A Gurmukhi to Shahmukhi Transliteration System

Gurpreet Singh Lehal
Department of Computer Science
Punjabi University, Patiala, India-147002.
gslehal@gmail.com

Abstract

Transliteration is a process wherein an input string in some alphabet is converted to a string in another alphabet, usually based on the phonetics of the original word. It is useful for Machine Translation, to create possible equivalents of unknown words, cross-lingual information retrieval systems and in development of multilingual resources. Punjabi is one of the unique languages, which are written in more than one script. In India, Punjabi is written in Gurmukhi script, while in Pakistan it is written in Shahmukhi (Urdu) script. This has created a script wedge as majority of Punjabi speaking people in Pakistan cannot read Gurmukhi script, and similarly the majority of Punjabi speaking people in India cannot comprehend Shahmukhi script. In this paper, we present a high accuracy Gurmukhi to Shahmukhi transliteration system. We have tried to overcome the shortcomings of the existing Gurmukhi to Shahmukhi transliteration systems and developed a system which can which can transliterate any Gurmukhi text to Shahmukhi at more than 98.6% accuracy at word level.

Introduction

Punjabi is one of the most widely spoken language in Indian sub-continent, with more than 100 million speakers in India and Pakistan. A unique feature of Punjabi is that it is written in two mutually incomprehensible scripts. In India Punjabi language is written in Gurmukhi script, while in Pakistan it is written in Shahmukhi (Urdu) script. This has created a script wedge as majority of Punjabi speaking people in Pakistan cannot read Gurmukhi script, and similarly the majority of Punjabi speaking people in India cannot comprehend Shahmukhi script. To break this script barrier, we have developed a transliteration system for transliterating Punjabi text written in Gurmukhi script to Shahmukhi.

Transliteration is a process wherein an input string in some alphabet is converted to a string in another alphabet, usually based on the phonetics of the original word. Transliteration is frequently used in Machine Translation, to create possible equivalents of unknown words, cross-lingual information retrieval systems and in development of multilingual resources.

Transliteration is usually classified into two directions. Given a pair (s,t) where s is the original word in the source language and t is the transliterated word in the target language, forward transliteration is the process of phonetically converting s into t, and backward transliteration is the process of correctly generating s given t(Lin and Chen, 2002). Backward transliteration is more challenging than forward transliteration. While forward transliteration can accomplish the mapping through table-lookup, backward transliteration is required to disambiguate the noise produced in the forward transliteration and estimate the original word as close as possible. The Gurmukhi to Shahmukhi transliteration is different from standard transliteration schemes in the sense that the Gurmukhi word has to be converted to Shahmukhi with exact spellings, since the language is written in both Gurmukhi and Shahmukhi. This fact has largely been ignored by existing systems, which have treated the transliteration problem as forward transliteration and used character mappings and dependency rules to convert Gurmukhi word to Shahmukhi without any consideration to spellings.

In this paper we present a fairly good accuracy Gurmukhi-Shahmukhi transliteration system, where we have taken special care to retain the correct spellings in the transliterated Shahmukhi text.

Gurmukhi and Shahmukhi scripts: a brief overview

Shahmukhi is a right-to-left script and the shape assumed by a character in a word is context sensitive. The Nastalique script, a cursive, right-to-left context-sensitive and a highly complex writing system is normally used for Shahmukhi orthography. It has 35 simple consonants, 15 aspirated consonants, one character for nasal

"Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009"

sound, 15 diacritical marks, 10 digits and other symbols. Gurmukhi derives its character set from old scripts of the Indian Sub-continent. It is a left-to-right syllabic script. 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 (Malik, 2006; Malik, 2005; Jawaid and Ahmed, 2009). Below is the mapping chart for Gurmukhi and its respective Shahmukhi character(s).

