Issues in Hindi to English and Marathi to English Machine Transliteration of Named Entities

M L Dhore
BRACT's Vishwakarma Institute of Technology, Pune

S K Dixit
Walchand Institute of Technology, Solapur, India

Ruchi M Dhore
PVG's College of Engineering and Technology, Pune, India

ABSTRACT

Almost all transactions ranging from various domains such as travel, shopping, insurance, entertainment, hotels, appointments etc. are available through Internet based applications. Needless to say, all these applications require the knowledge of English. As Internet users are growing day by day, it is logical to say that, there is a great demand to develop tools and applications to support Indian languages for them. The solution to provide local language support in the web based commercial applications is Machine Translation which can be used to translate static labels on web form and Machine Transliteration to transliterate dynamic user inputs from local language into the default language English. It is challenging to transliterate names and technical terms occurring in the user input across languages with different alphabets and sound inventories. This paper focuses important issues which frequently occur in Hindi to English and Marathi to English named entities machine transliteration.

General Terms

Machine Transliteration

Keywords

Allophones, Language of Origin, Multiple Transliterations, Orthography, Phonology

1. INTRODUCTION

For the last two decades, most of the work has been carried out by using English as a source language and Asian languages as the target languages. This work focuses on the specific issues of machine transliteration of Hindi to English and Marathi to English which are previously less studied language pairs. Machine transliteration is usually used to support the machine translation (MT) and cross-language information retrieval (CLIR) to convert the named entities.

India is a multilingual country with 22 constitutional officially recognized languages and 11 different scripts used in different regions spread across the country. Twenty two constitutional Indian Languages are: Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri, Maithili, Sindhi, Kashmiri, Santhali, Assamese, Manipuri, Bangla, Oriya, Gujarati, Punjabi, Telugu, Kannada, Tamil, Malayalam and Urdu [1]. The earliest writing in India was in the ancient Brahmi script, as seen in rock inscriptions. Most of the Indian languages (Indo-Aryan and Dravidian) use scripts which have evolved from the ancient Brahmi script. As a result the orthographic rules for writing text in these scripts are more or less the same, even if their scripts are totally distinct. These scripts are cumulatively referred to as "Indic Scripts". In India, there are 1,650 dialects but till today English continues to be ubiquitous as a mode of communication in higher education, judiciary, bureaucracy, and the corporate sector. Yet, a mere 5 percent of the population speaks English as their first language and a marginally higher percentage is comfortable with it as the primary mode of communication. Against this, the 2001 census figures suggest that close to 40 percent of the Indian population speaks Hindi or one of its variants as their first language. Despite the plurality of languages and scripts, the underlying grammar is largely similar across languages with a common vocabulary of 40 to 80 percent. An interesting point to note is that only 17.25 percent of the total population in the country can read or write in the English language [2]. Marathi is also one of the widely spoken languages in India especially in the state of Maharashtra. Both Hindi and Marathi use the "Devanagari" script and draw their vocabulary mainly from Sanskrit. It clearly indicates the need of Hindi to English and Marathi to English machine transliteration to support web based applications.

2. RELATED WORK

Transliteration of words from a source language to a target language is mostly considered to be a problem of generation of target language equivalents. The major techniques for transliteration can be broadly classified into two categories, viz grapheme-based and phoneme-based approaches. In the year 1998, Knight presented a phoneme-based statistical model using finite state transducer and transformation rules [3]. In the year 2002, Yaser and Knight used a grapheme-based approach that maps English letter sequences to Arabic letters [4]. In 2003, Abdul Jaleel and Larkey developed a simple, statistical technique for building an English- Arabic transliteration model using Hidden Markov Model [5]. In the same year, Goto used a maximum entropy based model for English-Japanese transliteration which predicts the Japanese equivalents for English chunks using their contextual information probabilities [6]. In 2004, Li presented a joint source-channel n-gram model for English-Chinese transliteration using orthographic alignments obtained from an English-Chinese bilingual dictionary [7]. In the context of Indian languages, Aswani and Gaizauskas (2005) have used a transliteration similarity based technique to align English-Hindi parallel texts. They used character based direct correspondences between Hindi and English to produce possible transliterations [8].

In 2006, Katre has discussed cross-cultural usability issues of Hindi and English for the mobile phones [9]. In 2008, Janet Pierrehumbert and Rami Nair have discussed the issues of the allophones for speech to text conversion [10]. In the same year, Karthik Gali has brought into notice many linguistic issues for named entity recognition for Indian languages [11]. In 2009, Abbas Malik has noted the issues of native sounds and native Hindi spellings for Urdu Hindi Transliteration [12]. In the same year, Ankit Aggarwal provided the error analysis due to foreign origin and half consonants for English to Hindi transliteration [13]. In 2010, Manoj Kumar Chinnakota and Om P. Damani discussed the issues of multiple transliterations, origin misclassification and schwa related errors for Persian to English transliteration [14]. In 2011, Sarvnaz Karmi discussed the common challenges in machine transliterations with respect to script specification, missing sounds, transliteration variants [15]. In the same year, Umair has provided the challenges in designing input method editors for Indian languages considering back-transliteration [16]. The direct transliteration of Hindi to English is quite difficult due to the various factors. This paper describes the practical issues related with Hindi to English machine transliteration with respect to script specification, missing sounds, transliteration variants, native sounds, multiple transliterations, loan words, language of origin, spelling variations and pattern of suffixes.

