TRANSLITERATION BY ORTHOGRAPHY OR PHONOLOGY FOR HINDI AND MARATHI TO ENGLISH: CASE STUDY

M L Dhore[1] R M Dhore[2] P H Rathod[3]

[1,3]Department of Computer Engineering, Vishwakarma Institute of Technology, Pune

[2]Pune Vidhyarthi Girh's College of Engineering and Technology, Pune

ABSTRACT

e-Governance and Web based online commercial multilingual applications has given utmost importance to the task of translation and transliteration. The Named Entities and Technical Terms occur in the source language of translation are called out of vocabulary words as they are not available in the multilingual corpus or dictionary used to support translation process. These Named Entities and Technical Terms need to be transliterated from source language to target language without losing their phonetic properties. The fundamental problem in India is that there is no set of rules available to write the spellings in English for Indian languages 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 and in turn the translation process. Major issue noticed by us is the transliteration of named entities consisting three syllables or three phonetic units in Hindi and Marathi languages where people use mixed approach to write the spelling either by orthographical approach or by phonological approach. In this paper authors have provided their opinion through experimentation about appropriateness of either approach.

KEYWORDS

Machine Transliteration, Phonology, Orthography, Devanagari, Hindi, Marathi, Syllable, Phonetic

1. INTRODUCTION

Hindi is the official national language of India and spoken by around ~500 million population of India. Marathi is one of the widely spoken Indo-Aryan languages in India especially in the state of Maharashtra and border areas of nearby states. Marathi and Hindi languages are written using Devanagari script. Marathi is spoken by more than 71.9 million (6.99% population of India) Indian populations. Transliteration is the conversion of a word from one script to another script without losing its phonological characteristics. For the direct transliteration of Hindi and Marathi to English *named entities*, major issues are difference in writing scripts, missing sounds, multiple transliterations, spelling by orthography or phonology, allophones, conjuncts, affixes, acronyms, loan words, incorrect syllabification, consonant 'r' in the conjuncts and schwa identification and deletion [1]. Henceforth named entity is denoted as *NE* and named entities denoted as *NEs.*

This paper focuses on the issues of making spellings for *NEs* of Hindi and Marathi languages in English and that to only the *NEs* of length three *aksharas* (phonetic units/syllable). Akshara is the single phonetic unit of orthography in Devanagari script. It is the minimal articulatory unit of speech in Hindi and Marathi. One *akshara* with or without diacritic in Hindi and Marathi is referred as a one phonetic unit or one syllable in the following sections of this paper. For

DOI : 10.5121/ijnlc.2013.2501

example: *NE* /माणिकराव (Mānikrāv)/ is made up of following five *aksharas*(syllables/phonetic units)

/मा(Mā) + णि(ni) + क(k) + रा(rā) + व (v)/

There are very few *NEs* made up of using only one phonetic unit or syllable. After the exhaustive analysis of *NEs* as a part of doctoral research, it is found that nearly 33% *NEs* used in India are made up of three syllables. Most of the people write their name in English strictly according to orthography of Hindi and Marathi while other makes the use of phonology.

2. RELATED WORK

Two broad approaches for machine transliterations are Grapheme-based and Phoneme-based. Grapheme-based transliteration follows an orthographic method and maps the source language graphemes/characters directly to the target language graphemes/characters. This approach obtains the transliteration using orthographic method. Phoneme-based transliteration follows the phonetic process where transliteration is treated as a conversion from source grapheme/character to source phoneme followed by a conversion from source phoneme to target grapheme/character. Second approach obtains the transliteration using phonology method.

C-DAC, NCST and Indictrans Team are the major player in the machine transliteration of Indian languages. C-DAC provided their technology based on ISCII in 1980 in the form of hardware based card called GIST. NCST developed a phonemic code based scheme for effective processing of Indian languages in 2003 [2]. Table 1 shows the various models and approaches used for the period 1994-2013 for selected refereed journal papers.