Table 1. Gurmukhi to Shahmukhi Mapping

Gurmukhi	Shahmukhi
ਅ	ا
ਆ	آ
ਇ	اِ
ਈ	ای
ਉ	اُ
ਊ	اُو
ਏ	اے
ਐ	اَے
ਓ	اَو
ਔ	اَو
ਕ	ک/ق
ਖ	کھ
ਗ	گ
ਘ	گھ
ਙ	نگ
ਚ	چ
ਛ	چھ
ਜ	ج
ਝ	جھ
ਞ	نج
ਟ	ٹ
ਠ	ٹھ
ਡ	ڈ
ਢ	ڈھ
ਣ	ن
ਤ	ط / ت
ਥ	تھ
ਦ	د
ਧ	دھ
ਨ	ن
ਪ	پ
ਫ	پھ
ਬ	ب
ਭ	بھ
ਮ	م
ਯ	ے
ਰ	ر
ਲ	ل
ਲ਼	ل
ਵ	و
ਸ਼	ش
ਸ	ث / ص / س
ਹ	ح / ہ
ਾ	ا / ہ / ◌ٰ
ਿ	◌ِ
ੀ	ی
◌ੁ	◌ُ
◌ੂ	◌ُو
◌ੇ	ے
◌ੈ	◌َے
◌ੋ	و
◌ੌ	◌َو
ਖ਼	خ
ਗ਼	غ
ਜ਼	ذ / ظ / ض / ز
ੜ	ڑ
ਫ਼	ف
◌ੰ	ں / ن
◌ੱ	◌ّ
◌ਂ	ں / ن

As can be seen from Table 1, there are certain Gurmukhi characters which have multiple equivalent Shahmukhi characters. For example, the character ਤ is mapped to ت and ط. Similarly, the character ਜ਼ is mapped to four Shahmukhi characters (ض/ظ/ذ/ز).

A statistical analysis on the Shahmukhi corpus was performed to determine the percentage of times the Gurmukhi character was mapped to the similar sounding Shahmukhi characters (Table 2). The Shahmukhi characters with highest frequency are selected for default mapping.

Table 2. Frequency of Occurrence of multiple mapping characters

Gurmukhi Character	Equivalent Shahmukhi Characters	% Frequency	Default
ਹ	ہ ح	92.12% 7.88%	ہ
ਸ	س ص ث	92.58% 7.41% 1.39%	س
ਕ	ک ق	91.29% 8.71%	ک
ਤ	ت ط	95.49% 4.51%	ت
ਜ਼	ز ض ظ ذ	62.12% 14.87% 13.91% 9.10%	ز

Issues in Gurmukhi-Shahmukhi Transliteration

As already discussed above, the main issue in Gurmukhi-Shahmukhi transliteration is to convert the Gurmukhi word to Shahmukhi with correct spellings. The existing systems for Gurmukhi-Shahmukhi transliteration are mostly rule based and thus do not have good transliteration accuracy as they do not check for correct Shahmukhi spellings. Two transliteration systems available on internet have been taken for comparison with our system (http://www.puran.info/PMT/PMT.aspx; http://parc.cdac.in/utrans.htm). The main challenges involved in development of a good accuracy system are:

1. **Multiple Mappings :** As already discussed in above section, there are certain Gurmukhi characters which are mapped to multiple Shahmukhi characters. The five such Gurmukhi characters are shown in Table 2. Two more characters ੰ and ਂ are mapped to both ں and ن . But these characters do not pose any problem since the character ں always come at the end of the word. For other cases we choose the character having higher frequency of occurrence. But one obvious problem is that the lesser frequent characters never get mapped. A statistical analysis of corpus reveals that these lesser frequently occurring similar sounding Shahmukhi characters have 1.45% frequency of occurrence. It was also found that the average word length of a Shahmukhi word is 3.55, which means that we can roughly predict that any rule based system which does not take care of these characters, will have 1.45*3.55 = 5.15% error rate at word level, due to multiple mappings. Thus to have a high accuracy system, it is necessary to solve this problem.
2. **No exact equivalent mappings in Gurmukhi for some Shahmukhi characters:** There are certain Shahmukhi characters such as ع (ain) and ء (hamza) for which there no exact equivalent Gurmukhi characters. Though *hamza* can be generated using character combination rules, but there are no specific rules for *ain*. As for example, consider the following cases:

