Hybrid Transliteration Framework for Kashmiri : Managing Diacritic Subjection in English-to-Native Script Conversion

Sameer ul Rahman[1], Er. Shilpa[2]

[1]PG-Scholar ., School of Computer Science and Engineering,Rayat Bahra University Mohali, [2]Asst. Prof., Computer Science and Engineering Department, Rayat Bahra University Mohali

Abstract:

In a world where connections between people are growing, effective communication between languages is essential. Machine translation systems play a pivotal role in facilitating seamless text conversion between languages, enabling communication in diverse fields such as entertainment, education, and commerce. However, certain linguistic elements, such as named entities, require specialized treatment to preserve their phonetic integrity across languages. Transliteration, the process of converting words while maintaining pronunciation and phonetics, addresses this need. This research focuses on developing a transliteration system from English to Kashmiri, recognizing the importance of maintaining phonological characteristics during script conversion. Leveraging natural language processing techniques, the system aims to efficiently convert text while ensuring readability for individuals from diverse linguistic backgrounds. The approach combines phoneme-based and grapheme-based transliteration methods with diacritic subjection to accurately represent sounds and characters in the target script. By bridging linguistic barriers and enhancing cross-linguistic communication, this research contributes to the accessibility and inclusivity of information across different language communities.

Keywords: *Machine Transliteration, Diacritics, Named Entities, Natural Language Processing, Phoneme, Grapheme, Kashmiri Language, Script Conversion.*

1. Introduction:

The linguistic tapestry of our world comprises approximately 6500 diverse languages, each contributing to the richness of human culture. Yet, this diversity often poses a challenge for individuals from disparate regions to effectively communicate. Transliteration emerges as a potent solution to bridge this gap, facilitating the exchange of ideas and fostering cross-cultural understanding.

Machine Transliteration serves as the automatic method employed by algorithms to transcribe alphabets or syllables of words from one script or language to another. It finds utility in various applications such as Question-Answering, Information Extraction, Foreign Language learning, Machine Translation, and Data Mining. For languages with different writing styles and fonts, like English and Kashmiri, transliteration becomes essential due to the disparity in scripts.

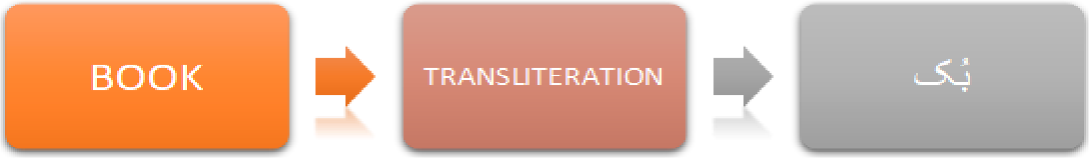


Figure 1 : Transliteration

For Named entities (NE) and Out-Of-Vocabulary (OOV) words, transliteration is indispensable in the development of machine translation systems, especially for language pairs like English-Kashmiri. English, with its Latin (Roman) script, contrasts with Kashmiri's Perso-Arabic script, requiring specialized transliteration approaches [1]. Despite the existence of alternative scripts like Devanagari and Sharada for Kashmiri, they are rarely used today.

In essence, transliteration intersects with Natural Language Processing (NLP), a field pivotal in artificial intelligence and computational linguistics. NLP enables computers to interpret and understand human language, narrowing the gap between human conversations and machine understanding. By leveraging NLP techniques, transliteration can be enhanced to ensure precision and consistency in converting phonemes and graphemes across scripts.

Transliteration plays a vital role in cross-language information retrieval, machine translation, and data mining, where maintaining the pronunciation and phonetic integrity of words is paramount. However, challenges persist due to differences in writing styles, phonetic variations, and the varying number of graphemes across languages [2].

Through the integration of transliteration and NLP, we embark on a journey to redefine artificial intelligence, enabling seamless interaction between humans and computers across linguistic boundaries. This interdisciplinary approach holds the promise of unlocking new possibilities in communication, fostering inclusivity, and advancing the frontiers of linguistic technology.

2. Problems In Transliteration

Transliteration is useful in Cross language information retrieval, Machine translation, Data mining, etcand is a part of Natural Language Processing (NLP) . When we translate a sentence frome one script to another script (i.e, from source script to target script) the target script should be transliterated not translated. For example if “Danish” in a document refers to the name of a person then it should remain Danish in all the languages and it should not get translated. For example, in kashmiri language, we write “کُل“. This in English may be written as “kul”. The letter "k" in English is synonymous in sound to as of the letter "ک "in kashmiri and so on [3].

