Conversion between Scripts of Punjabi: Beyond Simple Transliteration

Gurpreet Singh LEHAL[1] *Tejinder Singh SAINI*[2]
(1) DCS, Punjabi University, Patiala
(2) ACTDPL, Punjabi University, Patiala

gslehal@gmail.com, tej@pbi.ac.in

ABSTRACT

This paper describes statistical techniques used for modelling transliteration systems between the scripts of Punjabi language. 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 (Perso-Arabic) script. Shahmukhi script has its origin in the ancient Phoenician script whereas Gurmukhi script has its origin in the ancient Brahmi script. Whilst in speech Punjabi spoken in the Eastern and the Western parts is mutually comprehensible, in the written form it is not so. 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 an advanced and highly accurate transliteration system between Gurmukhi and Shahmukhi scripts of Punjabi language which addresses various challenges such as multiple/zero character mappings, missing vowels, word segmentation, variations in pronunciations and orthography and transliteration of proper nouns etc. by generating efficient algorithms along with special rules and using various lexical resources such as Gurmukhi spell checker, corpora of both scripts, Gurmukhi-Shahmukhi transliteration dictionaries, statistical language models etc. The proposed system attains more than 98.6% accuracy at word level while transliterating Gurmukhi text to Shahmukhi. The reverse part i.e. transliterating from Shahmukhi text to Gurmukhi is more complex and challenging but our system has achieved 97% accuracy at word level in this part too.

KEYWORDS: n-gram language model, Shahmukhi, Gurmukhi, Punjabi, Machine Transliteration, Word disambiguation, HMM

Proceedings of COLING 2012: Posters, pages 633–642,
COLING 2012, Mumbai, December 2012.

1 Introduction

There are more that six thousand living languages in the world and some languages are written in different scripts in different regions of the world. The multitude of foreign languages and mutually incomprehensible scripts of the same language pose a barrier to information exchange. Incidentally, the existence of Shahmukhi and Gurmukhi scripts for Punjabi has created a script barrier between the Punjabi literature written in India and in Pakistan. Notably, 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 available in both scripts. Hence, a machine transliteration system that overcomes script barriers is needed to handle these Punjabi scripts with different origins, different direction of writings, different set of alphabet, and different kind of writing system conventions. Already some work in this direction has been reported by Malik, 2006; Saini and Lehal, 2008; Saini et al., 2008 and Lehal, 2009.

2 Transliteration Issues with Punjabi Scripts

- **Missing Short Vowels in Shahmukhi Script:** Most Semitic languages in both ancient and contemporary times are usually written without short vowels and other diacritic marks, often leading to potential ambiguity (Nelken and Shieber, 2005). Similarly, in the written Shahmukhi script, it is not mandatory to put short vowels. In our findings, Shahmukhi corpus has just 1.66% coverage of short vowels ُ[ʊ] (0.81415%), ِ[ɪ] (0.7295%), and ◌ (0.1234%) whereas the equivalent ਿ[ɪ] (4.5462%) and ੁ[ʊ] (1.5844%) in Gurmukhi corpus has 6.13% usage. This leads to potential ambiguous transliteration from Shahmukhi to Gurmukhi script.
- **Multiple Mappings:** It is observed that there are multiple possible mappings between the two scripts. The Shahmukhi characters Vav و [v], Yeh ی[j] and noon ن[n] have shown vowel-vowel, vowel-consonant and consonant-consonant mapping in Gurmukhi script. On the other hand, Gurmukhi characters ਹ[h], ਸ[s], ਕ[k], ੜ[ɽ] and ਜ਼[z] have multiple similar sounding character in Shahmukhi.
- **Missing Script Maps:** There are many characters or symbols in the Shahmukhi script, corresponding to which there are no characters in Gurmukhi, e.g. Hamza ء [ɪ], Do-Zabar ً [ən], Do-Zer ٍ [ɪn], Aen ع[ʔ] etc.
- **Word Boundary Issues**: Like Urdu, Shahmukhi is written in Nastalique style. Due to Nastalique style and irregular use of space, Shahmukhi word segmentation has both space omission and space insertion problems (Durrani and Hussain, 2010; Lehal, 2009, 2010). The space within a word is used more as a tool to control the correct letter shaping rather than to consistently separate words and many times the user omits word boundary space between the consecutive Shahmukhi words when the first word ends with a non-joiner character.
- **Shahmukhi Word with Izafat Form:** There are many compound words or combinations of Shahmukhi words written as a multi-word expression in Gurmukhi script e.g. وزیرِاعظم, ਵਜ਼ੀਰ-ਏ-ਆਜ਼ਮ/vazīr-ē-āzam/; قتلِ عام, ਕਤਲ-ਏ-ਆਮ/katal-ē-ām/.
- **Foreign or Complex Spelling Words:** Shahmukhi words including foreign words have typical spellings such as اسکول, ਸਕੂਲ /sakūl/; اسٹوڈیو, ਸਟੂਡਿਓ /saṭūḍiō/; انویسٹمنٹ,