   ਔਰਤ -> عورت

   ਸ਼ੁਰੂ -> شُرُوع

   ਰਫ਼ੀ -> رفیع

   ਅਲਵਿਦਾ -> الوداع

   It is not possible to specify the rules in above examples for generation of *ain* in Shahmukhi words and it depends from case to case. From the statistics, we find that *ain* has 0.42% of frequency of occurrence and thus any pure rule based system, which cannot properly generate *ain* in Shahmukhi will further have 0.42*3.55 = 1.49% error rate at word level.
3. **Missing nukta symbols in Gurmukhi text** : Punjabi language has borrowed many words from Arabic, Persian etc. Five consonants (ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼) were added to the original 35 characters in Gurmukhi to accommodate the sounds from these languages. As can be seen, these characters were created by adding the nukta(dot) symbol at the feet of the

existing symbols (ਸ ਖ ਗ ਜ ਫ). But over the years, the usage of these characters particularly, ਖ਼ ਗ਼ ਜ਼ ਫ਼, has been on decline as many Punjabi speakers do not make a distinction between ਖ ਖ਼, ਗ ਗ਼ and ਫ ਫ਼ . In fact from the analysis of Gurmukhi and Shahmukhi corpus, we found that these four characters had combined character frequency of occurrence of only 0.17% as compared to 1.35% for their counterparts in Shahmukhi. The result is that most of the words in Gurmukhi are now written without nukta symbol. The symbol ਸ਼ is an exception. As an example, we found that the word ਫ਼ਕੀਰ occurs 69 times in the Gurmukhi corpus while the word ਫਕੀਰ occurs 147 times in the Gurmukhi corpus. It is interesting to see that even though ਫ਼ਕੀਰ is the correct spelling but still the word ਫਕੀਰ has higher frequency of occurence.The reader can easily make out the word even if it is not written without nukta, but if that word is converted to Shahmukhi using character to character based mapping it results in wrong spellings. So ਫਕੀਰ will be transliterated to پھکِیر in Shahmukhi, which is wrong while the actual transliteration is فقیر , which is obtained if the correct Gurmukhi spellings ਫ਼ਕੀਰ are used. This puts a constraint on the system that either the user should supply the correct Gurmukhi spellings for nukta characters, or the system should automatically correct such errors if proper spellings in Shahmukhi are to be generated.

4. **Difference between pronunciation and orthography:** In certain cases, the Gurmukhi words are written with short vowels, while they are pronounced with long vowels. The equivalent words in Shahmukhi are also written with long vowels and so the rule based mapping system which converts those short vowels in Gurmukhi Shahmukhi give wrong results in such cases. Some examples of such words are ਗੁਰੂ, ਬਿਮਾਰ and ਖ਼ੁਰਾਕ. They are pronounced as ਗੂਰੂ ਬੀਮਾਰ ਖ਼ੂਰਾਕ but written with short vowels, while the corresponding words in Shahmukhi are written with long vowels as خوراکـگورو بیمار respectively. For transliteration of such words, we have to device special methods to handle these cases.
5. **Transliteration of proper nouns:** The transliteration of Urdu proper nouns such as names of persons and places from Gurmukhi to Shahmukhi poses another challenge. Many times the spellings of such words in Shahmukhi are typical and it is not possible to formulate transliteration rules for generation of such spellings. As for example consider the Gurmukhi words ਅਬਦੁੱਲਾ ਬੁਸ਼ਰਾ ਰਹਿਮਾਨ and ਹੈਦਰਾਬਾਦ, which are written in Shahmukhi are عبداللہ بُشریٰ رحمٰن حیدرآباد but which will yield wrong results if transliterated using the usual rule based mapping system.

Gurmukhi-Shahmukhi Transliteration System Architecture

The system architecture of the Gurmukhi-Shahmukhi transliteration developed by us is displayed in Fig. 1. The system has been to take care of the issues raised in the last section. As can be seen from Fig. 1, the complete system is divided into 3 stages: pre-processing, processing and post-processing.