3. ISSUES IN TRANSLITERATION

For the automatic or direct transliteration of Hindi to English named entities following sixteen issues are found.

- Script Specification
- Missing Sounds
- Multiple Transliterations
- Spelling by orthography or Phonology
- Allophones
- Spelling Variations
- Capitalization
- Language of Origin
- Short Vowel and Long Vowel
- Conjuncts
- Affixes
- Acronyms
- Loan Words
- Incorrect Syllabification
- Consonant ‘r’ in the conjuncts
- Schwa Identification and Deletion

3.1 Script Specification

The possibility of different scripts between the source and target languages is the problem that transliteration systems need to tackle. Hindi uses the Devanagari script whereas English uses the Roman script. Devanagari script used for Hindi has total 12 pure vowels and two additional loan vowels taken from the Sanskrit and one loan vowel from English while according to Cambridge Advanced Learner's Dictionary English has only five pure vowels but, the vowel sound is also associated with the consonants w and y [17]. Table 1 show 15 vowels used in Devanagari script along with their matra signs.

Table 1. Vowel and its Matras in Devanagari

Vowel	Matra	Vowel	Matra
अ	No matra	ॠ	◌ॄ
आ	◌ा	ए	◌े
इ	◌ि	ऐ	◌ै
ई	◌ी	ओ	◌ो
उ	◌ु	औ	◌ौ
ऊ	◌ू	अं	◌ं
ऋ	◌ृ	अ:	◌:
Loan Vowel		ऑ	◌ॅ

There are 34 pure consonants, 5 traditional conjuncts, 7 loan consonants and 2 traditional signs in Devanagari script and each consonant have 14 variations through integration of 14 vowels while in Roman script there are only 21 consonants. The 34 pure consonants and 5 traditional conjuncts along with 14 vowels produce 546 different alphabetical characters [18]. Table 2 shows the 34 pure consonants in Devanagari script. The consonant /ळ/ is available only in Marathi language.

Table 2. Pure Consonants in Devanagari Script

क	ख	ग	घ	ङ
च	छ	ज	झ	ञ
ट	ठ	ड	ढ	ण
त	थ	द	ध	न
प	फ	ब	भ	म
य	र	ल	व	श
ष	स	ळ	ह	

Table 3 shows the 5 traditional conjuncts, 7 loan alphabets, 2 traditional signs and 2 special nasal signs in Devanagari script [19-21].

Table 3. Supplementary Consonants and Signs

Traditional Conjuncts	क्ष त्र ज्ञ श्र द्य
Additional Consonants	ड़ ढ़
Loan Consonants	क़ ख़ ग़ ज़ फ़
Traditional Signs	ॐ श्री
Special nasal	◌ं ◌ँ

3.2 Missing Sounds

All the languages in the world have their own phonetic structure, and symbols of the language script correspond to different phonetics. If there is a missing phonetic in the letters of a language, single phonetic are represented using digraphs and tri-graphs. Transliteration systems need to take care of both the convention of writing the missing phonetics in each of the languages involved in transliteration, and the method to exporting the phonetics from one language to the other.

For the few vowels and consonants in Devanagari script, there are no equivalents available in English. In such cases, back transliteration becomes difficult to get correct word in the source language.

3.2.1 Native Vowel Sound

Devanagari vowel /ऋ/ (ṛ) has no direct equivalent in English and is pronounced somewhere in between /ri/ and /ru/, as /ri/ sounds in the English word crystal and as /ru/ sounds in the English word crucial. Similarly, /ऌ/ is also like /ऋ/ and is pronounced somewhere in between /li/ and /lu/ similar to glycerin. Following forward transliteration of two Devanagari words is same.

/रितु/ (person name) is transliterated as /Ritu/ as well as /ऋतु/ (season) is also transliterated as /Ritu/. As the transliteration is same for both the words, it creates the confusion for back-transliteration whether to map /Ri/ to /रि/ or /ऋ/.

3.2.2 Retroflex and Dental Plosive

Hindi consonants /त/ (dental plosive) and /ट/ (retroflex) both are transliterated as /t/ in English. The consonant /t/ in English has more closer sound to /ट/ in Hindi. If it the case then there is no direct equivalent in English for the /त/ in Hindi. It creates a problem in back transliteration from English to Hindi whether to map a /t/ in English to /त/ or /ट/ in Hindi. The same problem occurs for the retroflex /ठ/, /ढ/, /ण/ and dentals /थ/, /ध/, /न/. For example

- The same ambiguity occurs for the surname /धोरे/ and /ढोरे/ where both are transliterated as /Dhore/ in English. The first author of this paper is /Dhore/ ('ढोरे') whose name is always displayed on Indian Railway's reservation chart as /धोरे/.

