REGIONAL LANGUAGES CHARACTER MAPPER AN OPEN SOURCE APPROACH

Hardeep S. Jawanda *

Pardeep S. Cheema**

1. Introduction:

Gurmukhi and Shahmukhi are two scripts of Punjabi language. These two are cognitive languages spoken in similar fashion, yet dissimilar in written scripts. Gurmukhi is mostly used in Indian part of Punjab and Shahmukhi is mostly used in Pakistan Punjab. When Punjabi is spoken, there is no change baring different dialects. On the contrary when the same speech is written in Gurmukhi it has different appearance then when written in Shahmukhi. This is because as per the consonants in the script that it uses, different symbols are placed for same speech in different script. 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.Majority of the Punjabi historical literature is available in Shahmukhi script , as Gurmukhi script was unavailable at the time. After the invent of Gurmukhi , lot of work has been there, and the Great Religious work “Guru Granth Sahib Ji” is written in Gurmukhi. Interaction is always required due to exchange needs that exist between these scripts, more so in educational context.The exchange of literature requires the knowledge of both the scripts. There are few people that know both the scripts and are able take advantage of the rich literature that is available in both scripts. But majority of people are deprived of at least one literature part. If they know Shahmukhi , then they are able to read literature available in Shahmukhi and are deprived of Gurmukhi literature. Similarly , if they know Gurmukhi, then

* Deptt of CSE, Guru Nanak Dev Polytechnic, Ludhiana.

** Deptt of CSE, SLIET Longowal , Sangrur

they are able to read literature available in Gurmukhi and are deprived of Shahmukhi literature.The interaction between script-based languages to enhance the available knowledge and information of regional languages is a considerable tool in sustaining the heritage and development of languages. The envisaged "Regional Language character Mapper under Open Source Technologies" (RLCMOS) is a significant tool to map the information from one regional language to another regional language. To allow the wide spread usage of language and its development, RLCMOS is kept as an open source. This scenario is more applicable in cases where two regional languages are near to each other and yet need translation, and this is exactly the requirement between Gurmukhi and Shahmukhi. Both are identified forms of Punjabi across border between India and Pakistan. There is recognized approximate phonetic equivalence between both languages. RLCMOS will act as useful for machine translation, cross-lingual information retrieval, multilingual text and speech processing between both languages.

The number of Punjabi speakers as indicated in various census and studies have grown substantially over the period of time and table below shows approximate number of Punjabi users,

Sr. No.	Region	Speakers
1	India(1994 IMA)	25,690,000
2	Malaysia (1993)	43,000
3	Kenya (1995)	10,000
4	Bangladesh (1961 census)	9,677
5	Fiji	1,167

Table 1. Number of Punjabi Speakers

A population of Punjabis have also immigrated to Britain, Canada and US, thus putting it on global map and native language of approximately 110 million people (Abbas Malik, 2006).

A Comparison between Gurmukhi and Shahmukhi can be drawn to see that how much these two much scripts are related or unrelated.The character sets of Shahmukhi and Gurmukhi are entirely different from each other and originated from different sources. Gurmukhi was initially formed by Guru Angad Dev (second Sikh Guru, 1504 – 1552) in the 16th century and contained 35 consonants, and at that time derived its character set form Landa script. It is written from left to right and unlike Shahmukhi its characters do not assume different shapes and also do not have small and capital forms. This form of Gurmukhi is not in use currently and a modern version of Gurmukhi has replaced it. The present Gurmukhi has 41 consonants, 9 vowelssymbols, 2 symbols for nasal sounds, 1 symbol that duplicates the sound of any consonant, 3 subjoined forms of the consonants Ra, Ha and Va and 1 post-base from of Ya (Bhatia 2005).

Genesis of Shahmukhi character set is from Persian/Arabic Its use started influencing Punjabi with the spread of Muslim rulers in Indian Subcontinent. Shahmukhi script is written from right to left and the shape assumed by a character in a word is context sensitive. It has 49 consonants, 16 diacritical marks, 16 vowels, etc. (Abbas Malik 2005).