Also, the pronunciation of the word should remain same. Thus this makes transliteration is a diffiucult task as the number of alphabets in the different languages are different and each of the alphabet have different phonotic sounds.During transliteration, the equivalent phonemes of source and target scripts are replaced.

While transliteration, there are number of problems due to phonems of the scaracters, phonetic sounds, difference in the number of the vowels and cononents etc.

Some basic problems in transliteration are listed below:

Different languages are made up of alphabets and every set of alphabets have different number of vowels and consonents. So, each vowels and consonents have different phonemes. Thus character matching is not directly applied here while transliteration.

LANGUAGE	VOWELS	CONSONANTS
ENGLISH	5	21
KASHMIRI	23	29

Table 1: Number of Vowels and Consonants in English and Kashmiri scripts

We already know that all languages have different sounds for their characters and also some of therse sounds are created by using digraph (two characters) and trigraph(three characters) i.e, by combining two or three characters of one language so that the phonems will be same of the other language.For example, letters like "se" and "jim" in english represent "ث" and "ج".

Kashmiri characters represent phonetic nuances absent in English	An English counterpart for Kashmiri characters	Phonetic Graph
ث	se	digraph
ج	jim	trigraph

Table 2: An example digraph and trigraph characters in English and Kashmiri scripts

Silent letters also create problems while transliteration. For example, the words (usually greek and latin) create difficulities in pronunciation, like the word "pseudomonas" in which the first letter "P" remains silent and thus it becomes difficult to judge. To overcome this, the origin of the word should be kept as one of the aspect for transliteration.

Sometimes the letters in one language represent two or more letters in other language, so the phonemes of one language are different for two are more letters of the other language. For example, the letter "D" in english is equivalent to "د" and "ڈ" of the kashmiri language.

3. Literature Survey

The field of Transliteration employs cutting-edge techniques and advanced engineering to simplify complexity and enhance computer systems' interaction, interpretation, understanding, and management of human language. Transliteration methods such as Machine Learning, Natural Language Processing Models, and Data Mining are commonly utilized. Here's an overview of various transliteration models:

Mir Aadil et al. [1] demonstrates the utilization of a Phoneme-Based model for transliterating English into Kashmiri, offering potential applications in information extraction and machine translation for this language pair. Testing the system on medical domain and Wikipedia-based English text yielded an overall accuracy of 86%.

K. Naren Sai Krishna et al. [4] employs two Recurrent Neural Networks, specifically an Encoder and a Decoder, to represent a word in Devanagari script (Hindi). The accuracy of the model notably improves with the integration of an attention mechanism alongside the Encoder-Decoder Model.

Sitender et al.[5] published a survey article that delves into new research directions that promise to enhance the development of high-quality Machine Transliteration Systems (MTS). It adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach, incorporating tools and evaluation methods tailored for English, Hindi, and Sanskrit languages.

Sahinur Rahman Laskar et al. [6] have introduced a method based on transliteration to augment phrase pairs, demonstrating enhancements in multimodal translation tasks. Achieving the second-best performance on the challenge test set for English to Hindi multimodal translation, they attained a BLEU score of 39.30 and a RIBES score of 0.791468.

Harish, B.S et al.[7] examined the diverse approaches and methodologies introduced by researchers in the realm of Indian regional language processing. They analyzed tasks such as machine translation, Named Entity Recognition, Sentiment Analysis, and Parts-Of-Speech tagging, considering Rule-based, Statistical, and Neural-based approaches.

Sayantan Paul et al. [8] developed frameworks for language-independent machine transliteration using neural network-based deep learning architectures, including Convolutional Sequence to Sequence and Recurrent Neural Network models, yielding promising results across multiple languages.

Nizar Habash et al. [12] investigated the conversion of unconstrained Egyptian Arabic into standard word order, achieving a 69% reduction in deviations from the benchmark through a two-phase process involving glyph and term transformations along with syntactic tagging.

Nadi Tomeh et al. [13] introduced a universal yet tailored approach for rectifying spelling mistakes at the glyph level, focusing on character-level analysis while incorporating term-level and contextual data. Their system significantly reduced error rates in Egyptian Arabic text by 65% compared to existing models.

Mitesh, M et al. [14] proposed a compositional machine transliteration model allowing for flexible integration of transliteration elements to improve performance between languages lacking primary comparable corpora. They applied two configurations, Serial and Parallel, demonstrating superior performance over traditional transliteration models.

