Hindi to Urdu Conversion: Beyond Simple Transliteration

Bushra Jawaid, Tafseer Ahmed
University Of Malta, Malta, Universitaet Konstanz, Germany
bushrajd84@hotmail.com, tafseer@gmail.com

Abstract

This paper incorporates a detailed analysis of existing work on Hindi to Urdu transliteration systems and finds the enhancements they required. It lists the issues that are beyond the scope of character by character mapping. The issues include multiple same sound Urdu characters against one Hindi character. Moreover, it deals with the issues when the same word or words are written in two different ways. The paper lists the differences in pronunciation, spelling and writing style. It presents solution to these issues that goes beyond transliteration.

1. Introduction

Urdu and Hindi are considered as different styles of the same language. These languages share grammar and differ in vocabulary and writing script. Urdu uses more Arabic and Persian words and is written in Nastaleeq script. Nevertheless, Hindi uses more Sanskrit words and is written in Devnagri script.

In conversation, Urdu and Hindi are intelligible. Television programs and cinema films are watched in the both languages communities without the need of translation. A Pakistani Urdu speaker understands the Indian Hindi films, and an Indian Hindi speaker understands Urdu programs. The problem arises when a person tries to read written text of the other language. Most of the people cannot read script of the other language.

A considerable amount of work is done on Hindi to Urdu transliteration. CRULP [1] and Malik [2] has discussed and implemented issues of Hindi to Urdu transliteration. There are two fundamental goals of this paper. The first goal is to find problems / short comings in the models / implementations of [1] and [2], and to propose solutions of these problems. The second goal is to find whether any accurate character by character Hindi to Urdu transliteration will be enough for the Urdu reader to read transliterated Hindi text comfortably or we need to deal the issues which are beyond transliteration.

2. Hindi-Urdu transliteration: a brief review

It has already mentioned that both languages use different scripts for writing. Here we discuss these scripts briefly.

Hindi is written in devnagri script and it is read and written from left to right. All consonants in Hindi inherit [ə] sound. All the vowels in Hindi are attached to the top or bottom of the consonant or to an [◌ा] vowel sign attached to the right of the consonant, with the exception of the [ि◌] vowel sign, which is attached on the left [5]. Hindi has 29 non-aspirated, and 15 aspirated consonants, and 11 vowels (*svara*) [2]. A syllable (*akshara*) is formed by the combination of zero or one consonants and one vowel. [5]

Nastalique script is read and written from right-to-left. Nastalique, a cursive, context-sensitive and a highly complex writing system, is widely used for the Urdu orthography. The shape assumed by a character in a word is context sensitive. The Urdu alphabet contains 35 simple consonants, 15 aspirated consonants, one character for nasal sound, 15 diacritical marks, 10 digits and other symbols. [2]

Below is the consonant chart for Hindi and its respective Urdu character.

Table 1: Mapping of Hindi and Urdu consonants

Devnagri Consonants		Urdu Consonants	
Letter	Name	Letter	Name
क	KA	ک	Kaaf

ख	KHA	کھ	Kaaf-Hay
ग	GA	گ	Gaaf
घ	GHA	گھ	Gaaf-Hay
ङ	NGA	ن	Noon
च	CA	چ	Chay
छ	CHA	چھ	Chay-Hay
ज	JA	ج	Jeem
झ	JHA	جھ	Jeem-Hay
ञ	NYA	یاں	
ट	TTA	ٹ	Ttay
ठ	TTHA	ٹھ	Ttay-Hay
ड	DDA	ڈ	Ddaal
ढ	DDHA	ڈھ	Ddaal-Hay
ण	NNA	ڑاں	
त	TA	ت/ ط	Tay/ Toay
थ	THA	تھ	Tay-Hay
द	DA	د	Daal
ध	DHA	دھ	Daal-Hay
न	NA	ن	Noon
प	PA	پ	Pay
फ	PHA	پھ	Pay-Hay
ब	BA	ب	Bay
भ	BHA	بھ	Bay-Hay
म	MA	م	Meem
य	YA	ے	Bari-Yeh
र	RA	ر	Ray
ल	LA	ل	Laam
व	VA	و	Wow
श	SHA	ش	Sheen
ष	SSA	ش	Sheen
स	SA	ث/ س/ ص	Seen/ Saay/ Suad
ह	HA	ح/ ھ/ ہ	Hay

Below is the Vowel chart for Hindi and Urdu.