The numeric characters of both languages are represented from 0 to 9 by separate set of characters.

The scope of RLCMOS can be easily extended to other cognate languages. This software "Regional Language Character Mapper Under Open Source" - RLCMOS is precisely doing this job of mapping the information from one regional language to another regional language. Any valuable information available in one regional language unless translated into another regional language will not be available to others and will be confined to people knowing that language only. Open source community is taking this task very seriously and self driven developers , enthusiasts and take every opportunity to spread and port regional language knowledge. Rural and Adult education are two areas where due to regional confinement the due knowledge is not imparted in the right way. Computer has made life of many, very easy. Like information sharing across the globe, educational presentations, audio – video lectures, softcopy of chapters, animated books, e-books etc. The main challenge is to choose two regional languages that are near to each other and yet need translation. Although there are thousands of regional languages to choose from for conversion, in my view Urdu to Punjabi, Hindi to Punjabi or English to Punjabi translation combination are good to explain the concept of this research.

Another major step forward that has been taken in this research is the "Total Open Source" approach, i.e. the softwares that are used to develop, debug and test the transliteration software under this research are all open source and , ultimately the transliteration software that is developed (the final application) is also kept as Open Source. Slowly but steadily Open Source (a concept of freedom) has become relevant to this present scenario, especially when we are dealing with education in rural area. Majority of the softwares in the field of transliteration are propriety and closed softwares i.e cost is also involved and the source code is also not available.As any software which is of a great use , will change hands and many people would use it. Out of these people many would be software developers, language experts,core users and they will having lot of suggestions and ideas. If source code is available developers, language experts , users can themselves change the code to incorporate the ideas. Further to this a continuous chain reaction of continuous improvement can occur in the transliteration software. Education written matter available in either of the scripts can be translated to save time and cost. But if the transliteration software itself cost more then it may not be viable to help the education field at low cost. Therefore use of Open Source to develop another open source software is the key to success.

2. Purpose:

The main purpose of this research is to bridge the gap between the two scripts on the pillars of Open Source by using machine transliteration. To support this purpose we attempt to design a new simple translator for Punjabi-Urdu (Shahmukhi) which will enable us to convert open source educational Punjabi to Urdu for teaching/translation purpose and use Open Source Technologies for low cost / free transliteration alternative.

3. Transliteration:

Transliteration is generation of target language from the input language / source language.

Most of the methods used for this purpose are based on statistical approaches. The major techniques for transliteration can be broadly classified into two categories, viz grapheme-based and phoneme-based approaches. Transliteration takes into account writing style, consonants,

grammar, names of source as well as of target language. With the adventmodern time foreign language words are incorporated into other languages and these need not to be transliterated as these words are accepted in their original form in both source and target language. Another thing that is taken into account while transliteration is proper nouns like person name, place name etc. These words are to be recognized first and then proper transliteration be done.Most of the work done on this field is confined to well known languages of the world. Work of translation for Indian regional languages is very less. The urban/rural population of India need translated work as the masses of urban. It is the rural India that need translated work as rural masses only know their regional language and approx. 70% of Indian Population live in rural area.

4. Challenges:

Transliteration task is a mammoth task and comes with lot of challenges. Various challenges that are being tackled are :-

Handling Non Dictionary Words:- The script and the pronunciation decide the vocabulary of a language. The script once written needs hardly any change, but pronunciation keeps changing over the years. Therefore new words come up while speaking. Thus these new words are not present in the dictionary. These words than are not translated. These can only be translated once they are added into the dictionary.

Less work in this field :- Punjabi literature , dictionaries and other material in Punjabi is not in abundance thus limiting the scope of such a research .

Increased usage of Computer:- Internet usage has increased phenomenally. Rural India is seeing the internet boom. As the reach of internet and computers has probed the rural India, this has led to growth in the number of Regional Indian Language content available on the web. It has become a challenge to produce a single content in hundreds of regional languages.