Taraka et al. [15] treated transliteration as a translation problem and employed Statistical Machine Translation (SMT) techniques for English-Hindi language pairs, achieving a precision of 46.3%.

Pankaj et al. [16] implemented a Statistical Machine Translation model to transliterate proper nouns from Punjabi into English. The method focuses on transliterating proper nouns from Gurumukhi script into their corresponding English names. The model was tested on a diverse set of names, with a sample size of 1000, achieving a precision of 97%.

Manikrao et al. [17] have introduced a phoneme-based system for transliterating Indian proper nouns into English, employing a solely consonant-based methodology. Their approach utilizes a hybrid technique, combining rule-based and metric-grounded methods to omit schwa. This direct approach does not rely on training bilingual databases and demonstrates a deep understanding of word structure in the Devanagari script, specifically targeting Hindi and Marathi to English transliteration.

Muhammad Ghulam Abbas et al. [18] created a Punjabi Machine Transliteration System designed to transliterate expressions from the Shahmukhi dialect to the Gurmukhi dialect. The core concept involves glyph mappings and reliance regulations, as mere character mappings at the onset are insufficient for the system's effectiveness. The developed model achieves over 98% precision in traditional literature and 99% in modern literature.

Abdelmajid et al. [19] proposed a method for identifying and translating Arabic named entities, utilizing a characterization model, a set of bilingual lexica, and transducers. These tools are instrumental in identifying linguistic and dialectal nuances associated with Arabic named entities. Their resources are deemed reusable, as evidenced by the trial and assessment validation conducted on the NooJ platform.

Nasreen et al. [20] presented a statistical approach for training an English to Arabic transliteration model without heuristics or lexical knowledge, showing effectiveness in transliteration tasks.

G.S.Josan et al. [21] initially employed a baseline method involving character-to-character matching. Subsequently, they evaluated this approach against a statistical method for transliteration, utilizing a Noisy channel model. They further suggested potential enhancements to their system, such as refining the language model through adjustments in alignment heuristics and maximum phrase length. Additionally, they proposed enhancing the syllable similarity score for better performance.

These studies highlight the diverse approaches and methodologies employed in machine transliteration across various languages and dialects.

Ref.	**Year**	**Dataset**	**Method**	**Observations**
[10]	2017	English-Hindi Lexicon	Character-level conversions	The word precision of the suggested transliteration software has been determined to be 70.22%, compared to 58.73%
[8]	2018	N/A	Neural networks	A Convolutional Sequence-to-Sequence neural network model and a Recurrent Neural Network have been shown to yield satisfactory results in multilingual machine transliteration.
[12]	2014	Egyptian -Arabic Lexicon	Character-level conversions	Reduction in error rate compared to the input standard, and enhancement beyond the aforementioned state-of-the-art model.
[13]	2013	Arabic - Egyptian Lexicon	Complete syntactic tagger,Script transformations.	A two-stage process can reduce deviations from the benchmark by 69%, facilitating the sequential handling.
[16]	2013	English-Gurmukhi Lexicon	Machine Translation.	The model is evaluated with diverse names, and an additional 1000 names are tested, resulting in a precision rate of 97%.
[17]	2012	Marathi-English,Hindi Lexicon.	Hybrid model based on rule and metric system.	The approach is straightforward, without requiring any bilingual database for training, and demonstrates a comprehensive understanding of word structures in the Devanagari script.

[19]	2011	Arabic name Lexicon.	Transducers,Bilingual Script, (NOOJ).	Their resources are independently reusable, as confirmed by the trial and validation assessment of the model.
[14]	2010	Marathi and Kannada,English-Hindi Lexicon.	CLIR network.	Demonstrated that a Cross-Language Information Retrieval (CLIR) network integrated with a compositional transliteration model consistently outperforms one integrated with a direct transliteration model.
[15]	2009	Hindi-English Scripts.	Beam search-based decode,SMT architecture.	The provided model demonstrates that these approaches can be effectively utilized for machine transliteration tasks, achieving a precision rate of 46.3%.
[18]	2006	Gurmukhi-Shahmukhi Lexicon.	Reliance regulations and glyph mappings.	The developed model achieves over 98% precision on traditional literature and 99% precision on modern literature.
[20]	2003	Arabic-English lexicon.	Elected n- gram pattern.	Evaluated both the statistically-trained model and a simpler hand-drafted model on a trial subset of named entities from the Arabic AFP corpus, confirming that they outperform two online translation platforms.

Table 3 : Comparative examination of current methodologies.