Table 2: Mapping of Hindi and Urdu vowels

Hindi Vowels		Urdu Vowels	
Letter	Diacritical Mark	Letter	Vowel
अ		ا	ə
आ	◌ा	آ	a
इ	◌ि	ِ	ɪ
ई	◌ी	ی	i
उ	◌ु	ُ	ʊ
ऊ	◌ू	وُ	u

ऋ	ृ	ر + ِ (consonant + vowel)	r̥
ए	े	اے	e
ऐ	ै	اَے	æ
ओ	ो	اُ	o
औ	ौ	اُو	ɔ

In Table 2 we have listed all the Urdu vowel symbols or group of vowels against each Hindi vowel to represent Hindi vowel sounds. Only exception is vowel "ऋ" whose vowel sound maps on Urdu consonant and vowel character sounds. Current transliteration systems don't provide support for the independent form of this vowel. CRULP's output for the sample word ऋषि (rishi) is "ऋش".

Here we are writing down few sample words by reading those a reader can have an idea of difference in writing style of both languages.

दुकानदार	دُکاندار
नमस्ते	نمستے
टमाटर	ٹماٹر

3. Issues in Hindi-Urdu conversion

The paper discusses the issues in Hindi to Urdu conversion that are remained unsolved in CRULP's and Malik's system. To identify these problems we made a small survey of Hindi text available at [3] and [4]. We transliterated the Hindi text to Urdu using CRULP's Hindi to Urdu transliterator. The problems identified in the converted text are listed. We explored Malik's solution to find whether his algorithm and structures have solution of these problems. It was found that most of the problems are not solved by his model too.
The identified issues are of three types. The first type of issues has unsolved problems in character by character transliteration of Hindi text into Urdu script.
But there are issues that are beyond the scope of character to character transliteration. There are differences in writing style and vocabulary. We are presenting list of all these issues in the following text. Most of the Hindi presented in the following discussion is taken from [3] and [4] and few example sentences are constructed. In section 4, we will present solutions for these issues.

3.1. Transliteration between different scripts

After transliterating the Hindi text by exploiting the CRULP's transliterator and later on comparing those results with the expected output of Malik's system, we found following issues that remained unsolved in either one or both of the systems.

3.1.1. Same/similar sound character. In the following example sentence, the word "غلتی" is not transliterated correctly. In problem-word "t" sound character should be transliterated into "ط" instead of "ت".

(1) यह दोनों की **ग़लती** है

یہ دونوں کی **غلتی** ہے

Similar sound character problem always occurs due to multiple Urdu characters against one Hindi character, as can be seen in table 1. For the same reason, the wrong selection of character has often found for words that end on "ہ".

(2) तुम खाना **हमेशा** बहुत अच्छा बनाती हो

تُم کھانا **ہمیشا** بہُت اَچّھا بناتی ہو

In (2), for example, word "ہمیشا" is written with "ا" instead of "ہ". Table 4 gives the list of same sound Urdu characters.

Table 4: List of same sound Urdu characters

Sounds	Urdu Characters list for each sound Character	Default Characters for Transliteration
t sound	ط , ت	ت
s sound	ص, س, ث	س
z sound	ز, ذ, ض, ظ	ز

a sound	ا, ع, ء	ا
a sound	ا, ع, ء, ہ -at-end	ا

Systems that are built for Hindi to Urdu transliteration currently have fixed transliteration rules defined for same sound character mapping. Those rules map Hindi's same sound characters on default Urdu characters as defined in Table 4.

3.1.2. Characters similar in Shape. Transliteration errors that occasionally occur are primarily due to the charcters that are exactly identical in shape in Devnagri script and differs only by a dot addition. Errors are rarely found because of the missing dots and mostly due to the pronunciation differences between the speakers of the both languages.

