Problems of Reading and Transliterating Urdu into Hindi script: An Approach to Human and Machine Transliteration

Dr. Sabahuddin Ahmad
Department of Linguistics
Aligarh Muslim University, Aligarh

ABSTRACT

This paper briefly examines the orthographic structure of Urdu and difficulties involved in reading and transliterating Urdu script by both-human and machine. The Alphabetical inventory of Urdu represents only four vowel graphemes out of total 36 basic characters i.e. [ا،و،ی،ے]. Rest of the vowel sounds are represented by the diacritical marks.
These vowel graphemes are pronounced differently in different position and most of the time these graphemes depend upon the diacritic marks to represent the value of different vowels.

Most of the vowels are represented with the help of diacritic marks which are usually not maintained in regular writing. Absence of these diacritic marks poses lots of problems and error for both human and machine. In Devanagari, there is one to one correspondence between grapheme and the sound. Hindi speakers learning Urdu script face lot of problems in reading and writing Urdu script. The absence of diacritic marks in input text of Urdu also causes errors in the Devanagari output of the automatic transliteration system. (Malik et. al 2009) have shown in their research that the Hindi Word Language Model increases the accuracy of Urdu Hindi transliteration especially for Urdu input without diacritical marks but diacritical marks are crucial and necessary for Urdu to Devanagari transliteration.

Apart from this, a consonant phoneme is realized by a number of graphemes or combination of two graphemes in Urdu. These irregularities cause problems in reading and transliterating Urdu script. Empty graphemes of Urdu script also create problems in transliterating Urdu script into Devanagari script. This paper will explore and explain these problems of Urdu script with special reference to Devanagari.

Key Words: Transliteration, Graphemes, Consonant, Vowel, Diacritical Marks

I. Introduction

Urdu script, known as 'nastaliq' writing, is an adoption of Perso-Arabic script with certain addition and modification. The alphabetic inventory of Urdu, represent 36 basic characters (for some it is 35). Out of these 36 graphemes, there are only four characters/graphemes for vowel sounds, as shown in the table no. 1 below:

Urdu Vowel Graphemes	Name
ا	əlɪf
و	wao
ی	çoʈiː je
ے	bəɽiː je

Urdu hasalso diacritical marks which are either used alone or in combination with the above Urdu vowel graphemes to represent vowel sounds in Urdu script. These diacritical marks are shown in the table no. 2 below:

Urdu Vowel Diacritics	Name of the Graphemes	Vowel Sounds
◌َ	zəbər	/ə/
◌ِ	zer	/ɪ/
◌ُ	peʃ	/ʊ/
◌ٗ	ʊlʈa: peʃ	/u/
◌ٓ	məd	-

'◌ٓ' /məd/ does not occur alone to represent any sound. It is used above 'ا' /əlɪf/ to represent /a:/ sound as in the word **'**آم**'**/a:m/ 'mango'.

The vowel grapheme 'ا' /əlɪf/ in the initial position of the word works as an entity which has no sound in itself; it requires diacritic marks to represent any vowel sound.

I. [ا] 'Alif'+Diacritic: [ا] 'Alif' at initial position with different diacritics represents different vowel sounds:

(a) Lax close unrounded front vowel

[ا] with [◌ِ]= [اِ] - [इ]- /ɪ/ 'اِملی' /ɪmlɪ/

(b) Lax half-open unrounded mid vowel

[ا] with [◌َ]= [اَ] - [अ] - /ə/ 'اَنداز' /əndɑːz/

(c) Lax close rounded back vowel

[ا] with [◌ُ]= [اُ] - [उ]- /ʊ / 'اُلّو' /ʊllʊ/

(d) Tense open unrounded back vowel

[ا] with [~]= [آ] - [आ]- /ɑ/ 'آم' /ɑːm/

Apart from the diacritics, [ا] 'Alif' at initial position also needs other vowel graphemes to represent tense close unrounded front vowel /i/ and tense close rounded back vowel /u/ sound. Without these vowel graphemes these sounds cannot be represented:

II. [ا] 'Alif'+ Another Vowel Grapheme+Diacritic: There are some vowels sounds which need three Urdu characters for their representation in Urdu script.