Table 1. Various Models and Approached used for Transliteration

Author	Year	Language Pair	Model	Approach
Arbabi[3]	1994	Arabic-English	Phoneme/Phonology	Handcrafted Rules
Knight[4]	1998	Japanese-English	Phoneme/Phonology	Weighted Finite State Transducers
Wan[5]	1998	English-Chinese	Phoneme/Phonology	Syllabification
Lee[6]	1998	English-Korean	Grapheme/Orthography	Source Channel
Jeong[7]	1999	Korean-English	Phoneme/Phonology	HMM
Kang[8]	2000	English-Korean	Grapheme/Orthography	HMM
Kang[9]	2000	English-Korean	Grapheme/Orthography	Decision Trees
Jung[10]	2000	English-Korean	Phoneme/Phonology	Extended Markov
Oh[11]	2002	English-Korean	Phoneme/Phonology	Contextual Rules
Lin[12]	2002	Chinese-English	Phoneme/Phonology	Widrow-Hoff
Al-Onaizan [13]	2002	Arabic-English	Hybrid	Source Channel WFST
Paola[14]	2003	English-Chinese	Phoneme/Phonology	Festival Speech Synthesis
Goto[15]	2003	English-Japanese	Grapheme/Orthography	Maximum Entropy
Jaleel[16]	2003	English-Arabic	Grapheme/Orthography	Handcrafted Rules & Bi-gram
Gao[17]	2004	English-Chinese	Phoneme/Phonology	Source Channel

Li[18]	2004	English-Chinese.	Grapheme/Orthography	Joint Channel
Author	**Year**	**Language Pair**	**Model**	**Approach**
Bilac[19]	2005	Japanese-English Chinese-English	Hybrid	Source Channel, EM and WFST
Malik[20]	2006	Shahmukhi -Gurmukhi	Grapheme/Orthography	Handcrafted Rules
Ekbal[21]	2006	Bengali-English	Grapheme/Orthography	Modified Joint Source Channel
Oh[22]	2006	English-Korean English-Japanese	Hybrid	MEM, MBL and Decision Tree
Oh[23]	2007	English-Korean English-Japanese	Combined	SVM MEM
Ganesh[24]	2008	English-Hindi	Grapheme/Orthography	HMM CRF
Surana[25]	2008	English-Hindi English-Telugu	Phoneme/Phonology	DATM
Saha[1]	2008	Hindi-English Bengali-English	Phoneme/Phonology	Handcrafted Rules
Karimi[26]	2008	English-Persian Persian-English	Combined	Source Channel Voted Method
Martin[27]	2009	English-Korean English-Hindi English-Kannada English-Russian	Grapheme/Orthography	N-Grams, FST and WFST
Oh[28]	2009	English-Chinese English-Hindi English-Japanese English-Russian English-Korean	Grapheme/Orthography	CRF, Margin Infused Relaxed Algorithm (MIRA), EM
Sittichai [29]	2009	English-Chinese English-Hindi English-Japanese English-Russian English-Korean	Grapheme/Orthography	N-Gram HMM Linear Chain
Vijayanand [30]	2009	English-Tamil	Grapheme/Orthography	Handcrafted Rules
Chai[31]	2010	English-Thai	Grapheme/Orthography	Syllabification. Letter To Sound
Chinnakotla [32]	2010	Hindi-English English-Hindi	Grapheme/Orthography	Handcrafted Rules, CSM
Yu-Chun Wang[33]	2011	English-Korean	Hybrid	CRF
Ying Qin[34]	2011	English-Chinese Chinese-English	Hybrid	CRF
Najmeh Mousavi Nejad[35]	2011	Farsi-to-English	Hybrid	MEM
Kishorjit[36]	2012	Bengali -Meitei Mayek	Phoneme/Phonology	SVM
Dhore[37]	2012	Hindi-English	Phoneme/Phonology	CRF