Filling the Missing Script Maps:- There are many characters which are present in the Shahmukhi script, corresponding to those having no character in Gurmukhi.

Multiple Mappings:- It is observed that there is multiple possible mapping into Gurmukhi script corresponding to a single character in the Shahmukhi script

Mapping of Simple Consonants - Unlike Gurmukhi script, the Shahmukhi script does not follow a 'one sound-one symbol' principle. In the case of non-aspirated consonants, Shahmukhi has many character forms mapped into single Gurmukhi consonant.

Proper Noun conversion is another area that poses challenge due to no proper work under this.

5. Related Work:

Most of themethods used for transliteration purpose are based on statistical approaches. As transliteration is not a new problem. The major techniques for transliterationcan be broadly classified into two categories, viz grapheme-based and phoneme-based approaches. Knight et al. (1997) developed a phoneme-based, statistical model usingfinite state transducer that performed back-transliteration using transformation rules. Paola and Khudanpur (2003) used another phoneme-based approach using transformation based learning algorithm. Yaser and Knight (2002) used a grapheme-basedapproach that maps English letter sequences to Arabic letters. Abdul Jaleel and Larkey (2003) demonstrated a simple, statistical technique for building an English-Arabic transliteration model using Hidden Markov Model (HMM) and alignmentsobtained from GIZA++ (Och and Ney, 2003).

Some of the works under proper noun conversion is 'Named entity recognition' - Though Named Entity Recognition (NER) is a known research area (e.g. MUC-6 1995, Daille& Morin 2000), multilingual Named Entity Recognition is quite new (ACL-MLNER 2003, Poibeau 2003).

6. Methodology:

The methodology used to develop Open Source Transliteration Software in this research is very simple.Corpus and Rule Based Technique applied in this research. The input text is made free from any mistakes and converted to Unicode format. This is known as normalization. The advantage here is that this will reduce the text scanning complexity and makes the work available for internationalization. Tokenization is done first to find words in the sentence. The system is designed to do sentence level translation in order to speed up the process. Individual words or tokens are also extracted to find the equivalent in the target language.The RLCMOS

will use phased scheme during implementation, and move systematically from word, alphabet to grammatical correction during transliteration.