(a) Tense close unrounded front vowel : /i/

[ا]+[ی] +[◌ٖ]= [ई] - /i/ 'ایٖجاد' /iːdʒaːd/

(b) Tense close rounded back vowel: /u/

[ا]+[و] +[◌ٗ]= [ऊ] - /u/ 'اٗوپر' /upər/

The absence of diacritics from the above letters will change their nature and will be read as /e/ and /o/ instead of /i/ and /u/.

III. [ا] 'Alif'+ Another Vowel Grapheme: Tense half-close unrounded front vowel /e/ and tense half-open unrounded front vowel/o/ are represented by combining two vowel graphemes of Urdu. They do not need any diacritic to represent vowel:

(a) [ا]+[ے]= [ए] - /e/ 'اے' /e/

For example- /ek/ 'ایک', 'एक'

(b) [ا]+[و]= [ओ] - /o/ 'او' /o/

For example- /or/ 'اور', 'ओर'

II. Positional Variation in the form of Vowel Characters

One character i.e. [ا] is common in all the vowel characters discussed in the previous section but the occurrence of this [ا] is limited to the initial position of the word. In medial and final position of the word, [ا] does not occur with the diacritics and other vowel characters which involve in representing particular vowel sound. It is similar like matras e.g. [◌ी, ◌ु etc] of Hindi, which represent vowel sounds in middle and final position of the word. Consider following examples of Table 1:

Table 1

Devanagari Characters	Perso-Arabic Characters	IPA	Position	Examples
ई–ع	اِی	/i/	I	اِیمان 'ईमान' /ɪman/
ई	ئ	/i/	F	کئ 'कई' /kai/
◌ी	ِی	/i/	M	پِیلا 'पीला' /pila/
◌ी	ِی	/i/	F	کہِی 'कही' /kəhɪ/
उ	اُ	/ʊ/	I	اُس '' /ʊs/
◌ु	◌ُ	/ʊ/	M	کُچھ '' /kʊc^{h}/
ऊ	اوٗ	/u/	I	اوٗپر '' /upər/
◌ू	وٗ	/u/	M	جوٗن '' /dʒun/
◌ू	◌ُو	/u/	F	بھالوٗ '' /b^{h}alu/

Urdu script has more variations in its character representation than the Devanagari script, as shown in table 1. In Urdu, /i/ vowel is represented by four different sets of characters depending upon its position of occurrence whereas in Devanagari, it is represented by two characters only i.e. 'ई' and '◌ी'. Similarly other vowels have different forms of character sets according to the position. This is a tricky thing in itself to learn the reading and writing of the Urdu script, though this is not the problem at all for automatic transliteration now. An accurate mapping of the characters provides good output of the transliteration but still there are some features and characteristics of Urdu script which poses problems for both human and machine.

III. Features Posing Problems in Reading and Transliteration Urdu Script

Absence of diacritic marks from the script, multiple consonant characters for single sound, empty letters/grapheme, characters for special Urdu sounds are the major reasons for the complexity of the Urdu script.

3.1. Absence of Diacritic Marks

The diacritical marks, which play a very significant role in Urdu phonology, however, remain unmarked in regular text. Native speaker of Urdu realize them correctly but it is difficult for the non-native speakers or foreign learners. Without the diacritic marks words seem ambiguous to read and sometimes it is unpredictable. This is challenging for both human and machine. Consider the following examples of output taken from SANGAM- automatic transliteration tool (https://sangam.learnpunjabi.org/) developed by Punjabi University, accuracy of which is above 90%:

1. کیا اس کے پاس **جو** ہے۔ تمہارے پاس کتنے کلو **جو** ہے

IPA: kya ʊske pas dʒəɔ hɛ. tʊmhare pas kɪtne kɪlo dʒəɔ hɛ

ST[1]: क्या**उसके**पास**जो**है।तुम्हारेपासकितनेकिलो**जो**है।

IPA: kya ʊske pas **dʒo** hɛ. tʊmhare pas kɪtne kɪlo **dʒo** hɛ

OT: क्याउसके/इसकेपासजौहै।तुम्हारेपासकितनेकिलोजौहै?

"Does he has **barleys**. How many kilo **barleys** do you have?"