Rathod[38]	2013	Marathi/Hindi-English	Phoneme/Phonology	SVM

3. HINDI AND MARATHI

India is a multilingual country with 22 constitutional officially recognized languages and 11 different scripts used in different regions spread across the country. Hindi is the world's fourth most commonly used language after Chinese, English and Spanish. It is an Indo-Aryan language which is a branch of Indo-European languages spoken as a first or second language by almost ~500 million people in India, as well as other parts of Asia, Africa, America, Europe and Oceania. Marathi is the fourth most commonly spoken language after Hindi, Bengali and Telugu in India. There are 34 full consonants, 5 traditional conjuncts and 1 traditional sign in Devanagari script used for Hindi and Marathi languages and each consonant have 13 variations through integration of 13 vowels. The 34 pure consonants and 5 traditional conjuncts along with 13 vowels produce 507 different alphabetical characters [39]. The consonant /ळ/ is used only in Marathi Language and not in Hindi language. Table 2 shows the map used for Marathi to English transliteration.

Table 2. Transliteration Map

Consonant Phonemes and Traditional Conjuncts	क ka	ख kha	ग ga	घ ha	ङ ~ga	च cha	छ Cha	ज ja	झ jha	ञ ~ja
	ट Ta	ठ Tha	ड Da	ढ Dha	ण Na	त ta	थ tha	द da	ध dha	न na
	प pa	फ pha	ब ba	भ bha	म ma	य ya	र ra	ल la	व va	श sha
	ष Sha	स sa	ह ha	ळ La	क्ष kSha	ज्ञ jnya	द्य dya	श्र shra	त्र tra	ॐ om
Vowel Phonemes	अ a	आ AA	इ ii	ई II	उ uu	ऊ UU	ऋ rr	ए ee	ऐ ai	ओ oo
	औ OO	अं AM	अः AH							
Graphical Signs (Matras) for Vowel	ा ā ं aM	ि i ः aH	ी I	ु u	ू U	ृ Ru	े e	ै ai	ो o	ौ au

The basic consonant shape in the Indian script always has the implicit short vowel /अ(a)/ and hence there is no explicit matra form for the short vowel ‘a’. For example, a NE /रजत(Rajat)/ is linguistically written as below in Devanagari generic script.

/र्+अ+ज्+अ+त्+अ (r+a+j+a+t+a)/

However, there is equivalent matra available for all other 12 vowels as / ा – ā, ि-i, ी- ī, ु-u, ू-ū, े -e, ै- ai, ो- o, ौ-au, ं-aM, ः – aH /, which get attached to the basic consonant shape whenever the corresponding vowel immediately follows the consonant. The written form of a

basic consonant without the implicit 'a' vowel either has an explicit shape or it has the graphical sign '्', known as *Halant* in Hindi and *Virama* in Marathi get attached to its basic consonant shape (e.g. क्). This is referred as pure consonant form of writing Hindi or Marathi language. The *Halant/Virama* is the vowel अ (a) omission sign. It serves to cancel the inherent vowel अ of the consonant to which it is applied [39].

When Devanagari vowel phoneme / अ (a)/ is added to any generic consonant phoneme then the consonant phoneme is called full consonant. It is necessary to have inherent / अ (a)/ attached to consonant to add the phone of /अ (a)/ vowel to it. If inherent / अ (a)/ is not added to the independent consonant phoneme, it becomes very difficult to utter its transliteration in English as well as to obtain its back transliteration in Hindi and Marathi from English [2]. Unicode and ISCII character encoding standards for Indian scripts are based on full form of consonants.

3. WHY THREE SYLLABLE *NES*?

The 22670 *NEs* consisting male names, female names, place names and organization names are analyzed. It has been observed that the minimum length of the *NE* is one *akshara* (formed using 1 syllabic unit or phonetic unit) and maximum length is eight *aksharas*. There are very few *NEs* written using either only one phonetic unit or more than eight phonetic units. From the number of *aksharas* in the *NE*, 8 categories are made. One akshara is considered equivalent to one phonetic unit in the Devanagari word. It is observed that nearly 50% *NEs* used in India are the combination of two or three individual *NEs*.