4. Proposed Methodology

4.1 ALGORITHM

The algorithm outlines a comprehensive approach to data acquisition, preprocessing, mapping, and output generation, facilitating efficient transliteration from English to Kashmiri script while addressing challenges such as data redundancy and the scarcity of online literature.

1. Data Acquisition:

- Collect data sets to be used for the experiment.

2. Data Preprocessing:

- Preprocess the collected data to reduce redundancy.
- Perform preprocessing to achieve precision in terms of graphemes and phonemes.

3. Data Mapping:

- Convert each data element into corresponding tokens using the principal controller.

- The principal controller utilizes glyph mappings and diacritic dependency regulations for mapping.

4. Data Output:

- Generate output that isn't merely a letter-to-letter transformation.

- Ensure proper connotation of sounds in the output.

- Since Kashmiri and Urdu lack rich literature online, create a database from scratch.

- Utilize a machine transliteration procedure with glyph mappings and diacritic dependency regulations for real-time transliteration.

- Incorporate a third-party translator for immediate translation of each input term.

4.2 SYSTEM PROCESS

Before outlining the steps of the transliteration process, it's important to understand the intricacies involved in converting Unicode-encoded English text into Kashmiri scripts.

Data inputted by clients in English is parsed into individual alphabets for preprocessing. These alphabets are then mapped to their respective tokens or IDs, representing their transliterated counterparts. Phonemes of the terms are also considered to ensure precise output. The resulting transliterated alphabets or terms are combined to form meaningful words in the desired language. The steps involved are written below:

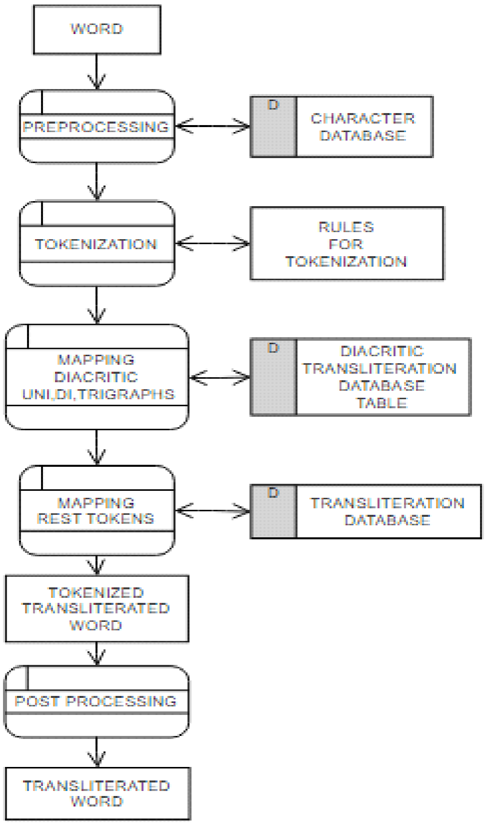


Figure 2 : Architecture and Mappings

1. Input text encoded in Unicode English is processed by the input word parser, which segments the text into individual English alphabets, referred to as tokens.