(3) भारती **ग़ज़लीं** बहुत पसन्द की जाती हैं

بھارتی **گزلیں** بہت پسند کی جا تی ہی

Table 5: Chart of similar shape Hindi characters

Urdu	Hindi	Transliteration Errors
ز , ج	ज़, ज	زبردست → جبردست
خ , کھ	ख़, ख	خوبی → کھُوبی
ف , پھ	फ़, फ	فضول → پھِزُول
غ , گ	ग़, ग	دِماغ → دِماگ

3.1.3. Nasalized sound character. In Urdu consonants chart we have a single character to represent nasalized sound known as "noon-ghunna" (ں).

Hindi script has "chandrabindu" (◌ँ) and "bindu" (◌ं) diacritics to represent nasalized sound. There are two problems in mapping of chandrabindu/bindu to Urdu script. The first problem arises when this nasalized sound character occurs in the middle of the word, as shown in (4):

(4) यह **इंडिया** गेट है

یہ اِںڈِیا گیٹ ہے

In Urdu when noon-ghunna comes in the middle of the word it is replaced by noon. Current transliteration systems map Hindi nasalized characters with noon-ghunna of Urdu irrespective of its position in the word.

The second issue is that few words in Urdu contain character "م" but in pronunciation they produce nasalized sound.

(5) जीटी रोड की **लंबाई** कतनी है?

جی ٹی روڈ کی **لںبائی** کتنی ہے؟

Hindi speakers write these words the way they pronounce it. That's why in the result of transliteration we get "noon-ghunna" instead of "م", as in (5).

3.1.4. Kasr-e-Izafat Issue. Kasr-e-Izafat is represented by (Zer) at the end of a word and is used to connect two words to form the compound word e.g. تاریخ پیدائش

Words having izafat symbol produces [e] sound effect during pronunciation. In devnagri script there is no concept of izafat to produce [e] sound that's why Indian native speakers use diacritical mark [◌े] whose independent form is [ए] in place of the diacritical mark [ि◌] whose independent form is [इ] for writting these words.

(6) **मीनारे** पाकिस्तान लाहौर में है

مینارے پاکِستان لاہورمیں ہے

Thus, this wrong diacritical marking as in (6) produces (ے) instead of Izafat sign (Zer). Solution of the above is not present in either of the two systems.

3.2. Different writing style

Even if a character by character mapping is modeled successfully, there remain few differences in writing conventions of Urdu and Hindi. These problems are beyond the scope of transliteration, and hence are not discussed in the two earlier works [1] and [2], but these should be addressed because the Urdu reader expects to read the text having Urdu conventions.

3.2.1. Native words. There is a difference in writing conventions of native Indic words in Hindi and Urdu. Problem has been found in those words which end up on vowel sound. Hindi language can have words that end on short vowel but Urdu language doesn't contain words that end on short vowel.

The first issue arises for the words which end up on short vowel character. In Urdu writing system when a short vowel comes at the end of the word it is written as a long vowel, as illustrated in example (7):

(7) नई दिल्ली में पार्लियामैण्ट और **राष्ट्रपति** भवन है

نئی دِلّی میں پارلِیامینٹ اؤر**راشٹرپتِ** بھون ہے

Observe that the word "راشٹرپتِ" has short vowel ending. CRULP's transliterator has not dealt with this issue whereas Malik [2] has presented his idea on a particular problem but this problem is also present in his system.

Hindi and Urdu share a basic common vocabulary that includes pronouns and auxiliaries. Errors in transliteration occur because of different writing conventions which have been followed for words that are being used by both language speakers from same set of vocabulary. As shown in (8): a pronoun "یہ" is transliterated as "یے".

(8) **ये** मरगिला की पहाड़ियाँ हैं

یے مرگِلا کی پہاڑِیاں ہیں

This issue is often found in words "یہ", "کِہ", "پَہ" and "وُہ". First three in pronunciations produce [e] sound that's why they are written with "ے". However, the pronoun "وُہ" is exception. It has been found in two different writing styles in today's Hindi text i.e. "وو" and "وے" which is transliterated form of "वो" and "वे".