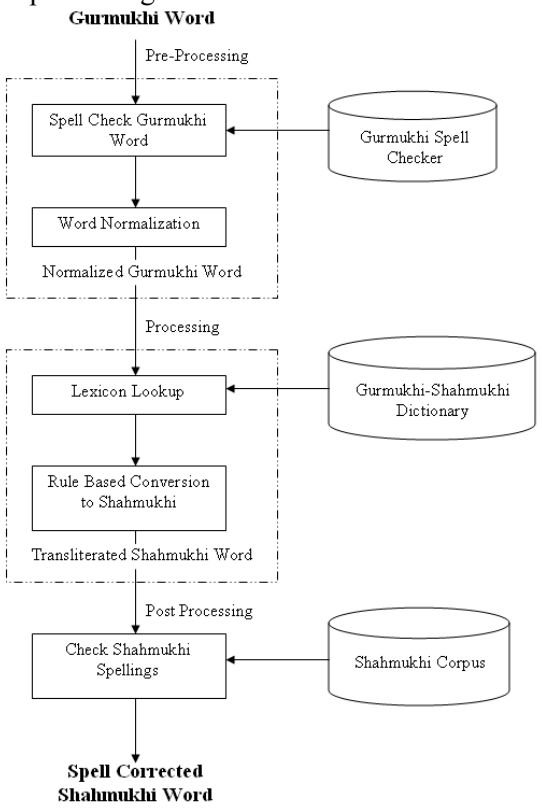

Figure 1: System architecture of Gurmukhi-Shahmukhi Transliteration System

In the pre-processing stage, the Gurmukhi word is cleaned and prepared for transliteration by normalizing the Gurmukhi word according to the Shahmukhi spellings and pronunciation. In the processing stage, the normalized Gurmukhi word is converted to Shahmukhi using a rule based system. The proper names and words with typical spellings are transliterated using a Gurmukhi-Shahmukhi lookup dictionary. In the post-processing stage, the spellings of the transliterated Shahmukhi word are corrected by using a Shahmukhi corpus.

Lexical Resources Used

The following lexical resources have been used:

Resource	**Details**
Gurmukhi spell checker	Root words : 41,253
Gurmukhi-Shahmukhi dictionary	Terms: 10,254
Shahmukhi Corpus	Total words :97,63,294 Unique words:1,93,679

The details of the three stages are:

1. **Pre-Processing :** The input to this stage is a Gurmukhi word in Unicode format. As already discussed above, the Gurmukhi word may be wrongly spelled because of the *nukta* related characters. These spelling errors may not appear so serious for Gurmukhi readers, but when transliterated as such they generate wrong Shahmukhi spellings. To solve this problem, the Gurmukhi word is sent to a Gurmukhi spell checker, which automatically corrects the *nukta* related errors. Thus the word ਗਜਲ will get converted to ਗ਼ਜ਼ਲ after being fed to the Gurmukhi spell checker. To handle the problem of variation in pronunciation and orthography of some Gurmukhi words, we have created a database of all such words along with their inflections. The Gurmukhi word is checked in this database and if found gets converted to the appropriate form. Thus the word ਗੁਰੂਆਂ gets converted to ਗੁਰੂਆਂ. At the end of this stage the Gurmukhi word correcetd for the spellings and normalised for the pronunciation based errors is generated. Thus for the input word ਖੁਸ਼ੀ , the output word will be ਖ਼ੁਸ਼ੀ after spell check and normalization.
2. **Processing :** In this stage the normalized word generated in pre-processing stage is converted to Shahmukhi by using letter to letter conversion mapping rules using the mapping Table 1 and Table 2. The Gurmukhi word is broken into its constituent Unicode characters and each Gurmukhi character is replaced by corresponding Shahmukhi character. For multiple mapping, the default character, with highest frequency of occurrence as mentioned in Table 2 is selected.

   As for example, the word ਕਮਰੇ will be converted to Shahmukhi as follows:

   ਕ + ਮ + ਰ + ੇ -> ک + م + ر + ے = کمرے

   If two vowels in Gurmukhi come together, then the character *hamza* is placed in between them in Shahmukhi.

   As for example for ਕੋਈ ,

   ਕੋ + ਈ -> کو+ئی and

   Similarly for ਆਓ

   ਆ + ਓ -> آ +ئو

   Besides, these simple mapping rules, some special pronunciation based rules have also been developed. It is not possible to lay down all the rules, but some of these rules are:

   ਇ + ਆ -> یا (ਲਾਇਆ ->لایا)

   ਿ + ਓ -> یو (ਵਾਲਿਓ -> والیو)

   ◌ੰ + ਪ -> مپ (ਪੰਪ -> پمپ)

   ◌ੰ + ਨ -> نّ (ਸੁੰਨ ->سُنّ)

   Though the above rules work fine for most of the common words, but they do not give proper results in the following cases:

   1. In case of multiple mappings of a Gurmukhi character to Shahmukhi characters. These rules may not map the correct Shahmukhi character.
   2. There are certain Shahmukhi characters like ع(*ain*) and *khadi zabar* for which there no exact equivalent Gurmukhi characters. The above rules will not be able to automatically generate these characters in final Shahmukhi output.