3.2.3 Plosive Sibilants

The sibilant /श/ (sh) in Hindi has no direct equivalent in English. The consonant /श/ sounds in the English word shut. Similar is the case for another sibilant /ष/ which has also no direct equivalent in English and it sounds in the English word shall. For example

- The words /शतक/ (Century) and /षटक/ (over in cricket) both are transliterated as /Shatak/.

3.3 Multiple Transliterations

The fundamental problem in India is that there are no sets of rules available to create the spellings according to the linguistics. People are writing different spellings for the same name at different places. This fact certainly affects the Top-1 accuracy of the transliteration process. It seems to be very difficult to develop a system which can really provide 100% accuracy. To generate multiple transliteration candidates for the same name following mapping table 4 is used.

Table 4. Mapping Table for Multiple Transliterations

Devanagari Consonant	English Equivalent	Top Preference for Hindi Transliteration
क	k , c, q	k
ख	k, kh	kh
ग	g, gh	g
ड	d, dh	d
व	v, w, o, b, bh	v
ब	b, bh,v,	b
श	sh,s	sh
क्ष	ksh, x	ksh
औ	au,ou	au
ई	i,e,ee,ey	i

Using Table 4 multiple transliteration candidates can be generated. Few examples are shown in Table 5.

Table 5. Multiple Transliterations

Devanagari	English Transliterations
चौधरी	Chaudhari, Chaudhary, Choudhari, Choudhary
वामनराव	Vamanrao, Vamanrav, Wamanrao, Wamanrav
लक्ष्मणराव	Laxmanrav, Laxmanrao, Lakshmanrav, Lakshmanrao
फडके	Phadake, Phadke, Fadake, Fadke

As shown in above table 4, a single Hindi consonant or vowel gets mapped to one or more character sequences in English. As a result of this, multiple transliterations can be generated. Due to this fact, it is very difficult to get Top-1 transliteration as per the user expectation most of the times. To avoid this problem a formulation of fixed set of rules is expected.

3.4 Spelling by Orthography or Phonology

It is observed that, in India there is a lot of confusion about the spelling of three aksharas named entity, having first two light syllables or having middle one as the light syllable [22]. Our observations after an analysis of student enrollment lists in colleges, BSNL directory and voter lists, shows that there are mixed approaches to write the spellings. Most of the people prefer to retain the schwa of second syllable. Figure 1 shows spelling by orthography using the syllable (denoted by σ) structure where 'O' denotes onset and 'N' denotes nucleus [23].

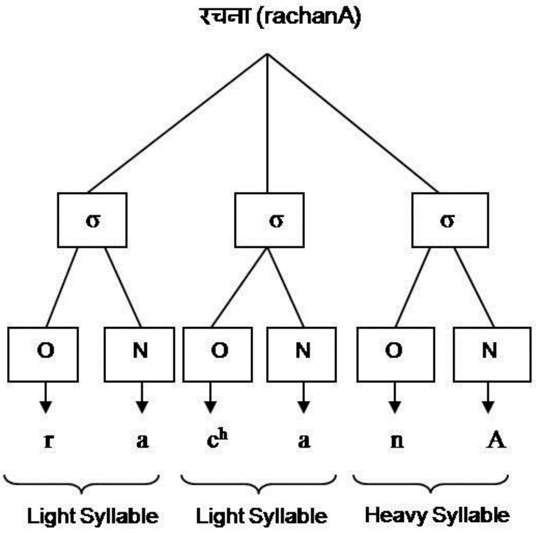

Figure 1. Spelling by Orthography

Figure 2 shows spelling by phonology.

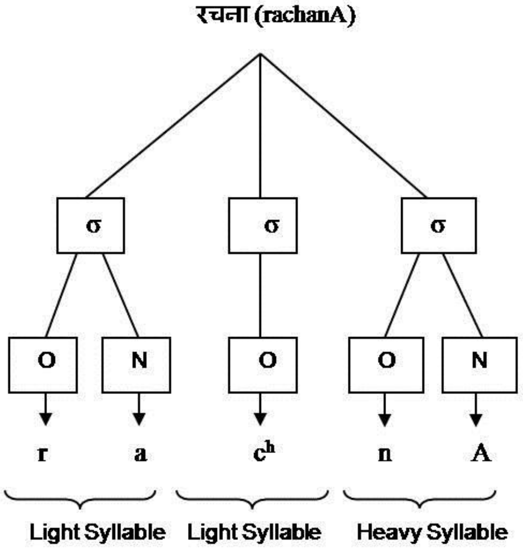

Figure 2. Spelling by Phonology

We found that some people write the spelling as per orthography and others by phonology. Table 6 shows a few examples of common surnames used in the state of Maharashtra, India.