For example:

NE /कुंभारगावतावडी/ (Kumbhārgāontāwadi- a place name) is formed using three NEs as /कुंभार (Kumbhār)/, /गाव (Gāon)/ and /तावडी (Tāwadi)/ respectively.

As the length of a *NE* increases, segmentation is required to find out the number of words used to form a *NE* in order to separate the rhythms within it and in turn number of phonetic units in each rhythm/segment. Table 3 depicts the analysis based on the number of segments and their lengths in *aksharas* for Hindi and Marathi *NEs*.

Table 3. Analysis

Length of NE	Named Entity	Segmentation	Segment Lengths
4	सुनयना(Sunaynā)	सु + नयना (Su + naynā)	1+3
	श्रीवर्धन(Shriwardhan)	श्री + वर्धन (Shrī + wardhan)	1+ 3
	रामचंद्र(Rāmchandrā)	राम + चंद्र (Rām + chandrā)	2+2
	रामराव(Rāmrāo)	राम + राव(Rām + rāo)	2+2
	धवलश्री(Dhawalshrī)	धवल + श्री (Dhawal + shrī)	3+1
5	भानुप्रताप(Bhānuprātāp)	भानु + प्रताप (Bhānu + pratāp)	2+3
	ओमप्रकाश(Omprakāsh)	ओम + प्रकाश (Om + prakāsh)	2+3

	किसनलाल(Kisanlal)	किसन + लाल (Kisan + lāl)	3+2
	मोहनदास(Mohandās)	मोहन + दास (Mohan + dās)	3+2
	रतनलाल(Ratanlāl)	रतन + लाल(Ratan + lāl)	3+2
	लक्ष्मणराव(Laxmanrāv)	लक्ष्मण + राव(Laxman + rāv)	3+2
6	शेजवळकर(Shejwalkar)	शेज+वळ+कर (Shej + wal + kar)	2+2+2
	शिवनारायण(ShivnārāyaN)	शिव+ नारायण (Shiv + nārāyaN)	2+4
	कमलकिशोर(Kamalkishor)	कमल+ किशोर (Kamal + kishor)	3+3
	जवाहरलाल(Jawāharlāl)	जवाहर+लाल (Jawāhar + lāl)	4+2
7	अहमदनगर(Ahmadnagar)	अह + मद + नगर(Ah +mad+nagar)	2+2+3
	गुरसहायगंज(Gursahāyganj)	गुर+सहाय+गंज(Gur+sahāy+ganj)	2+3+2
	मुरलीमनोहर(Muralīmanohar)	मुरली + मनोहर (Muralī+ manohar)	3+4
	नारायनस्वरूप(Nārāyanswarūp)	नारायन+ स्वरूप(Nārāyan+ swarūp)	4+3
	पुरुषोत्तमदास(Purushottamdās)	पुरुषोत्तम+दास(Purushottam + dās)	5+2
8	पुरुषोत्तमनगर(Purushottamnagar)	पुरुषोत्तम+नगर (Purushottam+nagar)	5+3
	कुंभारगावतावडी (Kumbhārgāvtāvadī)	कुंभार+गाव+तावडी (Kumbhār+gāv+tāvadī)	3+2+3
	लाखनवाडाबुद्रुक (Lākhanwādābudruk)	लाखन+वाडाबुद्रुक (Lākhan+wādābudruk)	3+5
	ब्रम्हपुरीपेडगाव (Brahmapurīpedgāv)	ब्रम्हपुरीपेड+गाव (Brahmapurīped +gāv)	6+2

Analysis in Table 3 depicts that most of *NEs* of length 4,5,6,7 and 8 are made of the two or three segments consisting of length 2 and 3. *NEs* having length 2 are always written in English using the orthography of Hindi and Marathi and full consonant approach. Few examples are shown below.