3.2.2. Short vowel for Vao. Words adopted by Hindi speakers from Urdu (Persio-Arabic) vocabulary have issues like wrong marking of "و" with short vowel pesh (ُ◌). This problem does not occur for all instances of vao in Urdu.

(9) क्या **हुआ**?

کیا ہُآ؟

Same problem has been found in words borrowed from English vocabulary.

(10) मैं अभी **पुलिस** बुलाती हूँ

میں اَبھی **پُلِس** بُلاتی ہُوں

In example (10) "پُلِس" is the transliteration of English word "police". This transliteration is correct according to its pronunciation but Urdu speakers write "police" using long vowel: as "پُولیس".

3.2.3. Persio-Arabic words. Urdu's vocabulary is comprised of many Persio-Arabic words. These borrowings have their own writing conventions which Urdu speakers generally follow.

When Hindi speakers write persio-arabic words they don't follow the writting scheme of these borrowings. Issues have been seen in the words that have silent-Alif. Hindi speakers neglect silent-alif part in writing because of the unawareness with the correct form of the word. The lack in familiarity with the correct form of the words is depicted as in (11):

(11) में **फ़िलहाल** अपने घर जा रहा हूँ

میں **فِلہال** اَپنے گھر جا رہا ہُوں

Problem-word in example (11) is "فِلہال" and its correct form is "فی الحال".

Another problem occurs with the persio-arabic words with "خ" sound. When "خ" is followed by [u] sound, they are mostly written with short vowel (ُ) (equivalent of Pesh) in Hindi text. This issue is discussed in detail in 3.2.2

(12) मुझे **ख़ुशी** है कि तुम मेरे पास हो

مُجھے **خُشی** ہے کِ تُم میرے پاس ہو

3.3. Multi-morphemes without space

If we analyze Hindi text we would find out frequent use of multi-morphemes words. These words are typically associated with the different problem-groups. Below we have discussed these groups and have morphologically correlated Hindi and Urdu grammars.

3.3.1. Case markers after pronouns (+ 'sab'). In Hindi and Urdu sentence structures we use case markers. These markers can be identified grammatically as non-semantic prepositions. In Hindi text, these case markers when use with the pronouns disowns their individual form and coalesce with the pronoun.

(13) **तुमने** फ़िल्म कभी नहीं देखी?

تُمنے فِلم کبھی نہیں دیکھی؟

In (13) case marker "نے" is joined with a pronoun "تُم". Nevertheless, the word "سب" is not a pronoun but it also follows the same principle. Hindi writers also write it in conjunction with the case markers. This is illustrated by the example in (14):

(14) **सबसे** ज़्यादा गर्मी राजस्थान में पड़ती है

سبسے زیادا گرمی راجستھان میں پڑتی ہے

In the above example, "سبسے" has subjected to the unnecessary joining.

3.3.2. gaa / gii / ge after verb. Hindi and Urdu grammar share the same auxiliaries to define aspect and tense. In Urdu language, auxiliaries and basic-verbs are written standalone; however, in Hindi text the verbs are concatenated with the future auxiliaries "gaa", "gii" and "ge".

(15) क्या अब मेरी पत्नी मुझे पहचान **सकेगी**?

کیا اَب میری پتنی مُجھے پہچان **سکیگی**؟

The verb "سکے" in (15) is followed by the future marker "گی". These two are written as a single word i.e. "سکیگی" and Urdu speakers are mostly not familiar with this form of the verb.

3.3.3. kar/ke after verb. Both Hindi and Urdu languages include conjunctions. These conjunctions are basically used to join two words or phrases together. Urdu language considers these conjunctions as individual word forms thus they are written separately. While in Hindi content conjunctions "kar" and "ke" are joined with the preceding verb.