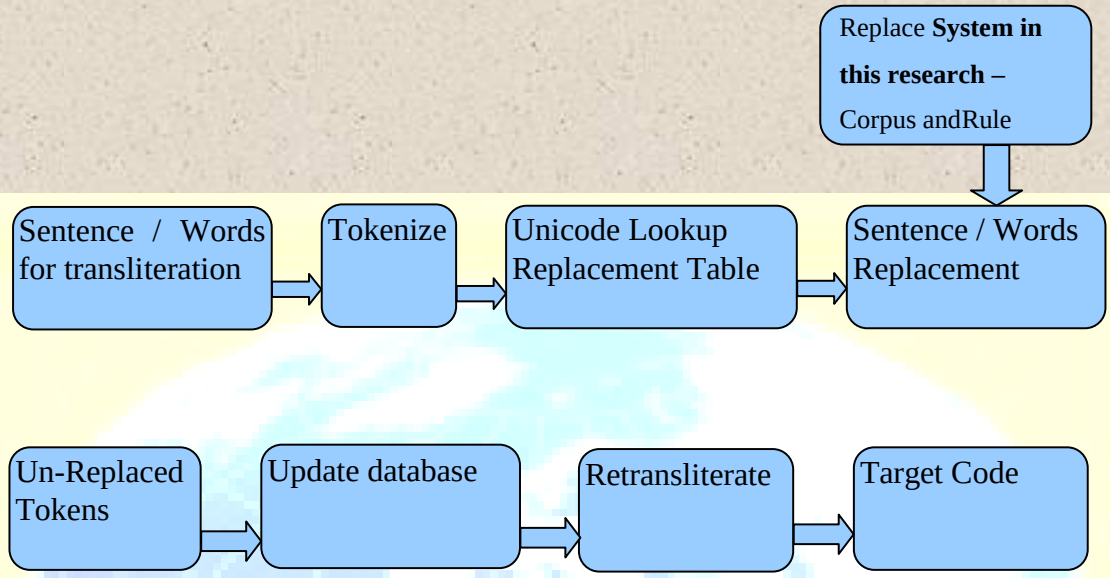

Figure 1 Transliteration Methodology

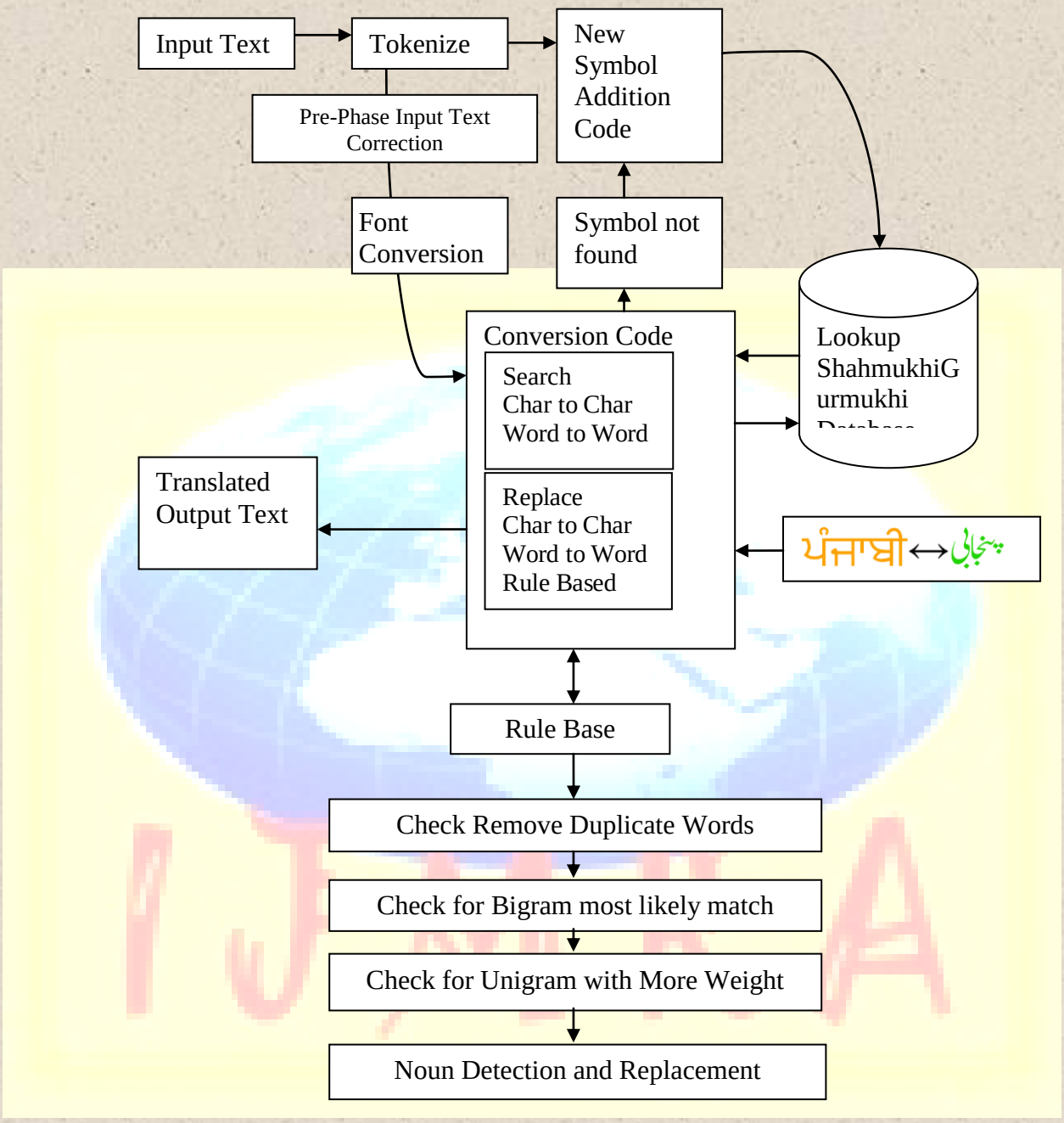

Figure 2 Detailed Structure of Transliteration System

7. RLCMOS Emphasis:

To increase the usability of RLCMOS and improve its educational development, it is proposed to keep it as open source.This step is advantageous to provide free access of code to users and allow add any additional features as deemed essential.Use of Java enables RLCMOS to be platform independent.