/श्याम/ is transliterated as /Shyām/ and not as /Shyāma/

/ प्राण/ is transliterated as /Prān/ and not as /Prāna

It is to note that, the short vowel 'a' if occurs at the end of transliteration is the *Halant/Virama* and empirically it is always deleted to pronounce the name in the proper manner.

NEs having length 3 are written in English either using the orthography or phonology of Hindi and Marathi. Few examples are shown below.

/ममता/ is transliterated as /Mamtā/ using phonology as well as /Mamatā/ using orthography.

/सरला/ is transliterated as /Sarlā/ using phonology as well as /Saralā/ using orthography.

There are many such *NEs* of length 3 which are written by the people using either approach. The transliteration system can be developed only using either approach. Therefore, the same error gets

gradually propagated for the *NEs* having length 4, 5,6,7 and 8, if *NE* contains the segment/segments of length 3. Following are the few examples according to their length in number of *aksharas*.

Examples of Devanagari *NEs* of length 4 *aksharas*:

- /सागरीका/ cannot be segmented but the root word is /सागर/. It would be transliterated as /Sāgrikā/ using phonology as well as /Sāgarikā/ using orthography.
- /भुवनेश/ cannot be segmented but the root word is /भुवन/. It would be transliterated as /Bhuvnesh/ using phonology as well as /Bhuvanesh/ using orthography.

Examples of Devanagari *NEs* of length 5 *aksharas*:

- /ममताबाई/ is the combination of 3+2 as /ममता/ +/ बाई/. It would be transliterated as /Mamtābāi/ using phonology as well as /Mamatābāi/ using orthography.
- /मानसीताई/ is the combination of 3+2 as / मानसी/ + /ताई /. It would be transliterated as /Mānsitāi/ using phonology as well as /Mānasitāi/ using orthography.

Examples of Devanagari *NEs* of length 6 *aksharas*:

- /सरलाकुमारी/ is the combination of 3+3 as /सरला/ + /कुमारी/. It would be transliterated as /Sarlākumāri/ using phonology as well as /Saralākumāri/ using orthography.
- /बायसामाऊली/ is the combination of 3+3 as /बायसा/ + /माऊली/. It would be transliterated as /Bāysāmāuli/ using phonology as well as /Bāyasāmāuli/ using orthography.

Examples of Devanagari *NEs* of length 7 *aksharas*:

- /मुरलीमनोहर/ is the combination of 3+4 as / मुरली/ + /मनोहर/. It would be transliterated as /Murlimanohar/ using phonology as well as /Muralimanohar/ using orthography.
- /जानकीनारायण/ is the combination of 3+4 as / जानकी/ + /नारायण /. It would be transliterated as /Jānkinārāyan/ using phonology as well as /Jānakinārāyan/ using orthography.

Examples of Devanagari *NEs* of length 8 *aksharas*:

- /कुंभारगावतावडी/ is the combination of 3+2+3 as /कुंभार/ + /गाव/ + /तावडी/ . It would be transliterated as /Kumbhārgāontāwdi/ using phonology as well as /Kumbhārgāontāwadi/ using orthography.
- /कवडगावजालना/ is the combination of 3+2+3 as / कवड/ + /गाव/ + /जालना/. It would be transliterated as /Kavadgāonjālnā/ using phonology as well as /Kavadgāonjālanā/ using orthography.
- /खिरडीगणेशपूर/ is the combination of 3+3+2 as / खिरडी/ + / गणेश/ + /पूर/. It would be transliterated as /Khirdiganeshpur/ using phonology as well as /Khiradiganeshpur/ using orthography.

From such examples, it is clear that no transliteration system can provide 100% Top-1 accuracy. For our experimentation, we have prepared a database of 4500 *NEs* of length 3 in Hindi and Marathi by using voter lists, Census lists and Telephone Directories and tested to check which approach is more appropriate.

