	ਹੁੰਦੀ-ਏ,86; ਹੁੰਦੀ-ਐ,125;	ਹੁੰਦੀ
8	-	-	hold ਐ	ਐ
9	No	ਅਜੇਹੀ{310} ਅਜਿਹੀ{1486} ਅਜਹੀ{4}	ਐ-ਜੇ,22; ਜੇ-ਅਜਿਹੀ,13;	ਜੇ
18	Yes	-	hold ਅਜਿਹੀ	ਅਜਿਹੀ
11	-	-	ਅਜਿਹੀ-ਗੱਲ,38; ਗੱਲ-ਕਰਨ,179; ਗਲ-ਕਰਨ,18;	ਗੱਲ
	EOS	-	hold ਕਰਨ	ਕਰਨ

Table 5. Post-Processing Status and output

Here, we see that the AFG has generated two additional forms ਅਜੇਹੀ{310} ਅਜਿਹੀ{1486} (table 5) for this token. These new forms are having additional diacritical marks of short vowels those are missing in the original form. Clearly, AFG has supplied the best possible forms. Next, we show how bi-gram manager will work on WGF tokens to generate final Gurmukhi token. In this model the next token will decide the selection of its previous one. Consider the case of second WFG token ਹੋਰ{14054} having bi-gram combinations with previous one as ਫੇਰ-ਹੋਰ with weight 12 and ਫਿਰ-ਹੋਰ with weight 20. Clearly, the token ਫਿਰ will produce as output not ਫੇਰ because ਫਿਰ-ਹੋਰ combination has higher weight than ਫੇਰ-ਹੋਰ. Similarly, this table shows found bi-gram weights and correspondingly decided Gurmukhi token as output.

7 Results and Discussion

The transliteration system was tested on a small set of poetry, article and story. The results are tabulated in Table 6.

As we can observe an average transliteration accuracy of 91.37% has been obtained. We got good transliteration with different inputs. The main source of error is the existence of vowel-consonant mapping between the two scripts. The Shahmukhi vowel characters Vav(و) and Yeh(ى) have mapping into Gurmukhi consonants Vava(ਵ) and Ya(ਯ) respectively. This kind of vowel-consonant mapping can not be resolved fully with dependency rules but can be minimized by refining the dictionary and phonetic code generation rules of AFG component. In other cases, system makes errors showing deficiency in handling those tokens which are not belonging to common vocabulary domain.

Type	Transliterated Tokens	Accuracy %
Poetry	3,301	90.63769
Article	584	92.60274
Story	3,981	90.88043
Total	**7,866**	**91.37362**

Table 6. Transliteration Results

8 References

Arbabi, Mansur, Scott M. Fischthal, Vincent C. Cheng and Elizabeth Bar. 1994. Algorithms for Arabic name transliteration. IBM Journal of research and Development, pp 183-193.

Haizhou Li, Min Zhang and Jian Su. 2004. A Joint Source-Channel Model for Machine Transliteration. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp 159-166.

Malik, M. G. Abbas. 2006. Punjabi Machine Transliteration. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp 1137-1144.

Y. Gal, 2002. An HMM Approach to Vowel Restoration in Arabic and Hebrew. Proceedings of ACL Workshop on Computational Approaches to Semitic Languages, pp 27-33.

Youngim Jung, Donghun Lee, Aesun Yoon, Hyuk Chul Kwon. 2004. Transliteration System for Arabic-Numeral Expressions using Decision Tree for Intelligent Korean TTS, volume 1. 30th Annual Conference of IEEE, pp 657-662.