Table 6. Spellings by Orthography and Phonology

Devanagari NE	Spelling By Orthography	Spelling by Phonology
नवले	Navale	Navle
वनवे	Vanave	Vanve
दगडे	Dagade	Dagde
मडके	Madake	Madke
चोपडे	Chopade	Chopde
कोलते	Kolate	Kolte
फाळके	Falake	Falke
तावरे	Tavare	Tavre

The reason behind the ambiguity of spelling is the deletion of schwa of second syllable, leads to consonant cluster in Hindi and Marathi. Another reason is the way the word is pronounced by an individual in various places. The origin of this problem is the lack of knowledge of prosodic features of spoken language, such as intonation, rhythm, and stress. It is true that the spellings are written using phonology but as shown in table 6, many people write the spelling as per orthography. If the spellings are written using phonology then the back transliteration may result in incorrect pronunciation. Few examples are

- Named Entity /*Garje/’* may be pronounced as /गर्जे/ as well as /गरजे/
- Named Entity /*Bansi/* may be pronounced as /बन्सी/ as well as /बनसी/
- Named Entity /*Falke/* may be pronounced as /फल्के/, /फलके/, /फळके/, /फाल्के/, /फाळके/ etc.

This problem occurs only with named entities consisting of three aksharas. In Hindi the named entities with two, four, five, six, seven and eight aksharas are very rhythmic in nature.

3.5 Allophones /v/ and /w/

The Devanagari alphabet /व/ gets mapped with two phones in English /v/ and /w/. They are the allophones for Devanagari alphabet /व/. These two variations of /v/ and /w/ in Hindi has long been noticed but today also there are no clear guidelines available from any authority about whether to map /व/ to /v/ or /w/. Few authors have carried out research in this regard, which is very useful but which is being used only by the researchers to carry out their research only [22]. We have found Top-1 accuracy get affected only because of improper mapping of /व/ to /v/ and /w/. Few examples are

- The village name /वडाळा/ is transliterated as /Wadala/ while another village name /वडगाव/ is transliterated as /Vadgaon/.

If we search on Google a word /Vaman/ which is a person name, we get first results as

- /Waman Hari Pethe/ which is /वामन हरी पेठे/

and third result as

- /Vaman Rao Pai/ which is /वामन राव पै/

It shows that the name /वामन/ is written as /Waman/ as well as /Vaman/. This kind of mapping of /व/ affects the accuracy of machine transliteration. In the state of Maharashtra of India, the Devanagari alphabet /व/ is mapped to /o/ if it appears as a last letter in named entity. For example, the correct transliteration of named entity /माणिकराव/ is /Manikrav/ but it is always written as /Manikrao/. Another interesting example is the name /राव भाई राव/ is transliterated as /Rav Bhai Rao/.

3.6 Spelling Variations

One other important issue is the way a particular named entity is transliterated differently in the various parts of India. Few examples are

- The name of month /श्रावण/ (also used as proper noun) is transliterated as /Shravan/ in western part of India while it is transliterated as /Sravan/ in northern part of India.
- The name of Indian idol /श्रीराम/ is transliterated as /Shriram/, /Sriram/ or /Shreeram/.

3.7 Capitalization

There is no concept of capitalization of leading characters of names in Indian languages unlike English and other European languages which plays an important role in identifying named entities for transliteration. From back transliteration point of view capitalization is very useful in the intermediate code of phonetic model. This model treats transliteration as a phonetic process rather than an orthographic process. Under such frameworks, transliteration is treated as a conversion from source grapheme to source phoneme followed by a conversion from source phoneme to target grapheme. For example

- The back transliteration of named entity /Manikrav/ would be /माणिकराव/, /मणिकराव/, /माणिकरव/ or /मणिकरव/. In rule based transliteration models always there is problem of mapping the vowel /a/ in English to either /अ/ or /आ/ in Devanagari. If the intermediate code for the name /माणिकराव/ is generated as /mAnikrAv/ then it would produce the correct back transliteration in Devanagari.

Same thing also happens with statistical methods, if the training data is not sufficient to learn all the patterns in the source language. Earlier the electricity bills were printed only in English in the Maharashtra state of India. Now a day’s bills are printed in bilingual format. It seems that the original database in English is transliterated in Marathi by using statistical method. The first author of this paper receives the electricity bill in bilingual format where the name /manikrao/ is transliterated as /मणिकराव/, but the correct name of author is /माणिकराव/.

3.8 Language of Origin

Some named entities are transliterated in multiple ways and each method is correct according to the context under consideration. These words are of foreign origin and are extensively used in local language. It is transliterated considering local context at certain instances while global context is considered at certain instances. This ambiguity affects accuracy considerably. For example

- The name /रज़ाक/ has the Arabic origin and should be transliterated as /Razaq/ but it has been observed that it is also written as /Razak/ in India. In Indian linguistic context the last letter /q/ is replaced by /k/. In India people from different origins and religions are living from hundreds of years. Therefore, providing the origin dependent transliteration is the major issue.

3.9 Short Vowel and Long Vowel

Hindi and Marathi languages have a set of orthography rules for writing Devanagari script. The rules include constraints which specify the context in which they are applicable, for examples, start of a word, ending of a word, after a vowel, after a consonant (AC), etc [88]. One of the rules is

- A word final akshara with the diacritic mark should be written using long vowel matra sign. Few examples are shown in Table 7.