4. SYSTEM ARCHITECTURE

The architecture of Hindi and Marathi to English transliteration system is shown in figure 1.

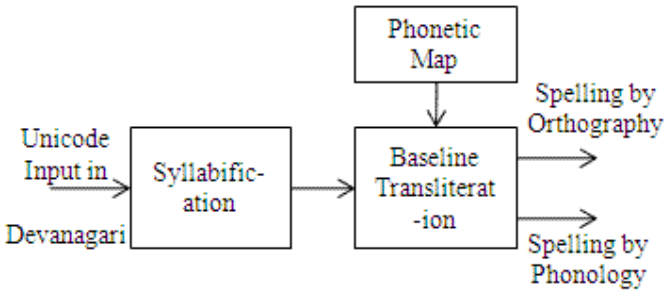

Figure 1: Architecture

4.1. Syllabification

As Unicode uses full consonant approach it treats Devanagari consonant phoneme and vowel phoneme as a separate units as shown below.

देवकी(Devaki) → द + े + व + क+ ी (C + V + C + C + V) where C is Consonant & V is Vowel

Table 4 shows few examples of Unicode based internal representation of Devanagari NEs.

Table 4. Internal Representation of Devanagari Script

Name Entity in Devanagari	Internal Representation using Unicode
राम(Rām)	र + ा + म
कृष्ण(KrishNa)	क + ृ + ष + ् + ण
श्रीराम(Shrīrām)	श + ् + र + ी + र + ा + म
विठ्ठल(Vitthal)	व + ि + ठ + ् + ठ + ल
धृतराष्ट्र(Dhrutrāshtra)	ध + ृ + त + र + ा + ष + ् + ट + ् + र
ज्ञानेश्वर(Dnyāneshwar)	ज + ् + ञ + ा + न + े + श + ् + व + र

BarahaIME (Input Method Editor) is used to accept the input in Devanagari. It supports only the Unicode. BarahaIME is used to type Indian language Unicode text directly into applications such as Internet Explorer, MS Word, Notepad, etc. When BarahaIME

program is started, it shows as an icon in the system tray at the bottom-right portion of the screen. It supports Kannada, Hindi, Marathi, Sanskrit, Tamil, Telugu, Malayalam, Gujarati, Gurumukhi, Bengali, Assamese, Manipuri and Oriya languages. IME Manager maps key codes of keyboard keys to characters in different language. When user presses any key, IME Manager reads the code of the key and language of the user. Based on this information it returns the mapped key to display on screen.

This feature of Unicode is very useful in the creation of Devanagari Phonetic Units. From internal representation of Unicode, phonetic units are formed for Devanagari names as shown below.

देवकी → द + े + व + क+ ी → । दे । व । की ।

Syllabification refers to the segmentation of source and target language NEs at phonetic level called as language Translation Units (TU). TU is equivalent to the syllabic unit or phonetic unit of the source or target language. Following algorithm is used to obtain the syllabic units of Devanagari *NE*. In the algorithm, V stands for vowel, C stands for consonant; HC stands for half consonant and G stands for nasal sound in Devanagari

Algorithm: Formation of Devanagari Phonetic Transliteration Units

Input: Devanagari *NE* in Unicode with their orthographic consonants and vowel phonemes

if first phoneme is V check for second phoneme
  if the second phoneme is V or G
    Combine it with first phoneme and mark it as first syllabic unit
  else
    Mark first V as a first syllabic unit (V)
  end if
end if
foreach next phoneme ***do***
  if the phoneme C is followed by HCVG or HCV or HCG or HC or VG or V
    Combine it and mark as a next syllabic unit and continue
  else
    Mark C as a next syllabic unit and continue
  end if
  if the phoneme C is followed by HCHCVG or HCHCV or HCHCG or HCHC or VG or V
    Combine it and mark as a next syllabic unit and continue
  else
    Mark C as a next syllabic unit and continue
  end if
end foreach

Output: Devanagari Phonetic Units or Source Transliteration Units (STUs).