3. These rules many times fail to transliterate Urdu proper nouns
The multiple mapping problem is handled in the post-processing stage, while for handling the zero mappings and transliterating proper nouns and other typical spellings, we create a separate database of Gurmukhi-Shahmukhi words. Some sample Gurmukhi words with typical Shahmukhi spellings stored in the database are :

ਫ਼ਤਵਾ	فتویٰ
ਖ਼ਸੂਸਨ	خصوصاً
ਸ਼ੁਰੂ	شروع
ਜ਼ਿਲ੍ਹਾ	ضلع
ਮੌਕਾ	موقع
ਇੱਤਲਾਅ	اطلاع

The Gurmukhi word to be transliterated is first searched in this database. If the word is found, then it is directly converted to Shahmukhi else it is converted using the above rule based mapping.

3. **Post - Processing:** This stage is primarily used to correct the spellings of the Shahmukhi words generated in the processing stage. The major source of spellings errors in the transliterated Shahmukhi words is the multiple character mapping in Shahmukhi. As for example the words ਸਲਾਹ, ਮਜ਼ਬੂਤ, ਕਿਸਮ and ਮਤਲਬ will be transliterated as سلاہ, کِسم مزبوت، and متلب while the actual spellings are صلاح, مضبوط, قسمٖ and مطلب respectively. To automatically correct the spellings, we have used a Shahmukhi corpus. For the Gurmukhi characters with multiple Shahmukhi mappings, word forms using all the possible mappings are generated and the word with the highest frequency of occurrence in the Shahmukhi corpus is selected. As for example, consider the word ਤਾਕਤ. Both the characters ਤ and ਕ have multiple Shahmukhi mappings. The Shahmukhi word generated in the processing stage is تاکت. From this word, all its forms are generated as follows:
طاقط طاکط طاکت طاقت تاقط تاکط تاقت تاکت .
A search for each of these words in the Shahmukhi corpus reveals that while the word طاقت has 2045 occurences, none of the other form has a single occurrence. Thus the word ਤਾਕਤ is transliterated to طاقت.

To illustrate how the Gurmukhi word is transformed in each of the three stages, we take the following sample sentence in Gurmukhi. The words which get modified in the next stage are highlighted in **bold**.

ਪੁਲਿਸ ਨੂੰ ਮੂਸਾ ਖਾਨ ਬਿਮਾਰ ਸਿਹਤ ਅਤੇ ਜਖਮੀ ਹਾਲਤ ਵਿਚ ਮਿਲਿਆ

After Gurmukhi spell checking in pre-processing stage, the sentence becomes

ਪੁਲਿਸ ਨੂੰ ਮੂਸਾ **ਖ਼ਾਨ** ਬਿਮਾਰ ਸਿਹਤ ਅਤੇ **ਜ਼ਖ਼ਮੀ** ਹਾਲਤ ਵਿਚ ਮਿਲਿਆ

The text after normalization becomes

ਪੂਲੀਸ ਨੂੰ ਮੂਸਾ ਖ਼ਾਨ **ਬੀਮਾਰ** ਸਿਹਤ ਅਤੇ ਜ਼ਖ਼ਮੀ ਹਾਲਤ ਵਿਚ ਮਿਲਿਆ

This normalized text is sent to the processing stage. In the processing stage, first each of the word is checked for typical spellings in the Gurmukhi-Shahmukhi dictionary. If found then they are directly changed to Shahmukhi, else transliterated using mapping rules.

Since the word ਮੂਸਾ has typical spellings the output after dictionary lookup is :

ਪੂਲੀਸ ਨੂੰ ਮُوسیٰ ਖ਼ਾਨ ਬੀਮਾਰ ਸਿਹਤ ਅਤੇ ਜ਼ਖ਼ਮੀ ਹਾਲਤ ਵਿਚ ਮਿਲਿਆ

The output from the processing stage after rule based transliteration is :

پولیس نوں مُوسیٰ **خان بیمار** سِہت **اتے زخمی** ہالت **وِچ مِلی**

The final output after running the spell checker in the post processing stage is :

پولیس نوں مُوسیٰ خان بیمار **صِحت** اَتے زخمی **حالت** وِچ مِلی

The outputs we got from the other existing systems are below. The wrongly transliterated words are highlighted in red colour.