Table 7. Examples of use of Long Vowels

Correct	Incorrect	Correct	Incorrect
मी (I)	मि	टोकरी (basket)	टोकरि
तू (You)	तु	अन्दरुनी (internal)	अन्दरुनि
पती (husband)	पति	कर्मचारी (worker)	कर्मचारि
ड़ाली (branch)	ड़ालि	हिंदुस्तानी (Indian)	हिंदुस्तानि
लकड़ी (wood)	लकड़ि	समझदारी (understanding)	समझदारि

For transliterating named entities in the Hindi and Marathi languages into English one of the major issues is whether to map the long vowel matra [◌ी / ī] to /i/ or /ee/. As per Devanagari linguistic rule long vowel /◌ी / should be replaced by /ee/ in English but for all examples in table 7, the last long vowel is always replaced by short vowel /i/.

Table 8 shows the correct and incorrect transliterations of mapping long vowel /ee/ according to the linguistics of Hindi and Marathi. The fact is, for the mapping of /ee/, there are a set of rules published by the Government, but nobody follows it.

Table 8. Examples of Use of Short / Long Vowels

Devanagari	English Transliteration
सीता	Seeta – Correct Spelling Sita – Incorrect Spelling
सिता	Sita– Correct Spelling Seeta – Incorrect Spelling
शितल	Shital – Correct Spelling Sheetal – Incorrect Spelling
शीतल	Sheetal - Correct Spelling Shital – Incorrect Spelling

It has been observed that many people are not following the rules of short vowels and long vowels during transliteration which affects the Top-1 accuracy of the machine transliteration.

3.10 Conjuncts

English script does not encode the conjuncts but they are pronounced in speech. But Devanagari script encodes conjuncts exactly the way one would pronounce it. In some names, the half consonants present in the name are wrongly transliterated as full consonants in the output word, and vice-versa. There is no problem in the forward transliteration but it may wrongly encode in back transliteration. For example

- The named entity /Asmita/ would get transliterated as /असमिता/ but the correct transliteration is /अस्मिता/.

3.11 Affixes

Name entities of the proper noun and locations have some common prefixes and suffixes in India. Few of the suffixes derived from the phonology of the English need to be retained while transliterating location names. But if these suffixes occur in proper nouns, they need to be transliterated as per the linguistic of Hindi or Marathi. For example

- The location name /खामगाव/ is transliterated as /Khamgaon/ where /gaon/ is the common suffix in India. But if the /गाव/ suffix occur at the start of the proper name, then it has to be transliterated as /gav/. The surname /गावस्कर/ should be transliterated as /Gavaskar/ and not as /Gaonskar/. Similarly, the location name /बंगलोर/ is transliterated as /Banglore/ but if /लोर/ appears as a suffix in the proper name like /सालोर/ then it should be transliterated as /Salor/ and not as /Salore/.

3.12 Acronyms

Acronyms are words formed from initial letters or a group of initial letters in a series of words. These words are generally transliterated according to the phonology of the Hindi or Marathi. But, such words need to be transliterated on character by character basis. For example

- An acronym /व्हीआयपी/ would be transliterated as 'vhiaipi'. But, the correct transliteration is 'VIP'. Also, the /पीईटीए/ which is the acronym of the 'People for the Ethical Treatment of Animals' would be transliterated as /piitie/ whereas the correct transliteration is /PETA/.

3.13 Loan Words

Loan words are the words taken from other languages. These words are always transliterated in English according to the phonology of their mother language. For example

- The named entity /बंगला/ is a loan word in Hindi taken from English. If this word is transliterated according to the phonology of Hindi, then its transliteration would be /bangla/. But the correct transliteration according to English phonology is /bungalow/.

3.14 Incorrect Syllabification

Syllabification is process of converting Named Entity into its phonetic units. Many a times incorrect syllabification leads to inaccuracy. For example

- If the Named Entity /गायत्री/ is syllabified as /गाय/ and /त्री/ then it gets transliterated as /Gaytri/ which is not correct transliteration. But, if we syllabify it as /गा/, /यत/ and /री/ then we get correct transliteration as /Gayatri/.

3.15 Consonant ‘r’ in the Conjuncts

The consonant /र/ (r) is special in the conjuncts where /r/ occurs as the first consonant and is written using a special ligature. In these conjuncts, the presence of r is indicated by a shape resembling a hook above the last consonant of the conjunct called as *rafar* matra. Few examples are

र + क = र्त   र + म = र्म   र + ण = र्ण   र + त = र्त

कर्म = क + र + म is transliterated as ‘Karma’ where consonant /m/ is followed by /r/.