2. کیا یہ الوّ ہے؟

IPA: kja yeh ʊllʊ hɛʔ

ST: क्यायेअलो्वहै

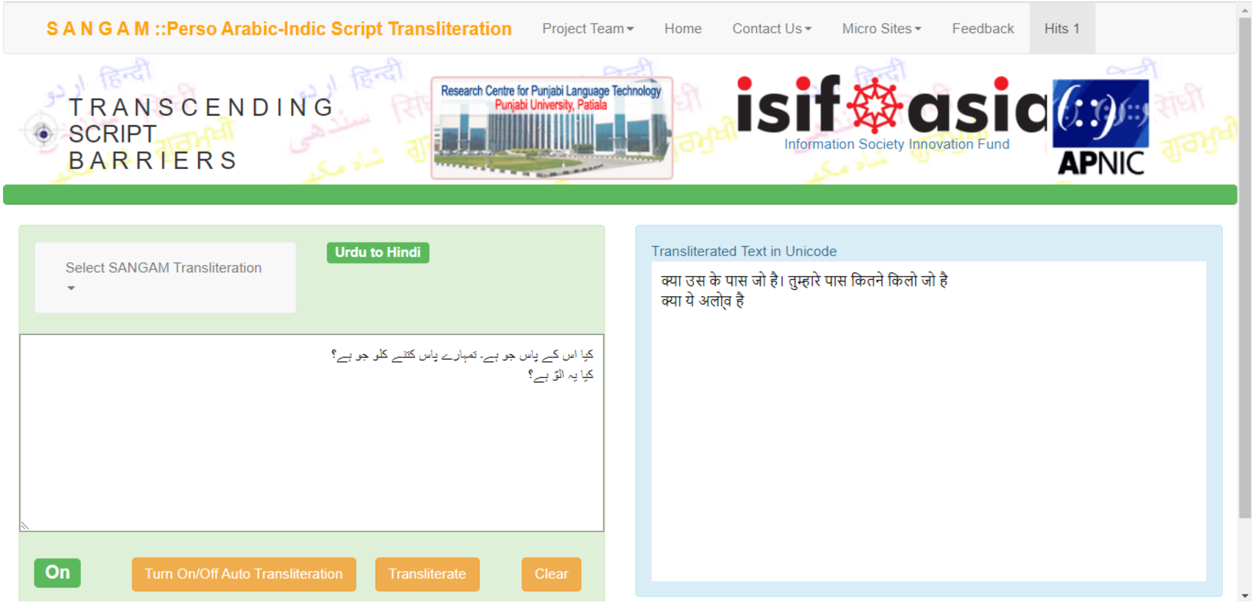
OT[2]: क्यायहउल्लूहै?

"is it an owl?"

In (1) and (2), we can see that two words i.e. 'جو' /dʒəɔ/ and 'الوّ' /ʊllʊ/ are transliterated as 'जो' /dʒo/ and 'अलो्व' /əlov/ respectively which are incorrect. Correct transliterations of these words are 'जौ' /dʒaɔ/ and 'उल्लू' /ʊllʊ/ respectively. It is only because of the absence of the diacritic marks (AIRAAB) for vowels.