ਇਨਵੈਸਟਮੈਂਟ /invaisṭamaiṇṭ/; جماعت, ਜਮਾਤ /Jamāt/; وکتی, ਵਿਅਕਤੀ/ Viaktī/; عبدالله, ਅਬਦੁੱਲਾ /abdullā/; رحمٰن, ਰਹਿਮਾਨ /rahimān/ etc.

- **Wrong Spellings due to Missing Gurmukhi Nukta Sign:** In order to accommodate foreign words from Urdu and Persian domain, five consonants (ਸ, ਖ, ਗ, ਜ, ਫ) of Gurmukhi alphabet are extended to ਸ਼[ʃ], ਖ਼[x], ਗ਼[ɣ], ਜ਼[z], ਫ਼[f] with Gurmukhi sign Nukta (pairin bindi). But over the years, the usage of these characters particularly, ਖ਼, ਗ਼, ਜ਼, and ਫ਼ has been on the decline as many Punjabi speakers do not make a distinction between ਖ ਖ਼, ਗ ਗ਼ and ਫ ਫ਼. The result is that most of the words in Gurmukhi are now written without nukta symbol. The symbol ਸ਼ is an exception. When this word is converted to Shahmukhi using character to character based mapping it results in wrong spellings.
- **Difference between Pronunciation and Orthography:** In certain cases, the Gurmukhi words are written with short vowels e.g. ਗੁਰੂ/gurū/, while they are pronounced with long vowels as ਗੂਰੂ /gūrū/. The equivalent words in Shahmukhi are also written with long vowels گورو /gūrū/. Therefore, simple rule based transliteration of such words resulting in wrong transliteration.
- **Ambiguity at word level:** There are many Shahmukhi words which map to multiple Gurmukhi words e.g. گل (ਗੱਲ /gall/, ਗਿੱਲ /gill/, ਗੁੱਲ /gull/, ਗੁਲ /gul/); تک (ਤਕ /tak/, ਤੱਕ /takk/, ਤੁਕ /tuk/) etc. Similarly, Gurmukhi word ਅਰਬ /arab/ has two Shahmukhi spellings with different senses as عرب (Arabia; native of Arabia) and ارب (one billion).

3 Punjabi Machine Transliteration System

The architecture of the Punjabi machine transliteration system is shown in Figure 1.

3.1 Rule-based Transliteration Model

Using the direct method, we have followed manual Consonant-Vowel (CV) approach for character alignments between the source and target scripts.

Dependency Rule for Shahmukhi	Gurmukhi	Example
Alef-Madda آ[ɑ] Vav with hamza و [o] at the beginning	ਆਊ	آؤٹ → ਆਊਟ (āūṭ)
Alef Madda آ[ɑ] followed by Vav و[o] at the beginning	ਆਵ	آواز → ਆਵਾਜ਼ (āvāz)
Alef ا[ɘ] followed by hamza ء [ɪ] and Choti Yeh ی[i] and Alef ا[ɘ] and Noongunna ں [ɲ]	◌ਾਈਆਂ	ودھائیاں → ਵਧਾਈਆਂ (vadhāīāṃ)

TABLE 1– Sample of some dependency rules for Shahmukhi characters

After that context dependent transformation rules are generated to resolve zero or multiple mappings into the target script (see Table 1). Similarly, special pronunciation based rules have

been developed for Gurmukhi characters while transliterating to Shahmukhi as shown in Table 2.