If /र/ follows a consonant in a cluster and is itself followed by a vowel, it is represented in different ways depending upon the shape of the preceding consonant. If the preceding consonant has a vertical line in it, then र is represented by the mark ’,’ placed at the lower part of the vertical line. Few examples are

क् + र = क्र   प् + र = प्र   म् + र = म्र   ग् + र = ग्र

चक्र = च+क+र is transliterated as ‘chakra’ where consonant /r/ is followed by consonant /k/. The same thing works for the cluster of /ट/ and /र/ which is written as /ट्र/. Few examples are: /महाराष्ट्र/ (State Name) and /राष्ट्रम/ (Kingdom, Nation). There is no difficulty in the forward transliteration but it creates the problem in back transliteration about the position of consonant ‘r’.

For example

- The named entity /मारके/ (surname in State of Maharashtra, India) is commonly transliterated as /Marke/ in English. During the back transliteration it gets transliterated as /मार्के/ which is not correct.

3.16 Schwa Identification and Deletion

Most of the written languages do not have a one-to-one correspondence between characters in their orthographies and the pronunciations of those characters in written words. In other words, an individual character in the orthography is not always equivalent to exactly one sound in a language. The schwa is the vowel sound in many lightly pronounced unaccented syllables in words of more than one syllable. In most languages, one or more vowels of the language which can occur in any syllable and which bears no stress or less stress, are called schwa. Schwa is the name for the most common sound in English. It is a weak, unstressed vowel sound and it occurs in many words. Any vowel letter can be pronounced as schwa in English and the pronunciation of a vowel letter can change depending on whether the syllable in which it occurs is stressed or not. The schwa is not any independent letter; instead it is related with the duration of vowel sound during the pronunciation. Schwa is represented by the symbol /ə/ as per the International Phonetic Association (IPA). In Devanagari script which is used as a standard to write Hindi and Marathi, there is an implicit अ in each consonant phoneme if it occurs without any other vowel matra in written form.

When a named entity written in Devanagari script is transliterated using English script, the implicit ‘अ’ attached to the single consonant either get mapped to ‘a’ or mapped to schwa /ə/ depending on whether the syllable is stressed or unstressed in the given Devanagari word. Like English the schwa is not related with all vowels in Hindi and Marathi, instead it is related only with the first vowel ‘अ’, which is inherently embedded in each consonant phoneme.

In Hindi and Marathi the problem is that the schwa is sometimes pronounced and sometimes not. For example, in the name /*Narhari/* (/नरहरी/ in Devanagari, /Nərəhəri/ in IPA) the schwa following letter ‘r’ is deleted in the pronunciation. Without any schwa deletion, not only with the two words sound very unnatural, but it will also be extremely difficult for the listener to distinguish between the two. Without the appropriate deletion of schwas, any speech output would sound unnatural. Consonants in written Hindi often carry annotations indicating the nature of the following vowel, which is not written separately. When there is no explicit marking, schwa is the default vowel, but this vowel does not always emerge in a word’s pronunciation. In Hindi (and Marathi) to English direct transliteration without corpus, identification and deletion of schwa plays a very critical role. According to Hindi and Marathi phonology, the stress pattern for the named entity /रामदेव/ is shown in table 9.

Table 9. Stress pattern for NE 'रामदेव’

Named Entity	**Syllabification**	**Stress Falls on**	**Less Stress Falls on**
रामदेव	[राम] [देव]	[रा][दे]	[म][व]

Therefore the correct transliteration would be ‘Ramdev’ and schwa after consonant ‘m’ gets deleted.

4. EXPERIMENTATION

We have developed a rule-based phonetic model using Linguistic approach. Total 15,244 Named Entities are tested and the Top-1 accuracy of our system is 74.139% and mismatch rate at Top-1 is 25.861%.

The test data includes personal names, surnames, and city and village names. The following notations are used for the evaluation metrics [23].

N: Total number of names in the test set

R_i: i-th reference name (input) in source language in the test set

$C_{i,\ k}$: k-th candidate transliteration (output) for i-th name in the test set (1 < k < 7)

Ki: Number of candidate transliterations produced by a transliteration system

The commonly used performance evaluation parameter ‘word accuracy’ denoted by ACC is used to test the performance of our method. It measures correctness of the first transliteration candidate in the candidate list produced by a transliteration system [100]. ACC = 1 means that all top candidates are correct transliterations i.e. they match one of the references, and ACC = 0 means that none of the top candidates are correct [46].

$$ACC = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1 \; if \; \exists Ri : Ri = C_{i,1} \\ 0 \;\; Otherwise \end{cases} \quad (1)$$

Table 10 and figure 3 shows the results of the 15224 names transliterated by using phonetic map table after pruning and deleting word final schwa.