[1] ST- SANGAM Transliterator

[2] OT- Original Transliteration

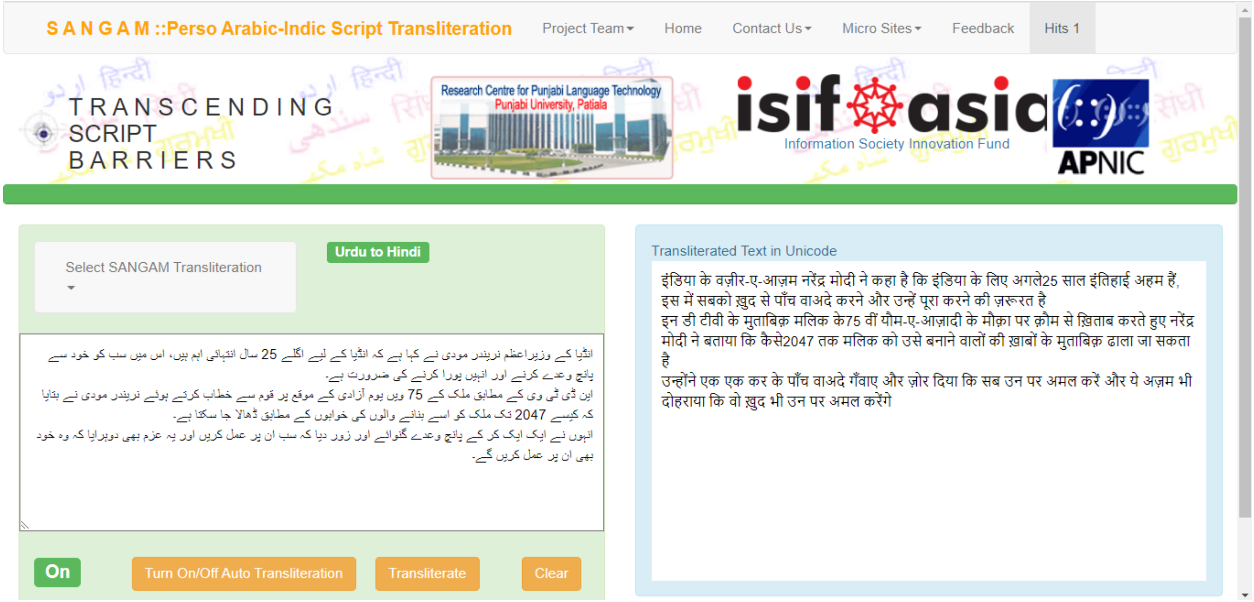

S A N G A M ::Perso Arabic-Indic Script Transliteration
Project Team
Home
Contact Us
Micro Sites
Feedback
Hits 1
TRANSCENDING SCRIPT BARRIERS
Research Centre for Punjabi Language Technology
Punjabi University, Patiala
isif asia
Information Society Innovation Fund
APNIC
Select SANGAM Transliteration
Urdu to Hindi
On
Turn On/Off Auto Transliteration
Transliterate
Clear

کیا اس کے پاس جو ہے۔ تمہارے پاس کتنے کلو جو ہے؟
کیا یہ الوٗ ہے؟

Transliterated Text in Unicode

क्या उस के पास जो है। तुम्हारे पास कितने किलो जो है
क्या ये अलो्व है

Some more result of transliteration by SANGAM


S A N G A M ::Perso Arabic-Indic Script Transliteration
Project Team
Home
Contact Us
Micro Sites
Feedback
Hits 1
TRANSCENDING SCRIPT BARRIERS
Research Centre for Punjabi Language Technology
Punjabi University, Patiala
isif asia
Information Society Innovation Fund
APNIC
Select SANGAM Transliteration
Urdu to Hindi
On
Turn On/Off Auto Transliteration
Transliterate
Clear

انڈیا کے وزیراعظم نریندر مودی نے کہا ہے کہ انڈیا کے لیے اگلے 25 سال انتہائی اہم ہیں، اس میں سب کو خود سے پانچ وعدے کرنے اور انہیں پورا کرنے کی ضرورت ہے۔
این ڈی ٹی وی کے مطابق ملک کے 75 ویں یوم آزادی کے موقع پر قوم سے خطاب کرتے ہوئے نریندر مودی نے بتایا کہ کیسے 2047 تک ملک کو اسے بنانے والوں کی خوابوں کے مطابق ڈھالا جا سکتا ہے۔
انہوں نے ایک ایک کر کے پانچ وعدے گنوائے اور زور دیا کہ سب ان پر عمل کریں اور یہ عزم بھی دوہرایا کہ وہ خود بھی ان پر عمل کریں گے۔

Transliterated Text in Unicode

इंडिया के वज़ीर-ए-आज़म नरेंद्र मोदी ने कहा है कि इंडिया के लिए अगले25 साल इंतिहाई अहम हैं, इस में सबको ख़ुद से पाँच वाअदे करने और उन्हें पूरा करने की ज़रूरत है
इन डी टीवी के मुताबिक़ मलिक के75 वीं यौम-ए-आज़ादी के मौक़ा पर क़ौम से ख़िताब करते हुए नरेंद्र मोदी ने बताया कि कैसे2047 तक मलिक को उसे बनाने वालों की ख़ाबों के मुताबिक़ ढाला जा सकता है
उन्होंने एक एक कर के पाँच वाअदे गँवाए और ज़ोर दिया कि सब उन पर अमल करें और ये अज़्म भी दोहराया कि वो ख़ुद भी उन पर अमल करेंगे

Input Text:

انڈیا کے وزیراعظم نریندر مودی نے کہا ہے کہ انڈیا کے لیے اگلے25 سال انتہائی اہم ہیں، اس میں سب کو خود سے پانچ**وعدے** کرنے اور انہیں پورا کرنے کی ضرورت ہے۔