Char1		Char2		Shahmukhi	Example		
ਇ [e]	+	ਆ[a]	→	یا	ਲਾਇਆ /lāiā/	→	یالا
ਿ [ɪ]	+	ਓ[o]	→	یو	ਵਾਲਿਓ /vāliō/	→	یوداں
ਂ [ɲ]	+	ਪ[p]	→	مپ	ਪੰਪ /pamp/	→	پمپ

TABLE 2 – Sample of some Pronunciation based Mapping Rules

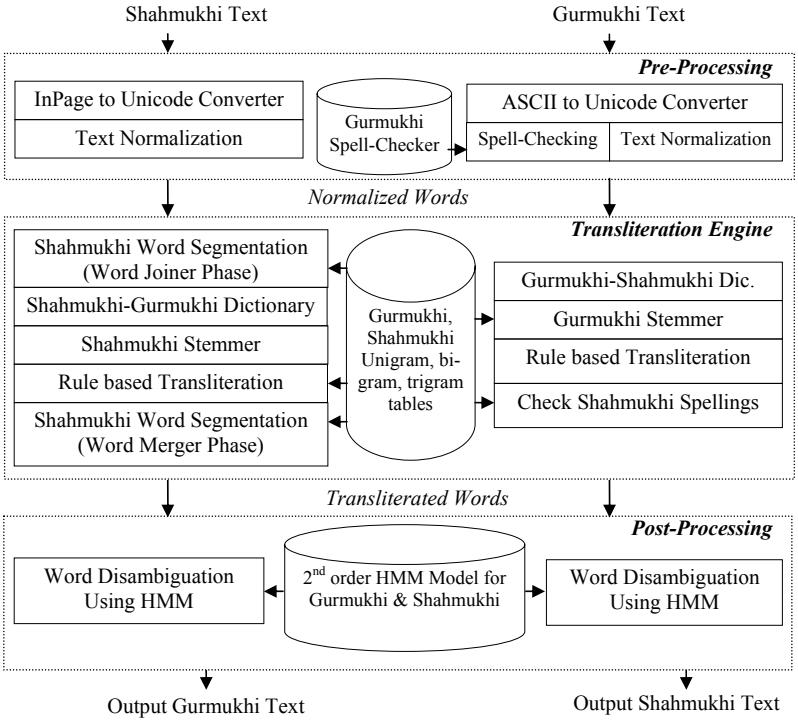

FIGURE 1– System Architecture

3.2 Transliteration using Lexical Resources

3.2.1 Pre-Processing

In the pre-processing stage input text is transformed into Unicode, cleaned and prepared for transliteration in the following manner:

Unicode Conversion: Shahmukhi text in InPage file and Gurmukhi text in traditional fonts is converted into Unicode.

Gurmukhi Spell Checker: Gurmukhi Spell-Checker is used to correct missing Gurmukhi nukta sign problem in Gurmukhi text as discussed earlier.

Text Normalization: The text normalization rules for input Shahmukhi text are formulated with reference to the Urdu Normalization Utility v1.0. **(**2009**)**. Like Urdu, the normalization of Shahmukhi characters is required for visually indistinguishable glyphs that have a different, but canonically equivalent, code point representation in Unicode character set. On the other hand, to overcome the pronunciation and orthographical differences, we normalize the Gurmukhi word by changing its orthography according to the Shahmukhi spellings and pronunciation after Gurmukhi spell-checking.

3.2.2 Transliteration Engine

Shahmukhi word Segmentation: As discussed by Lehal and Saini, (2011), the proposed transliteration model handles both types of word boundary issues at different phases. The first phase of transliteration handles space insertion problem and the space omission problem is addressed at the final phase of transliteration engine. On the other hand, Gurmukhi script is not affected with any segmentation problem.

Dictionary based Transliteration: A one to one Shahmukhi-to-Gurmukhi and Gurmukhi-to-Shahmukhi dictionary of the most frequent words are developed to speed up the transliteration process as well as to handle words with complex spellings as discussed earlier. In addition to this a special Shahmukhi-to-Gurmukhi bi-gram parallel resource is also developed for handling words with Izafat form (compound word) in Shahmukhi.

Light weight Stemmer for Punjabi Language: The size of any lexical resource is limited. It could happen as at times, though inflection may not be present in the respective script dictionary but its root word maybe present. In order to use this idea, we use a light weight stemmer to obtain the root word. Therefore, in our case, stemming is primarily a process of suffix removal. A list of common suffixes has been created. We have taken only the most common Gurmukhi and Shahmukhi suffixes such as ਕੋਲ, ਓਂ, ਿਓਂ, ੀਓਂ, ੇ etc and ے‘یں‘اوں‘وں etc.

Finally, rule-based transliteration is used for transliterating the input words that are not fruitfully processed by these developed lexical resources of the transliteration engine. We have proposed the following algorithm for character-level ambiguity and supplying missing short vowels.