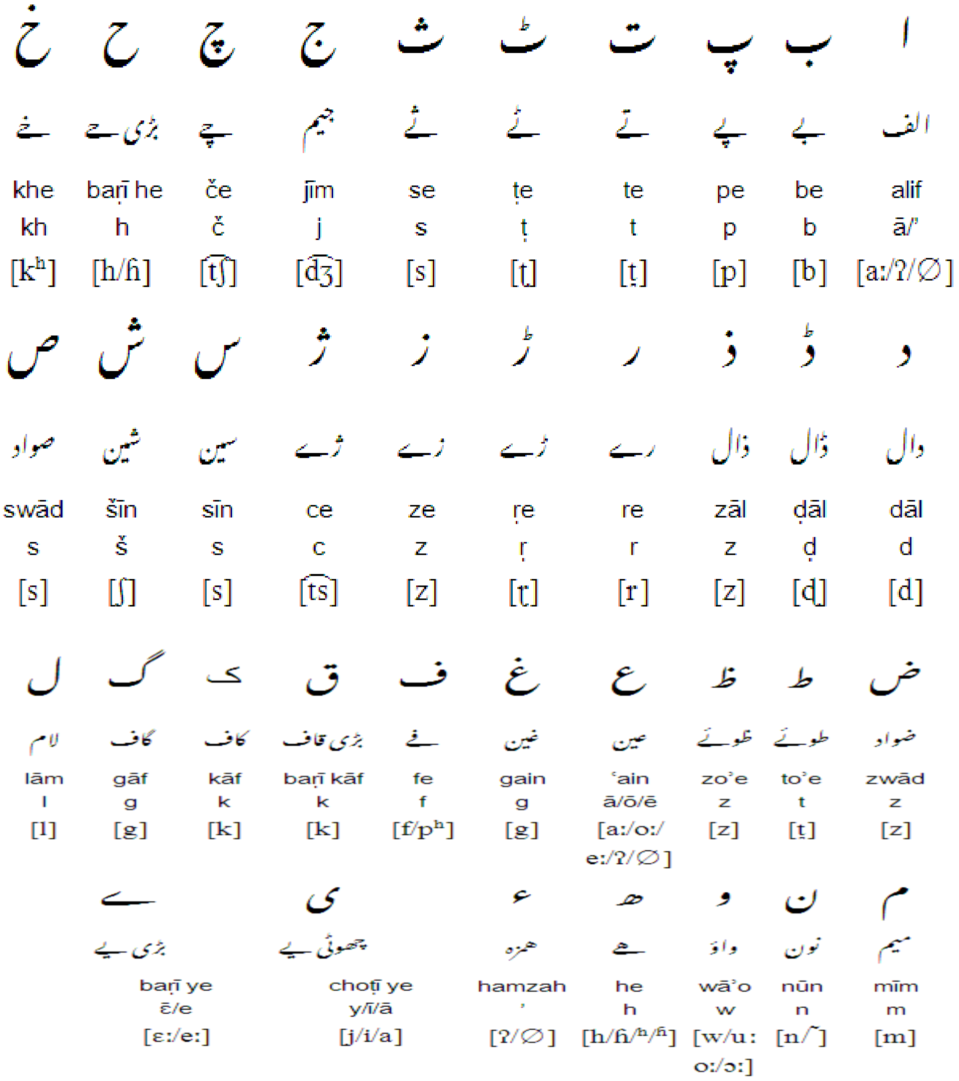
2. These tokens are then passed to the transliteration component, which converts them into corresponding Hindi, Urdu, and Kashmiri tokens using glyph mappings and diacritic dependency regulations controlled by the principal controller.

3. The generated output tokens are returned to the transliteration component. Once all token conversions are completed, the tokens are combined to form the required transliterated output text with diacritic implementation, generated by the output text generator.

4. The machine transliteration procedure manages token conversion and diacritic dependency regulation. It determines and resolves character dependencies while parsing input tokens based on their contextual arrangement.

5. Character dependencies are common due to diacritics in Hindi, Urdu, and Kashmiri scripts, making mapping these tokens from pure English script challenging. Dependency regulations fix character dependencies if present.

6. Tokens without character dependencies are transliterated directly using tokenization and glyph mappings.



خ	ح	چ	ج	ث	ٹ	ت	پ	ب	ا
خے	بڑی حے	چے	جیم	ثے	ٹے	تے	پے	بے	الف
khe	baṛī he	če	jīm	se	ṭe	te	pe	be	alif
kh	h	č	j	s	ṭ	t	p	b	ā/’
[kʰ]	[h/ɦ]	[t͡ʃ]	[d͡ʒ]	[s]	[ʈ]	[t̪]	[p]	[b]	[aː/ʔ/∅]

ص	ش	س	ژ	ز	ڑ	ر	ذ	ڈ	د
صواد	شین	سین	ژے	زے	ڑے	رے	ذال	ڈال	دال
swād	šīn	sīn	ce	ze	ṛe	re	zāl	ḍāl	dāl
s	š	s	c	z	ṛ	r	z	ḍ	d
[s]	[ʃ]	[s]	[t͡s]	[z]	[ɽ]	[r]	[z]	[ɖ]	[d]

ل	گ	ک	ق	ف	غ	ع	ظ	ط	ض
لام	گاف	کاف	بڑی قاف	فے	غین	عین	ظوئے	طوئے	ضواد
lām	gāf	kāf	baṛī kāf	fe	gain	ʿain	zoʾe	toʾe	zwād
l	g	k	k	f	g	ā/ō/ē	z	t	z
[l]	[g]	[k]	[k]	[f/pʰ]	[g]	[aː/oː/ eː/ʔ/∅]	[z]	[t̪]	[z]

ے	ی	ء	ھ	و	ن	م
بڑی یے	چھوٹی یے	ھمزہ	ھے	واو	نون	میم
baṛī ye	choṭī ye	hamzah	he	wāʾo	nūn	mīm
ɛ̄/e	y/ī/ā	ʾ	h	w	n	m
[ɛː/eː]	[j/i/a]	[ʔ/∅]	[h/ɦ/ʰ/ʱ]	[w/uː oː/ɔː]	[n/˜]	[m]

Figure 3 : English and Kashmiri diacritic mappings

Proper regulation and mapping of scripts containing diacritical traces are crucial to ensure accurate pronunciation and meaning in the transliterated text.The precision of the machine transliteration system relies on accurately determined and regulated diacritical traces. Insufficient diacritical traces significantly impact precision.

