8. SUMMARY

This thesis presents a complete Shahmukhi-to-Gurmukhi transliteration system with an impressive overall word accuracy of 96.94%. This Shahmukhi-to-Gurmukhi transliteration system is, to the best of our knowledge, the first of its kind that can run on Shahmukhi script with missing short vowels and other diacritics and excludes any manual intervention. Both online and offline versions of this transliteration software have been developed. The online version is available free-of-cost at department website http://s2g.learnpunjabi.org. Till date, we have more than 1200 registered users worldwide using the online system. Among these registered users, the maximum number is from India (47%) followed by Canada (15%) and USA (13%) as shown in Figure 47. A user can transliterate a text by typing it in a box provided at webpage or one can submit an InPage file directly whereby the system would automatically convert it into Unicode. A user can also transliterate any Unicode-based Shahmukhi website like www.wichaar.com or www.likhari.com etc., and can view it in Gurmukhi at the click of a button. The current system is one of the best Machine Transliteration Systems, enhancing networking between India and Pakistan and promoting the Punjabi language.

![Figure 47 Country-wise Usage of Shahmukhi-to-Gurmukhi Transliteration System](image)

In this chapter, we will summarize the achievements of the present research work. Directions for further research that can help to enhance this Machine Transliteration System have also been included.
8.1 Contributions

− **Handling the Missing Diacritical Marks and Character Level Ambiguity:** For the first time a statistics-based system for supplying missing diacritics and vowels for Shahmukhi script has been developed. We have proposed two algorithms for handling the missing diacritical marks and character level ambiguity along with hand-crafted contextual or positional dependency rules.

− **Lexical Resources:** The scarcity of resources of script pair in question like, availability of bilingual electronic dictionary etc., required the resources to be developed from scratch. So for the first time, we developed Shahmukhi and Gurmukhi scripts corpora, word frequency list, bigram, trigram tables and bilingual lexicon for machine transliteration.

− **InPage to Unicode Conversion:** InPage is the most popular Urdu word processor. The majority of the Shahmukhi soft data was found in the form of InPage software files. This soft data was required to be converted into its Unicode format. Therefore, a utility for conversion of InPage text to Unicode format has been developed and integrated into the pre-processing stage. This has provided flexibility to the proposed system and an additional facility for the user having the source text as InPage files.

− **Corpus Analysis and Comparison:** For the first time the two scripts have been compared at both character and word level to infer individual characteristics, similarities and differences. Comparison highlighted the significant fact that the first most frequent Gurmukhi and Shahmukhi character spans 8.89% and 13.57% of corpus, respectively. The top three characters of Shahmukhi cover 30% of its corpus and have multiple mappings into Gurmukhi character set. About 50% of the corpus is covered by first 10 characters of Gurmukhi vis-à-vis first 7 Shahmukhi characters cover the same in Shahmukhi corpus. Similarly, 95% of Shahmukhi corpus is covered by the first 28 characters whereas the same coverage in Gurmukhi corpus is made by the first 36 characters.

− **SoundEx for Gurmukhi Script:** For the first time, we have proposed a phonetic character level coding guide for Gurmukhi script in which Gurmukhi letters are mapped to 20 distinct phonetic code groups. It is an extension of Soundex technique, where Gurmukhi words are indexed phonetically rather than alphabetically. It is used for suggesting the missing vowels in the target script.
Shahmukhi Word Segmentation: For the first time the segmentation issues in Shahmukhi have been addressed. Statistics-based algorithms have been developed to handle space insertion and space omission problems. A novel technique of using the Gurmukhi training set to handle Shahmukhi segmentation issues has been developed.

Two Algorithms for Word Disambiguation: For the first time the problem of word level ambiguity present in Shahmukhi script has been addressed, using N-gram with small window context of size ±5 and HMM approaches.

Real Time Website Transliteration: A real time website transliteration facility has been developed. The user can convert any Unicode-based Shahmukhi website into Gurmukhi script with just a single click and can easily read and understand the matter published in Shahmukhi script. The preliminary idea of this approach has been demonstrated in the International Conference Coling-2008 [250]. According to our registered user Safir H. Rammah (www.apnaorg.com) they primarily rely on this online software to convert APNA's Punjabi magazine (published in both Gurmukhi and Shahmukhi scripts) Sanjh's Shahmukhi files to Gurmukhi. Similarly, Manzur Ejaz (President, Wichaar) appreciated this software which is very helpful for developing Gurmukhi Section of their popular Shahmukhi website www.wichaar.com (see Appendix-1).

Development of a Highly Accurate Shahmukhi-to-Gurmukhi Transliteration System: For the first time all the challenges posed by Shahmukhi script are handled, such as missing diacritics, multiple mapping or character ambiguity, word segmentation and word level ambiguity. In addition, script writing rules, character mappings, handcrafted contextual dependency rules, and heuristics are utilized in various phases of transliteration to attain high levels of accuracy. Very efficient data structures were developed to produce such a highly accurate and fast system implementation.

8.2 Recommendations for Future Work

Considering the present scenario, we believe that it is difficult to make any significant improvement beyond that achieved in our Shahmukhi-to-Gurmukhi transliteration system. However, there is always scope for improvement. Following are some of the future directions:
Increasing the Data Size and Its Quality
There is always a chance of improvement in the system performance upon improving the quality and size of the underlying system training data. Therefore, an increase in the size of current lexical resources and the quality of Shahmukhi and Gurmukhi script corpora will definitely yield better results.

Unicode-based Resources
Some early efforts in Gurmukhi script were found to be based on non-standard fonts that are designed without any consideration of its compatibility with Unicode standard. Therefore, there is an urgent need for developing various Unicode-based resources for promotion of Punjabi language and to get better results in future.

Integrating Word Morphology
The disambiguation task performed in the post-processing phase can be further improved by taking word morphological information into account in addition to the contextual information which is currently being used.

Foreign Words from Urdu Domain
A considerable number of Urdu words are found in Shahmukhi text (as foreign words) that have originated in Pakistan. A good online source for example is www.wichaar.com. Therefore, we can extend this system by adding an additional knowledge base component to translate the foreign Urdu words into equivalent Punjabi (Gurmukhi) language. This effort will increase the understandability of particularly those users having little or no knowledge of Urdu language. In fact, this is the case with the majority of Punjabi people using Gurmukhi script of writing.

Information Retrieval (IR) System
Yet further potential for future research is in developing an Information Retrieval (IR) System for the Unicode text of Punjabi language encoded in both Shahmukhi and Gurmukhi scripts.