Algorithm for Handling Short Vowels and Character-level Ambiguity: While transforming the Shahmukhi word token into Gurmukhi equivalent in the rule-based transliteration phase, we have proposed the following algorithm.
Step1: Convert Shahmukhi word to Gurmukhi by using predefined character mapping with dependency or contextual rules.
Step2: Format Gurmukhi word according to Unicode formatting like ਅ + ਾ → ਆ, ਅ + ੈ→ਐ and ਅ + ੇ → ਐ, ੳ+ੁ→ ੋ, ੳ + ੂ → ੌ, ੳ +ੇ → ੳ etc.
Step3: In the converted and formatted Gurmukhi word, at each valid character location, insert short vowels and generate unigram weighted list of all possible combinations.
Step4: Select the word with highest weight of occurrence.

For example, consider the Shahmukhi word سنگھ /saṅgh/ transliterated as ਸਿੰਘ /siṅgh/

Input characters:	س[s]	ن[n]	گھ[g^h]
Character mapping:	ਸ	ਨ | ◌ੰ | ਣ	ਘ
Supply short vowels:	ਸ | ਸੁ | ਸਿ	ਨ | ਨੁ | ਨਿ | ◌ੰ | ਣ | ਣੁ | ਣਿ	ਘ | ਘੁ | ਘਿ
Weighted list:	ਸਨਘ(0), ਸਿਨਘ(0), ਸੁਨਘ(0), ਸਨਿਘ(0), ਸਨੁਘ(0), ਸਿਨੁਘ(0), ਸੰਘ(547), ਸੁੰਘ(45), ਸਿੰਘ(55,338), ਸਣਘ(0),ਸਣਿਘ(0), ਸਣੁਘ(0), ਸਿਣੁਘ(0), ਸਣਘ(0) etc.		
Valid Unigrams:	ਸੰਘ(547), ਸੁੰਘ(45), **ਸਿੰਘ(55,338)**[most frequent]		

Similar approach is applied for handling the Gurmukhi characters with multiple Shahmukhi mappings. For example, consider the Gurmukhi word ਸਾਹਿਬ. It has two ambiguous character ਸ[s] → {س|ص|ٹ} and ਹ[h] → {ہ | ح}. The system will generate all the possible forms and then choose the most frequent صاحب (6432) unigram as output.

3.2.3 Post-processing

The word level ambiguity is still present in the transliteration output generated by transliteration engine. The ambiguous Shahmukhi word /mall/ل with missing diacritics has four valid Gurmukhi interpretations ਮੁੱਲ/mull/, ਮਿਲ/mil/, ਮਿੱਲ/mill/, and ਮੱਲ/mall/ within different contexts. On the other hand, the transliteration of Gurmukhi word ਹਾਲ has two Shahmukhi spellings with different senses as حال (state, condition, circumstance) and ہال (Hall; big room). But correct spellings can be selected after context analysis only. At the outset, all we have is the raw corpora for each script of Punjabi language. We have modelled 2nd order HMM for word level ambiguity as proposed by Thede and Harper (1999) for part of speech tagging. Rather than using fixed smoothing technique, they have discussed their new method of calculating contextual probabilities using the linear interpolation. The formula to estimate contextual probability $P(\tau_p = w_k \mid \tau_{p-1} = w_j, \tau_{p-2} = w_i)$ is:

$$P = k_3 . \frac{N_3}{C_2} + (1 - k_3) k_2 . \frac{N_2}{C_1} + (1 - k_3)(1 - k_2) . \frac{N_1}{C_0} \qquad (1)$$

where

$$k_3 = \frac{\log_2 (N_3 + 1) + 1}{\log_2 (N_3 + 1) + 2}; \qquad k_2 = \frac{\log_2 (N_2 + 1) + 1}{\log_2 (N_2 + 1) + 2}$$

N_3	Freq. of trigram $w_i w_j w_k$	C_2	Occurrence of bi-gram $w_i w_j$
N_2	Freq. of bi-gram $w_j w_k$ in corpus	C_1	Occurrence of unigram w_j
N_1	Freq. of unigram w_k in corpus	C_0	Total vocabulary

The disambiguation of ambiguous words ਹਾਲ and ਅਰਬ is performed using 2nd order HMM and output results are shown in Table 3. On the other hand, the HMM disambiguation for Gurmukhi word ambiguity is shown in Table 4.