8. Implementation:

The transliteration system is virtually divided into two phases. The first phase performs pre-processing and the second phase performs post-processing. In phase I ,input text is subjected to tokenizer ,these tokens are then processed for any consecutive repeated occurrences.Tokens are also checked for mistyped or unwanted characters.Unwanted characters are deleted and the input text made completely free from errors before transliteration. In second phase sentence / word lookup and replacement is done fromShahmukhi - Gurmukhi Dictionary. As discussed earlier implementation is done using open source tools , Java and MySql.

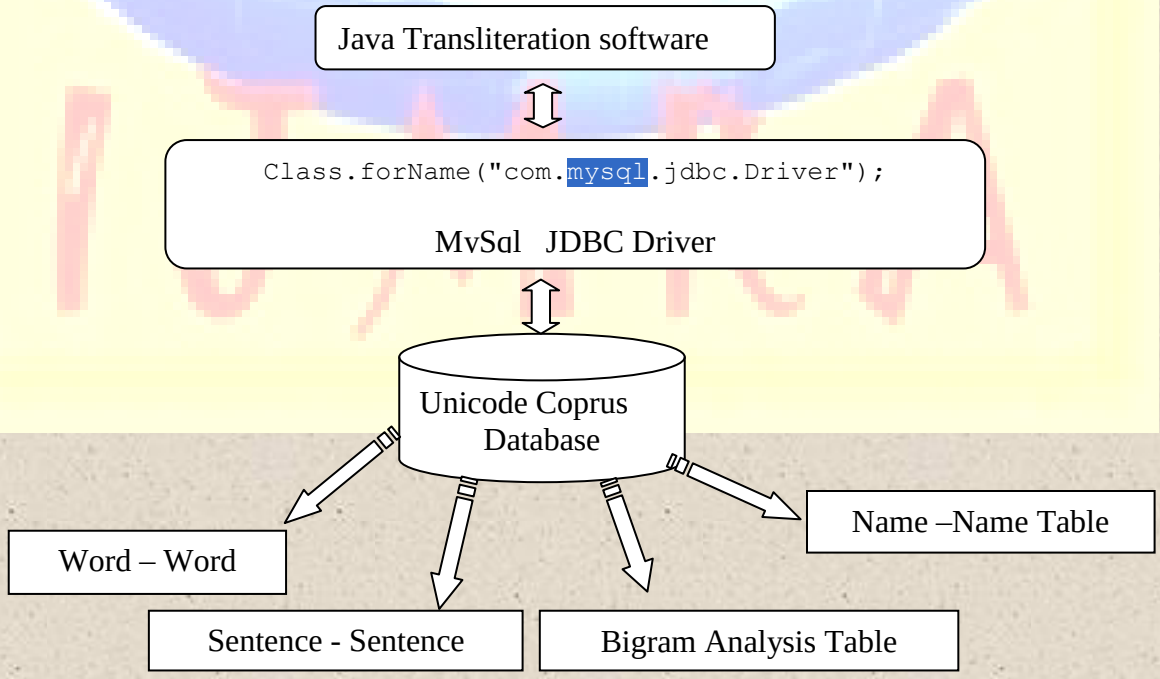

Figure 2 Implementation Overview Java – MySql

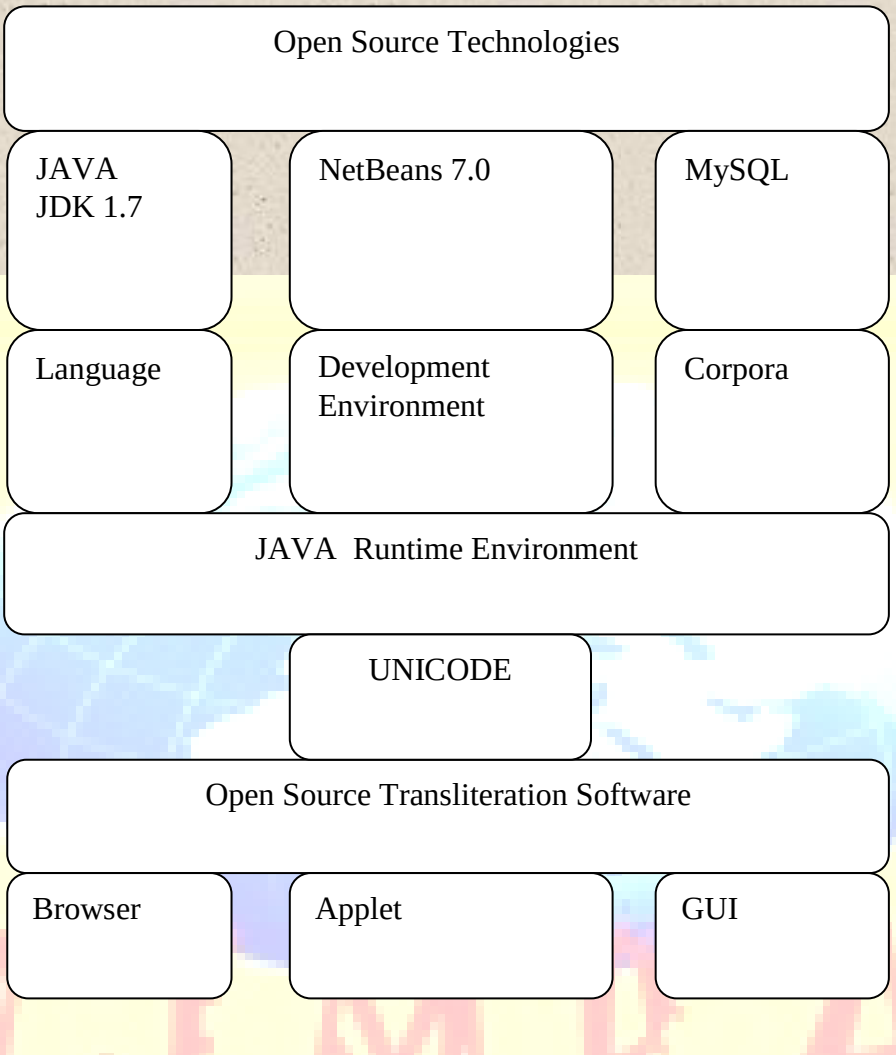

Figure 3 Structure of Open Source Transliteration Software Implementation

9. Major Results:

S.No.	Input Gurmukhi Text	Type	Output Rule Based (Macro)	Output Corpus and Rule Based (Java Code)
1	tYknflojIvrHyqusINglLsuxI	Sentence	Complete Transliteration	Complete Transliteration
2	aKLiqafr	word	Complete Transliteration	Complete Transliteration
3	skulborz	word	Near About 90 %	Half Only 50%
4	nvjoq	Noun (Name)	Near About 80%	Nil 0%

Table - 2 Results

The above table of results shows the comparison and the output of already available technique (Rule based Macro) provided by K.S. Panuufor testing and understanding and the basic open source transliteration software under this research. The results shows that total rule base is more effective and accurate. The combination of corpus and rule base as in our case tend to loose accuracy if corpus fails to deliver.