Table 10. Top-1 Accuracy Results

Length in Devanagari (Number of Aksharas)	Number of Named Entities	Number of Correct Transliterations (Top-1) & %	Number of Incorrect Transliterations & %
2	1839	1831 (99.564%)	8 (0.436%)
3	6061	6040 (99.635%)	21 (0.365%)
4	4780	2731 (57.133%)	2049 (42.867%)
5	1970	581 (29.492%)	1389 (70.508%)
6	497	94 (18.913%)	403 (81.087%)
7	61	10 (16.393%)	51 (83.607%)
8	16	0 (0%)	16 (100%)
	15224	11287 (74.139%)	3937 (25.861%)

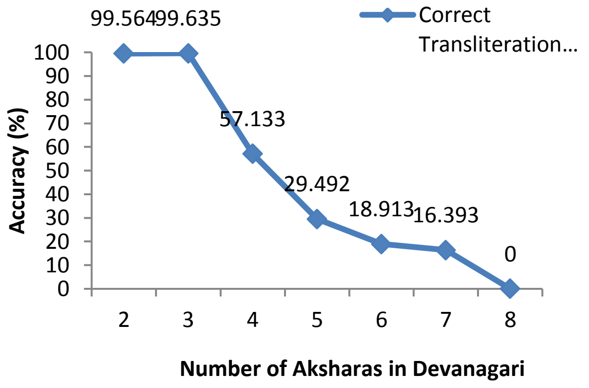

Figure 3. Top-1 Accuracy

The error analysis (25.861% = 100%) is shown in Table 11.

Table 11. Error Analysis

Nature of Error	Percentage
Missing Sounds	0.4 %
Spelling Variations and Multiple Transliterations	35.8%
Spelling by orthography or Phonology	19.6%
Allophones	23.4%
Language of Origin	6.2%
Short Vowel and Long Vowel	5.9%
Conjuncts	2.3%
Affixes	4.8%
Incorrect Syllabification	1.6%

Following are the results of Top-1 accuracy, Mean F-Score, Top-2 MRR, Precision and Recall
Word Accuracy in Top-1 (ACC) =74.139 %
Fuzziness in Top-1 (Mean F-score) =0.919
Mean Reciprocal Rank (MRR- Top-2) = 0.739
Precision = 0.947
Recall =0.897
Figure 4 depicts the results of Top-1 accuracy, Mean F-Score, Top-2 MRR, Precision and Recall.

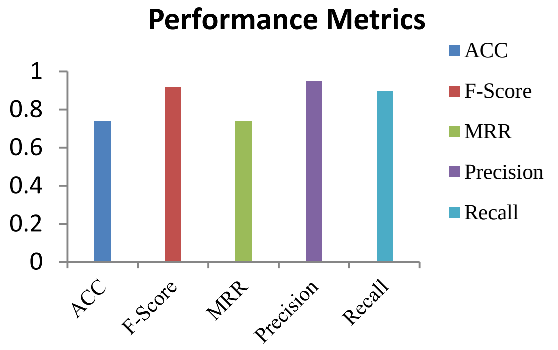

Figure 4. Performance Metrics

5. CONCLUSION

In this paper, we presented frequently occurred issues in machine transliteration of Hindi to English and Marathi to English. We have focused on the Top-1 accuracy gets affected by spelling variations as well as by the possibility of multiple transliterations. We found most of the error occurs due allophones [v] and [w] as there are no clear guidelines available about their usages in the transliteration. These errors can be reduced only by using training based statistical models.

Another major issue noticed by us is the transliteration of three aksharas (syllabic units) named entities in Marathi and Hindi where people use mixed approach to write the spelling either by orthography or phonology. The only solution to this issue is to present the spellings by both the approaches.

The errors caused by the language of origin can be reduced by using the special tags in the database about the origin of language. For the use of short and long vowels, set of rules is available, but nowadays most of the people are not following it. It is possible to reduce these kinds of errors if the spellings are formulated as per linguistic rules. If all above issues are taken into the consideration, it would definitely help in increase of accuracy of machine transliterations of Hindi to English and Marathi to English.

6. ACKNOWLEDGMENTS

We express our gratitude to Dr. Shashikant Halkude, Principal, WCE, Solapur, Dr. Sulbha Apte, Professor and Head, Department of Computer Engineering, WCE, Solapur, H K Abhyankar, Vice President VIT, Dr. Mukund Nalawade, Dr. V M Kale, Prof. Ashish Mujumdar Dr. Manasi Patwardhan and Dr. C M Mahajan of Vishwakarma Institute of Technology for their continuous encouragement to carry out the research in the area of Machine Transliteration. We also thank to Sandip Shinde, Milind Kulkarni, Santosh Shelke, Umesh Kodgule, Pravin Rathod, Alankar Dasre, Vijay Budkar, Aparna Mitragotri, Uday Jaunjal, Abhinandan Desai, Anil Ingole, Sanjay Aradhye, Ramrao Dhore, Satyam Manohar Kale, Vandana Gajanan Dhore, Varsha Vilas Bhise, Rekha Baburao Akhare, Chhaya Ashish Morkhade, Ganesh

Gadkar, Ananta Bankar, Anil Kamble, Ravindra Badhe, Ajay Shigwan for their moral support and valuable guidance of Marathi and Hindi linguistics. Special thanks to Tushar Sonwalkar who helped in the error analysis of machine transliteration.