Sr.	Before WSD	Ambiguity	After WSD
1	فیصل مختار حال وکھے میٹنگ دا اکھی ویکھیا حال	حال | ہال	فیصل مختار ہال وکھے میٹنگ دا اکھی ویکھیا حال
2	اس سال عرب دیشاں نال آپسی وپار چار عرب تک پھیلیا	عرب | ارب	اس سال عرب دیشاں نال آپسی وپار چار ارب تک پھیلیا

TABLE 3 – Shahmukhi Word Sense Disambiguation (WSD) using HMM

Sr.	Input Shahmukhi Text	Ambiguity	After WSD
1	تحصیل ترن تارن	{ਤੁਰਨ, ਤਰਨ}	ਤਹਿਸੀਲ **ਤਰਨ** ਤਾਰਨ
2	لوک اس طرحاں اس دی گرفت وچ	{ਉਸ, ਇਸ}	ਲੋਕ ਇਸ ਤਰਾਂ ਉਸ ਦੀ ਗ੍ਰਿਫ਼ਤ ਵਿਚ

TABLE 4 – Gurmukhi Word Sense Disambiguation (WSD) using HMM

4 Evaluation and Results

4.1 Step-by-Step Evaluation of Shahmukhi-to-Gurmukhi System

A set of ten examples from various online and offline sources are collected for step-by-step evaluation of the system stages. The size of each example ranges from 94 to 246 words per example and the total size of this collection is 1,422 words. The transliteration output from each evaluation stage of the system is manually evaluated. The transliteration steps and Accuracy of the system in the various evaluation stages are shown in Table 5.

Transliteration Steps	Evaluation Stages				
	1st	2nd	3rd	4th	5th
Rule-based approach	☑	☑	☑	☑	☑
Dictionary		☑	☑	☑	☑
Handling missing vowels and char ambiguity			☑	☑	☑
Word segmentation + Light weight Stemmer				☑	☑
Word disambiguation using HMM					☑
Transliteration Accuracy (%)	**47.63**	**87.69**	**92.44**	**95.46**	**97.04**

TABLE 5 – Step-by-Step Evaluation and System Accuracy

4.2 Step-by-Step Evaluation of Gurmukhi-to-Shahmukhi System

A set of eight examples are collected for step-by-step evaluation of the system stages. The size of this collection is 906 words. The transliteration steps and system accuracy with improvement are shown in Table 6 and Figure 4 respectively.

Transliteration Steps	Evaluation Stages		
	1st	2nd	3rd
Rule-based approach	☑	☑	☑
Dictionary + Light weight Stemmer + char ambiguity		☑	☑
Word disambiguation using HMM			☑
Transliteration Accuracy (%)	**75.42**	**97.46**	**98.03**

TABLE 6 – Step-by-Step Evaluation and System Accuracy

4.3 System Evaluation

Shahmukhi-to-Gurmukhi: The natural sources of Shahmukhi text are very limited. With this limitation we have identified the available online and offline sources and three different test sets are taken from different domains. The data Set-1 is a Shahmukhi book of having 37,620 words. The Set-2 consist of online articles, stories and current issues form www.likhari.org having total size of 39,714 words and the Set-3 is a collection news, articles, stories, novels, poetry etc. published on www.wichaar.com and having total size of 46,678 words. The output of the

system is manually evaluated by the person having the knowledge of both the scripts and has Punjabi language as a mother tongue. After manual evaluation the word accuracy is calculated as shown in Table 7. The overall transliteration accuracy of the system is fairly high at 97%. Amongst datasets, the word accuracy for the Set-3 (wichaar.com) is less than Set-2 (likhari.org) which in turn is less that the Set-1 (book). One contributory reason might be that the Pakistani dialect of Punjabi language is frequently used by the writers of wichaar.com. Another possible reason may be the diversity within the dataset.

Table 7 shows an average occurrence of 0.67% words marked as out-of-vocabulary (OOV) by the system. We call them OOV because while transliterating such words our system fails to identify them in any form and the output produced by the system is produced by a hybrid system based on rule-based conversion and a tri-gram character language model. We observed that these types of words mostly include words not present in system corpus, wrong input and foreign words mostly from English or Urdu domain. After manual evaluation of the OOV words with correct input, the average word level transliteration accuracy is calculated as 63.04% as shown in Table 7.