این ڈی ٹی ویکے مطابق **ملک** کے 75 ویں یوم آزادی کے **موقع** پر قوم سے خطاب کرتے ہوئے نریندر مودی نے بتایا کہ کیسے 2047 تک **ملک** کو اسے بنانے والوں کی **خوابوں** کے مطابق ڈھالا جا سکتا ہے۔

انہوں نے ایک ایک کر کے پانچ**وعدےگنوائے** اور زور دیا کہ سب ان پر عمل کریں اور یہ عزم بھی دوہرایا کہ وہ خود بھی ان پر عمل کریں گے۔

Output Text:

इंडियाकेवज़ीर-ए-आज़मनरेंद्रमोदीनेकहाहैकिइंडियाकेलिएअगले25 सालइंतिहाईअहमहैं,
इसमेंसबकोख़ुदसेपाँच**वाअदे**करनेऔरउन्हेंपूराकरनेकीज़रूरतहै

इनडीटीवीकेमुताबिक़**मलिक**के75 वींयौम-ए-आज़ादीके**मौक़ा**परक़ौमसेख़िताबकरतेहुएनरेंद्रमोदीनेबतायाकिकैसे2047 तक**मलिक**कोउसेबनानेवालोंकी**ख़ाबों**केमुताबिक़ढालाजासकताहै उन्होंनेएकएककरकेपाँच**वाअदेगँवाए**औरज़ोरदियाकिसबउनपरअमलकरेंऔरयेअज़मभीदोहरायाकिवोख़ुदभीउनपरअमलकरेंगे

Table 2: Result Analysis

S.No.	Input Words	IPA	No. of Occurrence in text	Wrong Outputs	Correct words	Reasons
1	وعدے	/vade/	2	वाअदे	वादे	wrong mapping/special character
2	این ڈی ٹی وی	/en di ti vi/	1	इनडीटीवी	एनडीटीवी	wrong mapping
3	موقع	/maɔqa/	1	मौक़ा	मौक़े	different convention of reading
4	ملک	/mʊlk/	2	मलिक	मुल्क	absence of diacritics (AIRAAB)
5	خوابوں	/xwabõ/	1	ख़ाबों	ख़्वाबों	wrong mapping
6	گنوائے	/ginvae/	1	गँवाए	गिनवाए	absence of diacritics (AIRAAB)

We can see in the table 2 that the words of S. No. (4) and (6) i.e. ‘ملک’ ‘country’ / mʊlk/ and ‘گنوائے’ ‘to make someone count’ /ginvae/ respectively are wrongly transliterated as ‘मलिक’ ‘name or title of the name’ /məlɪk/ and ‘गँवाए’ ‘lost’ /gãvae/ due to the absence of diacritic marks. Such errors affect the meaning of the text i.e. sometimes such errors totally change the meaning of a text, causes wrong information and misunderstanding.

3.2 Multiple Consonant Characters for One Sound

Like vowels, consonant sounds also have more than one character or graphemes to represent single sound. In other words, single sound has multiple correspondences in Urdu script but they are not positionally governed. These multiple characters have random distribution.

Table 3

Consonants	IPA	Hindi Alphabets	Urdu Alphabets			
Voiceless Dental Stop	/t/	त	ت	ط		
Voiceless Alveolar Fricative	/s/	स	س	ص	ث	
	/h/	ह	ه	ح		
Voiced Alveolar Fricative	/z/	ज़	ض	ذ	ز	ظ

Urdu has many loan words from Arabic and Persian that include graphemes from these languages, retained in the Urdu spelling. As a result, there are several different Urdu characters mapping to the same phoneme. These graphemes which represent same sound are of different origin so that it is hard to infer any pattern in their distribution. It causes errors in the spelling of words.