5. SYSTEM ARCHITECTURE

The server-side model comprises MySQL, utilizing PHP with XAMPP as the development environment. On the client side, HTTP and CSS technologies are employed.

All database operations are performed using MySQL within XAMPP. The database contains collections of letters from each language and their corresponding mappings with phonemes. PHP, in conjunction with JavaScript, facilitates dynamic interactions with the databases. The Yii framework of PHP is utilized for efficient development.

Apache, an open-source web server, handles the transfer of web content over the internet, processing queries and serving them via HTTP methods. It is commonly paired with PHP and is embedded in XAMPP by default.

XAMPP, an acronym for Cross-Platform, Apache, MySQL, PHP, and Perl, provides an open-source bundle of web solutions. It includes Apache server, MariaDB (formerly MySQL), PHP, and Perl, enabling local hosting for website testing before deployment to the main server. XAMPP creates a conducive environment for testing blueprints based on Apache, Perl, MySQL database, and PHP directly on the host network.

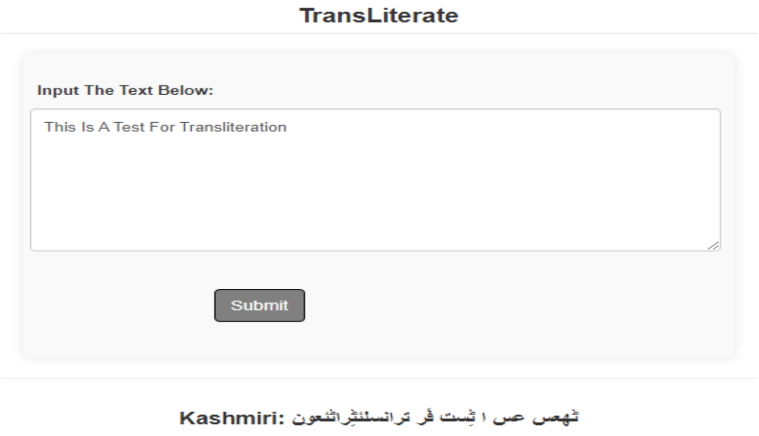
Yii, an open-source OOP PHP framework, follows the MVC pattern and offers robust features for web application development. It simplifies the development process by providing tools and libraries for common tasks, such as database access, form validation, and authentication.

PHP, a recursive acronym for Hypertext Preprocessor, is a widely-used scripting language in web development projects. It is known for its simplicity, flexibility, and broad support for various databases and web servers. PHP code is embedded directly into HTML, allowing developers to create dynamic web pages effortlessly.

Ajax (Asynchronous JavaScript And XML) facilitates asynchronous data exchange between the browser and server without requiring the entire webpage to reload. It enhances user experience by enabling interactive features such as auto-complete search, live chat, and real-time updates. By leveraging browser-embedded XML Http request objects along with JavaScript and HTML DOM manipulation, Ajax optimizes data transmission and improves responsiveness in web applications.

6. Results

After compiling selected input texts, the hybrid machine transliteration system successfully transliterates them Kashmiri texts. Subsequently, the output texts undergo meticulous testing to identify and rectify any errors or inaccuracies. Precision testing is carried out manually, with assistance from dictionaries and individuals proficient in the respective scripts. The system's performance was evaluated using a dataset comprising 1000 words, achieving an impressive accuracy rate of 90%. The primary challenge in achieving such precise results lies in eliminating ambiguity arising from diacritical traces and resolving them through established character dependency rules.