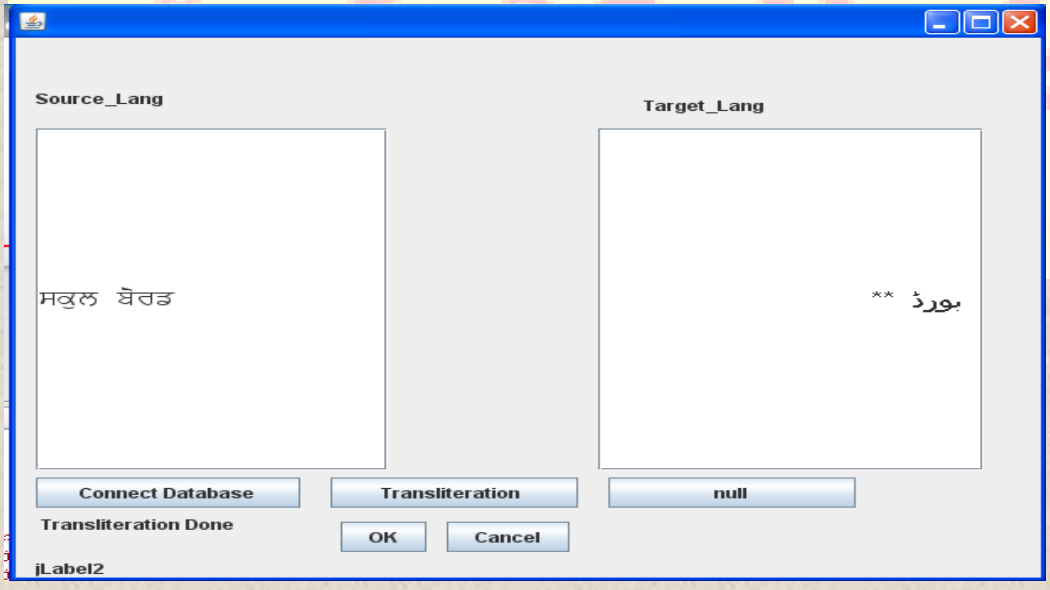

Figure 4 Screenshot of Result at table 2

10.Conclusion:

RLCMOS helps in progression of educational development of Punjabi language, resulting into breaking of barrier between its two forms namely Gurmukhi and Shahmukhi. It will alsoresult as a useful tool for switching between two languages. Being an open source and platform free, the usage will be unrestricted and thus invoking further development of the tool.

Acknowledgment:

Hardeep Singh Jawanda thanks Kirpal Singh Pannu for providing his analysis and database for use and testing of this work.

References:

- Abbas Malik, M. G. 2005. Towards Unicode Compatible Punjabi Character Set. In proceedings of 27th Internationalization and Unicode Conference, 6 – 8 April, Berlin, Germany.
- Abbas Malik, M.G., Panjabi Machine Transliteration, 2006, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1137–1144.
- Abdul Jaleel N., Larkey L. S.,2003, "Statistical Transliteration for English-Arabic Cross LanguageInformation Retrieval", CIKM '03: Proceedings of the Twelfth International Conference onInformation and Knowledge Management, ACM, New York, NY, USA, p. 139-146,
- ACL-MLNER 2003, Workshop on Multilingual and Mixed-language Named Entity Recognition, ACL 2003, Sapporo, Japan, http://acl.ldc.upenn.edu/acl2003/mlner
- Anil Kumar Singh —Sethuramalingam Subramaniam —TarakaRamaTransliteration as Alignment vs.Transliteration as Generation for Crosslingual Information Retrieval- TAL. Volume 51 – n°2/2010

- 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
- DailleBéatrice, Morin Emmanuel, (2000) Reconnaissance automatique des nomspropres de la langue écrite: les récentesréalisations, TraitementAutomatique des Langues (TAL), Vol. 41(3), pp.601-622, 2000.
- Och F. J., Ney H., "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, vol. 29, n° 1, p. 19-51, 2003.
- [Knight and Graehl, 1997] Kevin. Knight and Jonathan Graehl. Machine Transliteration. In Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pages 128–135, Somerset, New Jersey, 1997. Association for Computational Linguistics.
- Paola V. and Khudanpur, S. Transliteration of proper names in cross-language applications. In Proceedings of SIGIR. 2003, pp. 365-366