(16) फिर याद **चलकर** मेरे पास क्यों आती है

پھِریاد **چلکر** میرے پاس کیوں آتی ہے

In example (16) "چلکر" is problem-word in which "کر" has been connected with "چل" and this association has created new form of word which is neither a part of Urdu word list nor familiar for Urdu readers.

3.3.4. Compound words. Urdu language has compound words in its vocabulary. These words are based on multiple constituent words. These constituent words when written in compound form give new essence to the word. In Urdu, the general rule of writing compound word is to put space among the constituents however Hindi writers usually don't follow this convention; they write it by connecting the constituents together.

(17) रानी टैलिफ़ोन पर चाची से **बातचीत** कर रही है

رانی ٹیلِفون پر چاچی سے **باتچیت** کررہی ہے

Urdu glossary contains compound word "بات چیت" which is consisted of two consitutents. As we can see in (17) these constituents are merged together hence generated new construction for the same compound word: "باتچیت".

3.4. Difference in vocabulary

Although both languages share a lot of words that includes native (Indian), Arabic, Persian, and English loanwords but still Hindi and Urdu words differ at some places because one is using the words having Arabic or Persian origion while other is using the Sanskrit loanwords. Differences have been noticed in a close class words and also in a general vocabulary.

Table 6: Examples of difference in vocabulary in Urdu and Hindi

Close Class Words		Open Class Words	
Hindi	Urdu	Hindi	Urdu
دوارا	سے	ایتِہاسِک	تاریخی
کارن	وجہ سے	راجنیتی	سیاست

There is a difference in the use of pronoun "وُہ". In Urdu, it is used for both singular and plural. But in Hindi "وُہ" is used for third person singular and "وے" is used as third person plural.

4. Proposed solutions

We have listed all those problems in section (3) which we came across during transliteration process. Now we present the solution of these problems. We propose to use the available system i.e. CRULP's Hindi to Urdu transliterator as part of the solution. We introduce pre-processing (on Hindi text before transliteration) and post processing (on transliterated Urdu text) to get better transliteration.

The enhancements that are required to construct an improved transliteration system are given below.

4.1 Multi-morphemes without space

As presented in 3.3 and 3.2.3, there are multi-morpheme words that are written without space in Hindi, but in Urdu there is a space between the morphemes of these words. We need a solution that introduces the required space in transliterated Urdu word.

We are presenting the solution which can intensify transliteration result using operational systems. For the problem-groups discussed below, we will apply proposed techniques on transliterated text i.e. on Urdu lexicon.

4.1.1. Case markers after pronouns (+ 'sab'). To solve the issues presented in 3.3.1, we will maintain a separate list of pronouns, case markers and exception words discussed earlier to identify words that are joined with the case markers. To distinguish those words we will perform a character by character matching and as soon as a group of characters matches with any of the pronoun in the pronoun list we start checking for the remaining part of the word in a case markers list. If it matches successfully with any of the case marker in the list, both words gets splitted.

ارا + ہم = ہمارا (don't break)
نے + ہم = ہمنے (break)

4.1.2. kar/ke after verb. To classify verbs that are written in merged form with the conjunctions (as discussed in 3.3.3), we need to maintain a separate list of verbs in the root form e.g. "chal", "aa", "parh" etc. We will have a root-verb-conjunction list too that will have two words "kar" and "ke". We will start with character by character matching of the word in verb list; if we find a match for the verb initially then we start comparing rest of the characters in conjunction list. If both of these character groups of the word are successfully matched in their respective list, a space will be inserted between verb and conjunction.
It is important to note that we need a list of only the root form of the verbs because kar/ke comes only with the root form of the verb. For instance, aa-kar is correct, but aaye-kar is not.

4.1.3. gaa / gii / ge after verb. We discussed the issue of future marker in section 3.3.2. This issue is only related to the verbs in subjunctive form e.g. آئیں, جائیں, کھائیں. To resolve it, we require a list of subjunctive form of the verbs to detach a verb from auxiliaries' گا / گے / گی.
We need list of subjunctive verb because parhen-ge is being used in both languages, but parhte-ge is not. The solution of this issue will follow the same matching technique as described in 4.1.1 and 4.1.2

4.1.4. Multi-morphemes. To disambiguate multi-morphemes words that are written without space in Hindi, we will maintain a list of compound words. For creating this list we will select all the words from the available word list which have space among constituents. We will then remove all the spaces in between a compound word and will put the word without spaces and word with spaces in newly created compound words-list.