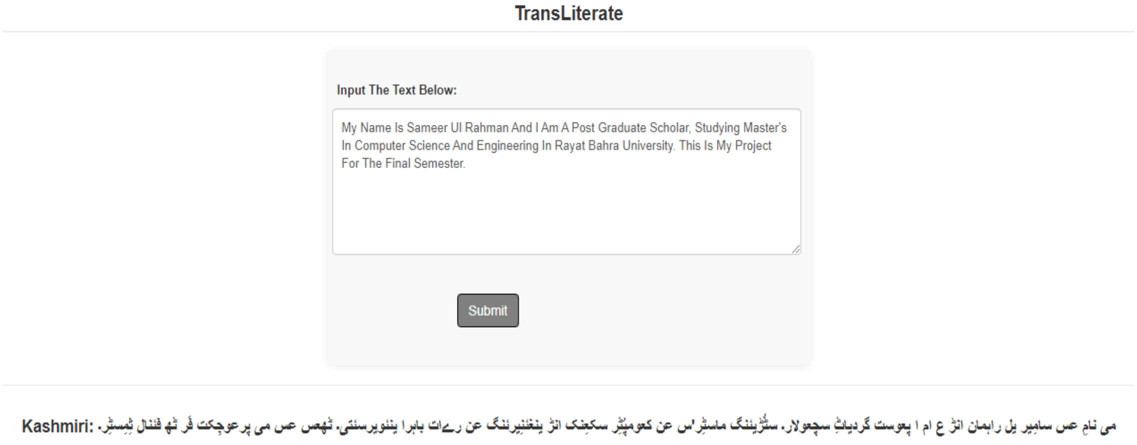
TransLiterate

Input The Text Below:

This Is A Test For Transliteration

Kashmiri: ٹھِس عس ا ٹِسٹ فٗر ٹرانسلِٹِراٹِشون

Figure 4 : Result obtained when applying the method to a uni and di-graph text.



TransLiterate

Input The Text Below:

My Name Is Sameer Ul Rahman And I Am A Post Graduate Scholar, Studying Master's In Computer Science And Engineering In Rayat Bahra University. This Is My Project For The Final Semester.

Kashmiri: می نامِ عس سامِیر یل راہمان انڑ ع ام ا پوسٹ گردیاٹِ سچولار. سٹڑینِنگ ماسٹرٗس عن کومپیٹر سکیِنک انڑ ینگنیرِننگ عن رےات بابرا ینِوِرسِٹی. ٹھِس عس می پروجِکٹ فٗر ٹھ فِنال ٹِمسٹر.

Figure 5 : Result obtained when applying the method to a paragraph.

Figure 4 and 5 showcases the result of transliterating input text into Kashmiri with diacritic subjection. The input text, in a English script or language, has been converted into Kashmiri using a transliteration method that incorporates diacritic marks for accurate pronunciation and representation.The user inputs an English word or a clause or a sentence, such as "My Name Is Sameer Ul Rahman," which is then dynamically transliterated into its corresponding form in Kashmiri, as depicted in the image. The input is entered into a text area, and the resulting transliteration is displayed dynamically for user convenience.

7. Conclusion And Future Directions

The primary objective of developing this application is to provide a platform for transliterating English input into Kashmiri output. Leveraging a hybrid approach, the transliterations closely resemble their intended forms. Furthermore, it facilitates access to data on Kashmiri vocabulary, which may be less commonly spoken and thus not readily available. The PHP and JavaScript combination ensures a user-friendly interface for seamless navigation.

7.1 Moving forward, several potential enhancements can be considered:

- Integration of a voice-to-text transliteration feature to enhance user convenience.

- Exploration of incorporating a camera scanner functionality, allowing users to transliterate words captured by the camera without manual input into the application.

- Implementation of a calculator tool to assess the accuracy of transliterated terms, providing users with feedback on the precision of the output.

7.2 Additionally, further research and development could focus on:

- Continuous improvement of transliteration accuracy through refining algorithms and language models.

- Expansion of language support to include additional regional languages, broadening the application's utility and reach.

- Collaboration with linguistic experts and native speakers to validate transliterations and ensure linguistic authenticity.

- Integration of machine learning techniques to enhance transliteration performance and adaptability to diverse input scenarios.

References

[1] Aadil, Mir Asger, M.2017/03/1558 English to Kashmiri Transliteration System - A Hybrid Approach 162 10.5120/ijca2017913418 International Journal of Computer Applications.

[2] Understanding the Processes of Translation and Transliteration in Qualitative Research", 2010 by Krishna Regmi, Jennie Naidoo, Paul Pilkington.

[3] S. Karimi, F. Scholer, and A. Turpin, "Machine transliteration survey," Computing Surveys (CSUR),vol. 43, p. 17, 2011.