7. REFERENCES

[1] Manoj Jain, Vijay Kumar and Manish Kumar Singh, Encoding of Indic script : ISCII & Unicode, Department of Information Technology, pp. 95-106, New Delhi

[2] Frost & Sullivan, Local language information technology market in India, TDIL, Department of IT, Ministry of Communications and Information Technology, India, pp. 1-128, 2003

[3] Knight Kevin and Graehl Jonathan, Machine transliteration. In proceedings of the 35th annual meetings of the Association for Computational Linguistics, pp. 128-135, 1998

[4] Al-Onaizan Y, Knight K, Machine translation of names in Arabic text. Proceedings of the ACL conference workshop on computational approaches to Semitic languages. 2002

[5] Nasreen Abdul Jaleel and Leah S. Larkey, Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the 12th international conference on information and knowledge management. pp: 139 – 146, 2003

[6] Goto I, Kato N, Uratani N, and Ehara T, Transliteration considering context information based on the maximum entropy method. In Proceedings of MT-Summit IX, pp. 125-132, 2003

[7] Li H, Zhang M, and Su J, A joint source-channel model for machine transliteration. In Proceedings of ACL , pp. 160-167, 2004Niraj Aswani, Robert Gaizauskas, Aligning words in English-Hindi parallel corpora, Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 115–118, Ann Arbor, June 2005. Association for Computational Linguistics, 2005

[8] Katre D.S. A Position Paper on Cross Cultural Usability Issues of Bilingual (Hindi & English) Mobile Phones in the Proceedings of Indo Danish HCI Research Symposium. 2006

[9] Janet Pierrehumbert and Rami Nair, Implications of Hindi Prosodic Structure, Northwestern Univesity

[10] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma. Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition. In Proceeding of workshop on Workshop on NER for South and South East Asian Languages, IJCNLP-2008

[11] Abbas Malik, Laurent Besacier, Christian Boitet and Pushpak Bhattacharyya, A Hybrid Model for Urdu Hindi Transliteration, Machine transliteration shared task, named entities workshop: shared task on transliteration, Singapore, pp. 177-185, 2009

[12] Ankit Aggarwal, Transliteration involving English and Hindi languages using syllabification approach, Thesis, Indian Institute of Technology, Bombay, Mumbai, 2009

[13] Manoj K. Chinnakotla, Om P. Damani, and Avijit Satoskar, Transliteration for Resource-Scarce Languages, ACM Trans. Asian Lang. Inform. Process. 9, 4, Article 14, pp 1-30, December 2010

[14] Karimi S, Scholer F, and Turpin, Machine transliteration survey. ACM Computing Surveys, Vol. 43, No. 3, Article 17, pp.1-46, April 2011

[15] Umair Z Ahmed Kalika Bali Monojit Choudhury and Sowmya, Challenges in Designing Input Method Editors for Indian Languages: The Role of Word-Origin and Context Proceedings of the Workshop on Advances in Text Input Methods, pp.1–9, Thailand, November 13, 2011.

[16] Design Guidelines (Sanskrit, Hindi, Marathi, Konkani, Sindhi, Nepali), Department of Information Technology, pp. 95-106, New Delhi, pp. 1-38

[17] Omkar N. Koul, Modern Hindi Grammar, Dunwoody Press, 2008

[18] S. P.Mudur, N. Nayak, S. Shanbhag, and R. K. Joshi, An architecture for the shaping of indic texts, Computers and Graphics, vol. 23, pp. 7–24, 1999

[19] Walambe M R, Marathi Shuddalekhan, Nitin Prakashan, Pune, 1990

[20] Walambe M R, Marathi Vyakran, Nitin Prakashan, Pune, 1990

[21] Janet Pierrehumbert and Rami Nair, Implications of Hindi Prosodic Structure, Northwestern Univesity.

[22] Naim R Tyson and Ila Nagar, Prosodic rules for schwa-deletion in Hindi Text-to-Speech synthesis, International Journal of Speech Technology, pp. 15–25, 2009

[23] Haizhou Li, A Kumaran Vladimir Pervouchine and Min Zhang, Report of NEWS 2009 Machine transliteration shared task, named entities workshop: shared task on transliteration, Singapore, pp. 1-18, 2009

AUTHOR'S PROFILE

M. L. Dhore (manikrao.dhore@vit.edu) has completed ME in Computer Science and Engineering from NITR, Chandigarh, India in 1998.Currently he is working as Associate Professor in Computer Engineering Department at Vishwakarma Institute of Technology, Pune, and Maharashtra, India. Presently he is pursuing his Ph.D. from University of Solapur, Maharashtra, India, in Computational linguistics. His areas of research interest are Machine Transliteration and Computer Networking.

Dr. S. K. Dixit (dixitsk1@yahoo.com) has received Ph.D. in Electronics from Shivaji University, Kolhapur in 2002. Currently he is working as Head and Professor in Department of Electronics and Telecommunication at Walchand College of Engineering, Solapur, Maharashtra, India.

Ruchi Dhore (ruchidhore@yahoo.com) is the student of second year Computer Engineering at Pune Vidyarthi Grih's College of Engineering and Technology, Pune, Maharashtra, India. Her area of research interest includes Text Processing and Pattern Searching. She likes to build her carrier in linear and non-linear pattern searching.