The transliterated word will be searched in the compound words-list, having both forms of the word; if a word matches with any entry in the list, the corresponding original compound word with spaces will be returned.

4.2. Word to word mapping

In the above section, we introduced the solutions in which we can find the need of a missing space in transliterated Urdu text using different kind of a word lists. This approach solves many problems, but there are remaining set of problems which do not have generalized solutions. These problems need a simple word to word mapping between Hindi words and Urdu equivalents.

In a word to word mapping we get wrong transliteration because of the words that have different spellings in both languages. Solution to this problem is to maintain a list of words that differ in spelling in both languages. This list will contain a proper Urdu lexicon against each Hindi word. Before transliteration we will map a word in the list, if a list contain that word, Urdu word with proper transliteration already defined in the list against problem-word will be selected and transliteration process will not be carried out for that particular word.

This method will solve the issues introduced in 3.2.1 (partially), 3.2.2 and 3.2.3 (partially).

4.3. Problems to be solved

The problem of same sound characters needs extensive efforts. The problem can be solved by generating all possible transliterations of the Hindi word having same sound characters. These candidate transliterations will be matched with a Hindi word list. We are not proposing the details of this solution here, but one can give an algorithm similar to "roman to Urdu transliteration using word list" algorithm presented in the same conference [7].

We can not find the optimal solution of 3.1.4 i.e. kasr-e-izafat problem. This problem can be partially solved through compound-word-list approach given in 4.1.4. It is observed that an average Hindi speaker use very few words that have kasr-e-izafat. It is not productive in Hindi and usually new words and phrases are not constructed by using kasr-e-izafat in Hindi. It is highly probable that if a Hindi writer used a word having kasr-e-izafat, it will be in an Urdu lexicon. If the lexicon has spelling/pronunciation too like [8], we can make a list of the compound words having two words and an izafat between them. If any transliterated Hindi word ends on bari-ye then the "word-space-next word" will be searched in the izafat-compound-word-list, if found then the word from the izafat-compound-word list will be the output.

Vocabulary is still a problem, as Urdu readers are not aware of all the words (specially the Sanskrit words) that are used in Hindi.

5. Conclusion

After reviewing the output of existing Hindi to Urdu transliteration system, we find that there are issues that are still needed to be solved. The main issues for a better transliteration are: slight difference in both languages script, difference in spelling of the same words due to unawareness with correct form or due to diverse writing style, alteration in original forms of the words (commonly being used by both languages chroniclers) due to variation in a pronunciation of the words. We suggested post processing of the transliterator output and solved issues caused by writing convention differences, by consulting specific word lists.

6. Acknowledgement

The authors are thankful to Tikaram Poudel for his help in analysis of Hindi script and identification of problems.

7. References

[1] *Hindi to Urdu Transliterator*, Center for Research in Urdu Language Processing (CRULP), Lahore. Available:http://crulp.org/software/langproc/h2utransliterator.html

[2] Malik, M. G. Abbas, *Hindi Urdu Machine Transliteration System*, MS thesis, Department of Linguistics, University of Paris 7, Paris, 2006.

[3] BBC Hindi website.
Available: http://www.bbc.co.uk/hindi/

[4] T. Afroze, *A door into Hindi*, website, North Carolina State University, Raleigh NC.
Available: http://taj.chass.ncsu.edu

[5] *Devnagri Script,* website.
Available: http://www.haryana-online.com/devnagri.htm

[6] *Urdu*, website.
Available: http://www.haryana-online.com/urdu.htm

[7] T. Ahmed,. "Roman to Urdu transliteration using wordlist", submitted to *Conference of Language and Technology 2009*, Lahore, 2009.

[8] *Urdu Online Dictionary,* website.
Available: http://www.crulp.org/oud/default.aspx