پُلِس نُون مُوسا کھان بِمار سِہت اتے جکھمی ہالت وِچ مِلیا
http://www.puran.info/PMT/PMT.aspx

پُلِس نُوں مُوسا کھان بِمار سیھت اتے جکھمي ھالت وِچ مِلیا
http://parc.cdac.in/utrans.htm

Another sample sentence when transliterated by the three systems gave the following output:

Gurmukhi	ਸ਼ਾਹ ਆਲਮ ਕੈਂਪ ਵਿਚ ਦਿਨ ਤਾਂ ਕਿਸੇ ਨਾ ਕਿਸੇ ਤਰ੍ਹਾਂ ਗੁਜ਼ਰ ਜਾਂਦੇ ਹਨ ਪਰ ਰਾਤਾਂ ਕਿਆਮਤ ਦੀਆਂ ਹੁੰਦੀਆਂ ਹਨ।
Our system	شاہ عالم کیمپ وچ دن تاں کسے نہ کسے طرحاں گزر جاندے ہن پر راتاں قیامت دیاں ہندیاں ہن۔
Puran	ساہ آلم گینپ وِچ دِن تاں کِسے نا کِسے ترہ٘اں گُجر جاندے ہن پر راتاں کیامت دِیاں ہُندِیاں ہن ۔
Utrans	شاہ الم گینپ وِچ دِن تاں کِسے نا کِسے ترھاں گُزر جاندے ھن پر راتاں کیامت دیاں ھُندیاں ھن ۔

Experimental Results

We have tested our system on 121 pages of text compiled from newspapers, books and poetry. The results are compared with Puran(http://www.puran.info/PMT/PMT.aspx) and Utrans(http://parc.cdac.in/utrans.htm), the two transliteration systems available on the net (Table 3).

Table 3. Transliteration Accuracy

System	Transliteration Accuracy
Utrans	83.45%
Puran	84.92%
Our System	**98.6%**

We observed that the main of sources of improvements in the transliteration accuracy over the existing systems have been:

1. Pre-processing stage, wherein the wrong Gurmukhi spellings are corrected and spellings of some of the words are modified according to their pronunciation.
2. Development of transliteration rules for special cases
3. Usage of Gurmukhi-Shahmukhi dictionary
4. Correction of Shahmukhi spellings with the help of a Shahmukhi corpus.

All these have been implemented for the first time in a Gurmukhi-Shahmukhi transliteration and have resulted in the development of a good accuracy transliteration system.

Failure Cases

The system fails for the following cases:

1. Words with typical spellings, which are not present in Gurmukhi-Shahmukhi dictionary and Shahmukhi corpus.
2. Gurmukhi words with multiple spellings in Shahmukhi. As for example the Gurmukhi word ਕਤਰਾ can get mapped to both قطرہ and کترا. Similarly the word ਅਰਬ has two forms in Shahmukhi عرب and ارب . The correct spellings can be selected after context analysis only.

Conclusion

In this paper we have presented a Gurmukhi to Shahmukhi Transliteration system with around 98.6% word accuracy. We have tried to overcome the shortcomings of the existing rule based Gurmukhi to Shahmukhi Transliteration systems. The various challenges such as multiple/zero character mappings, variations in pronunciations and orthography and transliteration of proper nouns etc. have been handled by generating special rules and using various lexical resources such Gurmukhi spell checker, Shahmukhi corpus, Gurmukhi-Shahmukhi transliteration dictionary etc.

The author will like to acknowledge the web support provided by Tejinder Singh Saini and linguistic support provided by Abdur Rashid, Anwar Chirag and Mohammad Sadiq.

Reference

Wei-Hao Lin and Hsin-His Chen, *"Backward Machine Transliteration by Learning Phonetic Similarity"*, proceedings of the 6th conference on Natural language learning, - Volume 20 , pp. 1-7 (2002).

M.G.A. Malik, *"Punjabi Machine transliteration"*, Proceedings of the 21st International Conference on Computational Linguistics, pp. 1137–1144 (2006).

M.G.A. Malik, *"Towards Unicode Compatible Punjabi Character Set"*, proceedings of 27th Internationalization and Unicode Conference, Berlin, Germany (2005).

Bushra Jawaid, and Tafseer Ahmed, *"Hindi to Urdu Conversion: Beyond Simple Transliteration"*, Proceedings of the Conference on Language & Technology, Lahore, pp.24-31 (2009).

http://www.puran.info/PMT/PMT.aspx

http://parc.cdac.in/utrans.htm