Test Data	Total Words	Found	OOV	Accuracy (Found)	Accuracy (OOV)
Set-1 (book)	37,620	99.468%	0.532%	98.49%	50.00%
Set-2 (likhari.org)	39,714	98.927%	1.073%	96.64%	50.00%
Set-3 (wichaar.com)	46,678	99.595%	0.405%	95.68%	87.5%
Total	**1,24,012**	**99.33%**	**0.67%**	**96.94%**	**63.04%**

TABLE 7– Word Accuracy with Test Data

Gurmukhi-to-Shahmukhi: We have tested our system on more that 100 pages of text compiled from newspapers, books and poetry. The overall transliteration accuracy of this system is 98.6% at word level, which is quite high and actually more then its reverse system. The major source of errors are typical and multiple spellings in Shahmukhi. The accuracy of this word disambiguation task is highly dependent on the training corpus. The accuracy of this system can be increased further by increasing the size of the training corpus and having plentiful of data covering maximum senses of all ambiguous words in the target script.

5 Conclusion

The paper proposes a transliteration system model between the scripts of Punjabi language and incorporates various challenges which were hitherto not dealt with by existing rule based system. The paper describes the proposed high accuracy Gurmukhi-to-Shahmukhi transliteration system which can transliterate any Gurmukhi text to Shahmukhi at more than 98.6% accuracy at word level. Both the systems are complex and challenging. The proposed Shahmukhi-to-Gurmukhi transliteration system has more than 97% accuracy at word level. The various challenges such as multiple/zero character mappings, missing vowels, word segmentation, variations in pronunciations and orthography and transliteration of proper nouns etc. have been handled by generating efficient algorithms along with special rules and using various lexical resources such as Gurmukhi spell checker, corpora of both scripts, Gurmukhi-Shahmukhi transliteration dictionaries.

References

Al-Onaizan, Y. and Knight, K. (2002). Machine transliteration of names in Arabic text. In *Proceedings of the ACL workshop on Computational approaches to Semitic languages*, pages 1–13, Philadelphia, PA.

Ananthakrishnan, S., Narayanan, S. and Bangalore, S. (2005). Automatic diacritization of Arabic transcripts for automatic speech recognition. In *Proceedings of ICON-05,* Kanpur, India.

Durrani, N. and Hussain, S. (2010). Urdu Word Segmentation. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL*, pages 528–536, Los Angeles, California.

Jelinek, F. and Mercer, R.L. (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland.

Katz, S.M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP*, 35(3): 400-401.

Lehal, G. S. (2009). A Two Stage Word Segmentation System for Handling Space Insertion Problem in Urdu Script, In *Proceedings of World Academy of Science, Engineering and Technology*, pages 321-324, Bangkok, Thailand.

Lehal, G. S. (2009). A Gurmukhi to Shahmukhi Transliteration System. In *Proceedings of ICON-2009: 7th International Conference on Natural Language Processing*, pages 167-173, Hyderabad, India.

Lehal, G. S. (2010). A Word Segmentation System for Handling Space Omission Problem in Urdu Script, In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) and 23rd COLIN*, pages 43–50, Beijing.

Lehal, G. S. and Saini, T. S. (2011). A Transliteration based Word Segmentation System for Shahmukhi Script. In *Proceedings of ICISIL*. Springer, Communication in Computer and Information Science, CCIS-139, pages 136-143, India.

Malik, M.G.A. (2006). Punjabi Machine Transliteration. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 1137-1144.

Naseem, T. and Hussain, S. (2007). Spelling Error Trends in Urdu. In *Proceedings of Conference on Language Technology (CLT07)*, University of Peshawar, Pakistan.

Nelken, R. and Shieber, S. M. (2005). Arabic Diacritization Using Weighted Finite-State Transducers. In *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages,* pages 79-86, Ann Arbor, Michigan.

Oh, J.-H., Choi, K.-S. and Isahara, H. (2006). A comparison of different Machine Transliteration Models. *Journal of Artificial Intelligence Research*, 27:119-151.

Saini, T. S., Lehal, G. S. and Kalra, V. S. (2008). Shahmukhi to Gurmukhi Transliteration System. In *Proceedings of 22nd international Conference on Computational Linguistics (Coling)*, pages 177-180, Manchester, UK.

Saini, T. S. and Lehal, G. S. (2008). Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach. *Research in Computing Science*, 33:151-162, Mexico.

Thede, S.M. and Harper, M.P. (1999). A Second-Order Hidden Markov Model for Part-of-speech Tagging, In *Proceedings of the 37th annual meeting of the ACL on Computational Linguistics,* pages 175-182.

Urdu Normalization Utility v1.0. (2009). Centre for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, Pakistan. Retrieved April 14, 2011 from http://www.crulp.org/software/langproc/urdunormalization.htm