Table 4

Urdu Graphemes	IPA	Urdu Words	Words in Devanagari	IPA	Meaning
ت	/t/	تماشا	तमाशा	/təmaʃa/	show
ط	/t/	طشتری	तश्तरी	/təʃtəri/	plate
س	/s/	سانس	सांस	/sãs/	breath
ص	/s/	صبر	सब्र	/səbr/	patient
ث	/s/	ثمر	समर	/səmər/	fruit
ح	/h/	حلوہ	हलवा	/həlwa/	sweet dish
ہ	/h/	ہلچل	हलचल	/həlcəl/	stir or bustle
ض	/z/	مضمون	मज़मून	/məzmun/	essay
ذ	/z/	ذرّہ	ज़र्रा	/zərra/	particle
ز	/z/	مزدور	मज़दूर	/məzdur/	labour
ظ	/z/	ظالم	ज़ालिम	/zalɪm/	cruel

We can see in the table 4 that the identification of correct graphemes which has multiple correspondences is a real challenge while transliterating the Devanagari text into Urdu text. Similarly, it is confusing for the Urdu learners and sometimes for native speakers too. Multiple characters are the major challenge for Devanagari to Urdu automatic transliterator. For the learners, practice is the only tactics to come out this difficulty.

3.3 Empty letters/graphemes

Urdu owing to its Perso-Arabic heritage displays an interesting phenomenon of silence of graphemes where either an individual grapheme or even a sequence of graphemes is silent in Urdu orthography. It means that the sounds of these silent characters are not transferred in the target script.

Table 5

Urdu Words	IPA (Pronunciation)	Letters forming the word	Silent sequence of Grapheme	Automatic Transliteration	Correct Transliteration
علی الصباح	/əlːssəbah/	ع، ل، ی، ا، ل، ص، ب، ا، ح	ل، ی، ا	अलस्सबाह	अल्स्सबाह
شمش الحق	/ʃəmʃʊlhʊda/	ش، م،ش، ا،ل، ہ، د، ا	ا	शमशउल-हक़	शमशुलहक़
قمرالزماں	/qːmrʊzzəmã/	ق، م، ر، ا، ل، ز، م، ا، ں	ا، ل	क़मरअल्ज़मां	क़मरूज़ज़मान
صباح الدین	/səbahʊddin/	ص، ب، ا، ح، ا، ل، د، ی، ن	ا، ل	सबाहउद्दीन	सबाहुद्दीन
اناالحق	/ənəhəq/	ا، ن، ا، ل، ح، ق	ا	अनालहक	अनलहक
قمرالہدا	/qəmrʊlhʊda/	ق، م، ر، ا، ل، ہ، د، ا	ا	क़मरअलहदा	क़मरुलहुदा
قمرُالہدا	/qəmrʊlhʊda/	ق، م، ر، ُ، ا، ل، ہ، د، ا	ا	क़मरुअलहदा	क़मरुलहुदा

3.4 Special Characters

Perso-Arabic Letter	**Devanagari**	**Roman Script**
ع	**--**	**A**
ء	**--**	**‘a**

Mehwish Leghari and Mutee U Rahman Arain, 2015

We can represent these two sounds by the character representing above sound in Roman script by making minor changes using diacritic marks but we don’t find any mean to represent them in Devanagari. They are alphabets, not diacritic marks but they don’t represent any sound alone.

Alphabets representing Common phonemes in Hindi-Urdu		Alphabets representing special Phonemes of Urdu	
ک	क	ق	क़
کھ	ख	خ	ख़
گ	ग	غ	ग़
پھ	फ	ف	फ़
ج	ज	ز، ذ، ض،ظ	ज़

Apart from multiple character we should also focus on the special alphabets representing special sound which is the beauty of the Urdu script

IV. Conclusion

To sum up, it can be said that the absence of ‘Airaab’ from the Urdu script poses major problems of ambiguity in reading Urdu script and transliterating Urdu script into another script e.g. Devanagari. Empty letters i.e. the letters which are not supposed to be pronounced as per the convention of the Urdu script are another source of error while reading Urdu text. They create problems in automatic transliteration as well. Feature of ‘multiple characters for one sound’ may not be much problematic in reading text but creates lots of confusion while writing in Urdu or transliterating any text into Urdu. This paper has highlighted the complexities of the Urdu script and discussed the problems which occur in reading Urdu script and the problems occur in automatic transliteration of Urdu script into any other script.

References

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

[2] Malik, A. et.al. (2009). “A Hybrid Model for Urdu Hindi Transliteration” in the Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pp. 177-185, Suntec, Singapore

[3] Ahmad, Sabahuudin. (2009). *UrduHarf-o-Saut*.2010.Delhi:Arshia Publication,

[4] ________________ (2014). *Imlanama :Takhreeb vo radde takhreeb*. Delhi: ArshiaPublication,

[5] Yamunà Kachru. Hindi. John Benjamins Publishing Co., (2006)