Working of Algorithm is illustrated in the Table 5.

Input in Devanagari is गोविंदस्वरुप (Govindswarup is a person name)

Input in Unicode representation = ग +ो + व+ ि + ं + द + स + ् + व + र + ु + प

Table 5. Formation of Devanagari Phonetic Units

Phoneme Sequence	Devanagari Phonemes	Algorithm	Phonetic Unit
1	ग	Combines ग + ो (CV)	गो
2	ो		
3	व	Combines व + ि+ ं (CVG)	
4	ि		विं
5	ं		
6	द	द (C)	द
7	स	Combines स + ् + व (CHC)	स्व
8	्		
9	व		
10	र	Combines र + ु (CV)	रु
11	ु		
12	प	प (C)	प

4.2. Transliteration Module

It is to note that the first vowel /अ/ in Hindi and Marathi is mapped to English letter 'a' (short vowel) while the second vowel /आ/ is mapped to 'ā' (long vowel as per IPA) in English. The alphabet 'a' in English is a short vowel equivalent to /अ/ which is also a short vowel in Hindi and Marathi while /आ/ in Hindi and Marathi is a long vowel and mapped to capital 'ā' in our phonetic scheme. Unicode and ISCII character encoding standards for Indic scripts are based on full form of consonants.

4.3 Transliteration by Orthography

This module maps each syllable in Hindi and Marathi into English by using full consonant based phonetic map. Phonetic map is implemented by using the translation memory and mapping is done by writing the manual rules. This mapping does not consider the phonological effects of individual syllable. It does not consider the metrical structure of the given word and simply maps the syllables using full consonant approach. Table 6 shows the spellings generated in English for the Devanagari NEs consisting of three *aksharas* using orthography.

Table 6. Spellings by Orthography

Named Entity in Devanagari	Syllabification	Mapping by Orthography	Transliteration /Spelling in English
आपटे	[आ] [प] [टे]	[ā] [pa] [te]	āpate
सरला	[स] [र] [ला]	[sa] [ra] [lā]	saralā

ममता	[म] [म] [ता]	[ma] [ma] [tā]	mamatā
दळवी	[द] [ळ] [वी]	[da] [la] [vi]	dalavi
फडके	[फ] [ड] [के]	[pha] [da] [ke]	phadake
रचना	[र] [च] [ना]	[ra] [cha] [nā]	rachanā
फाळके	[फा] [ळ] [के]	[fā] [la] [ke]	fālake
मानसी	[मा] [न] [सी]	[mā] [na] [si]	mānasi

4.4 Transliteration by Phonology

According to phonology of Hindi and Marathi there are three categories of vowels as short, long and diphthongs as shown below.

Short Vowels अ/a or ə/, उ/u/, इ /i/

Long Vowels आ/ ā/,ए /e/,ई/ī/, /ओ /o/,ऊ /ū/

Diphthongs ऐ/ai/, औ/au/

Generally, the location of word stress in Hindi and Marathi is predictable on the basis of syllable stress. Stress is related both to the vowel length and the occurrence of postvocalic consonant. According to Hindi and Marathi phonology literature there are three classes of vowels used for stress analysis but it is possible to obtain the stress analysis using only two classes as shown below.

Cμ - Light syllable (CV) where V is the only short vowel /a/ or schwa /ə/. L is used to denote light syllable henceforth.