[4] K. Naren Sai Krishna, P.Krishnanjaneyulu, K.Mahima, G. Krishna Vamsi, M. Tarun Vishnu, Transliteration of text from english to hindi,Journal of Engineering Science,Vol 11, Issue 4 , April/2020

[5] Malik, Sitender Bawa, Seema Kumar, Munish Phogat, Sangeeta 2021/09/133 A comprehensive survey on machine translation for English, Hindi and Sanskrit languages 14 10.1007/s12652-021-03479-0 Journal of Ambient Intelligence and Humanized Computing.

[6] Sahinur Rahman Laskar, Bishwaraj Paul, Partha Pakray, Sivaji Bandyopadhyay,
English-Assamese Multimodal Neural Machine Translation using Transliteration-based Phrase Augmentation Approach,Procedia Computer Science,Volume 218,2023,Pages 979-988,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2023.01.078.

[7] Harish, B.S., Rangan, R.K. A comprehensive survey on Indian regional language processing. SN Appl. Sci. 2, 1204 (2020). https://doi.org/10.1007/s42452-020-2983-x

[8] Soumyadeep Kundu, Sayantan Paul, and Santanu Pal. 2018. A Deep Learning Based Approach to Transliteration. In Proceedings of the Seventh Named Entities Workshop, pages 79–83, Melbourne, Australia. Association for Computational Linguistics.

[9] A. Diluni De Silva and A. R. Weerasinghe, "Masters Project Final Report (MCS) 2019 Project Title Singlish to Sinhala Converter using Machine Learning Student Name Supervisor's Name S E1 E2 For Office Use Only."

[10] Dhindsa, Baljeet. (2017). ENGLISH TO HINDI TRANSLITERATION SYSTEM USING COMBINATION-BASED APPROACH. International Journal of Advanced Research in Computer Science. 8. 609-613. 10.26483/ijarcs.v8i8.4801.

[11] P. Bhattacharyya, M. M. Khapra, and A. Kunchukuttan, "Statistical Machine Translation between Related Languages," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts, 2016, pp. 17–20, doi: 10.18653/v1/N16-4006.

[12] Noura Farra, Nadi Tomeh, Alla Rozovskaya, and Nizar Habash. 2014. Generalized Character Level Spelling Error Correction. In Proceedings of the Conference of the Association for Computational Linguistics (ACL), Baltimore, Maryland, USA.

[13] Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Atlanta, GA.

[14] A KUMARAN, MITESH, M. KHAPRA and PUSHPAK BHATTACHARYYA, September 2010 Indian Institute of Technology Bombay.Compositional Machine Transliteration.”work done during the author’s internship at Microsoft Research India”.

[15] Rama, Taraka & Gali, Karthik. (2009). Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem. 10.3115/1699705.1699737.

[16] Kumar, Pankaj. “Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns.” (2013).

[17] Dhore, Manikrao & Dixit, Shantanu & Dhore, Ruchi. (2012). Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis. 111-118.

[18] Malik, Muhammad Ghulam Abbas. (2006). Punjabi Machine Transliteration. 1.10.3115/1220175.1220318

[19] Fehri, Hela & Haddar, Kais & Ben Hamadou, Abdelmajid. (2011). Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model. 134-142.

[20] Jaleel, Nasreen & Larkey, Leah. (2003). Statistical transliteration for english-arabic cross language information retrieval. 139-146. 10.1145/956863.956890.

[21] Josan, Gurpreet & Lehal, Gurpreet. (2010). A Punjabi to Hindi Machine Transliteration System. International Journal of Computational Linguistics and Chinese Language Processing. 15. 77-102.

[22] Amarappa S, Sathyanarayana S (2015) Kannada named entity recognition and classification using conditional random fields. In: 2015 International conference on emerging research in electronics. Computer Science and Technology (ICERECT), IEEE, pp 186–191

[23] Dave S, Parikh J, Bhattacharyya P (2001) Interlingua-based english-hindi machine translation and language divergence. Mach Transl 16(4):251–304

[24] Ekbal A, Haque R, Bandyopadhyay S (2007a) Bengali part of speech tagging using conditional random field. In: Proceedings of seventh international symposium on natural language processing (SNLP2007), pp 131–136

[25] Patil N, Patil AS, Pawar B (2016) Survey of named entity recognition systems with respect to indian and foreign languages. Int J Comput Appl 134(16):88

[26] Ravi K, Ravi V (2016) Sentiment classification of hinglish text. In: 2016 3rd international conference on recent advances in information technology (RAIT), IEEE, pp 641–645

[27] Antony P, Soman K (2011) Machine transliteration for indian languages: a literature survey. Int J Sci Eng Res IJSER 2:1–8