Example: क (ka/kə) →CV

Cμμ - Heavy syllable (CVV, CCV, CVVN, CCVVN), where VV in pair is the long vowel, a single V is the short vowel and N is nasal sign. H is used to denote heavy syllable henceforth.
Examples:

का (kā) → CVV

प्र(pra) →CCV

प्रां(prān) →CVVN

The CVC or CVVC are closed syllables and they also form Heavy syllable after combination.
Examples:

कर(kar) →CVC

कार(kār) →CVVC

This approach considers the stress of individual syllable whether it is light syllable or heavy syllable according to the phonology of Hindi and Marathi and the way it is pronounced. The schwa is the vowel sound in many lightly pronounced unaccented syllables in words of more than one syllable. It is represented by /ə/ symbol [40]. When a *NE* written in Devanagari script is

transliterated using Roman script, the implicit अ/a/ attached to the single consonant either get mapped to short vowel 'a' or 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/a/, which is inherently embedded in each consonant phoneme. The schwa of unstressed syllable remained in the transliterated output need to be removed for the appropriate pronunciation of word according to phonology of Hindi and Marathi. The schwa identification and deletion is done by applying stress analysis.

It is observed that, in India there is a lot of confusion about the spelling of three *aksharas NE*, having first two light syllables or having middle one as the light syllable [41]. Most of the people prefer to retain the schwa of second syllable while other removes the schwa of middle less stress syllable to retain the properties of phonology. Table 7 shows *NEs* having their middle syllable as unstressed or light syllable.

Table 7. *NEs* with Unstressed Middle Syllable

आपटे	फडके	मडके	साधना	नवले	फणसे	तायडे	कामदे	भेलके
सरला	रचना	वनवे	देवकी	जबडे	सालवे	तावरे	जावळे	फेगडे
ममता	फाळके	शेळके	केतकी	कुमठे	झावरे	कामना	आसरे	देवळे
दळवी	मानसी	पवळे	देवळे	कोलते	रोकडे	वारके	बारले	चोपडे
वंदना	सांगली	साधना	सायली	माऊली	गायत्री	बाफना	तारळे	पावसे
गोमती	रेवती	मालती	मुरली	बायसा	तनया	गेनबा	वाळके	रावते

According to phonology, Table 8 shows the syllabification and stress patterns of Devanagari NEs which are made up of three *aksharas*/syllables.

Table 8. Stress Analysis by Phonology

Named Entity	**Syllabification**	**Stress Falls on**	**Less Stress Falls on**
आपटे	[आ] [प] [टे]	[आ] [टे]	[प]
सरला	[स] [र] [ला]	[स] [ला]	[र]
ममता	[म] [म] [ता]	[म] [ता]	[म]
दळवी	[द] [ळ] [वी]	[द] [वी]	[ळ]
फडके	[फ] [ड] [के]	[फ] [के]	[ड]
रचना	[र] [च] [ना]	[र] [ना]	[च]
फाळके	[फा] [ळ] [के]	[फा] [के]	[ळ]

मानसी	[मा] [न] [सी]	[मा] [सी]	[न]

According to phonological properties, the inherent short vowel /a/ of middle light syllable is treated as schwa /ə/ and hence schwa gets removed in the final transliteration if the last syllable is heavy syllable. Table 9 shows the transliterations of *NE* consisting of three syllables using phonology.

Table 9. Spelling by Phonology

Named Entity in Devanagari	**Syllabification**	**Mapping by Phonology**	**Transliteration /Spelling in English**
आपटे	[आ] [प] [टे]	[ā] [pə] [te]	āpte
सरला	[स] [र] [ला]	[sa] [rə] [lā]	sarlā
ममता	[म] [म] [ता]	[ma] [mə] [tā]	mamtā
दळवी	[द] [ळ] [वी]	[da] [lə] [vi]	dalvi
फडके	[फ] [ड] [के]	[pha] [də] [ke]	phadke
रचना	[र] [च] [ना]	[ra] [chə] [nā]	rachnā
फाळके	[फा] [ळ] [के]	[fā] [lə] [ke]	fālke
मानसी	[मा] [न] [सी]	[mā] [nə] [si]	mānsi

5. EXPERIMENTATION

Figure 2 depicts the snapshot of experimentation showing transliteration generated for the Devanagari *NE* /ममता/ using orthography as well as phonology where /ममता/ is transliterated as /Mamata/ using orthography and /Mamta/ using phonology.

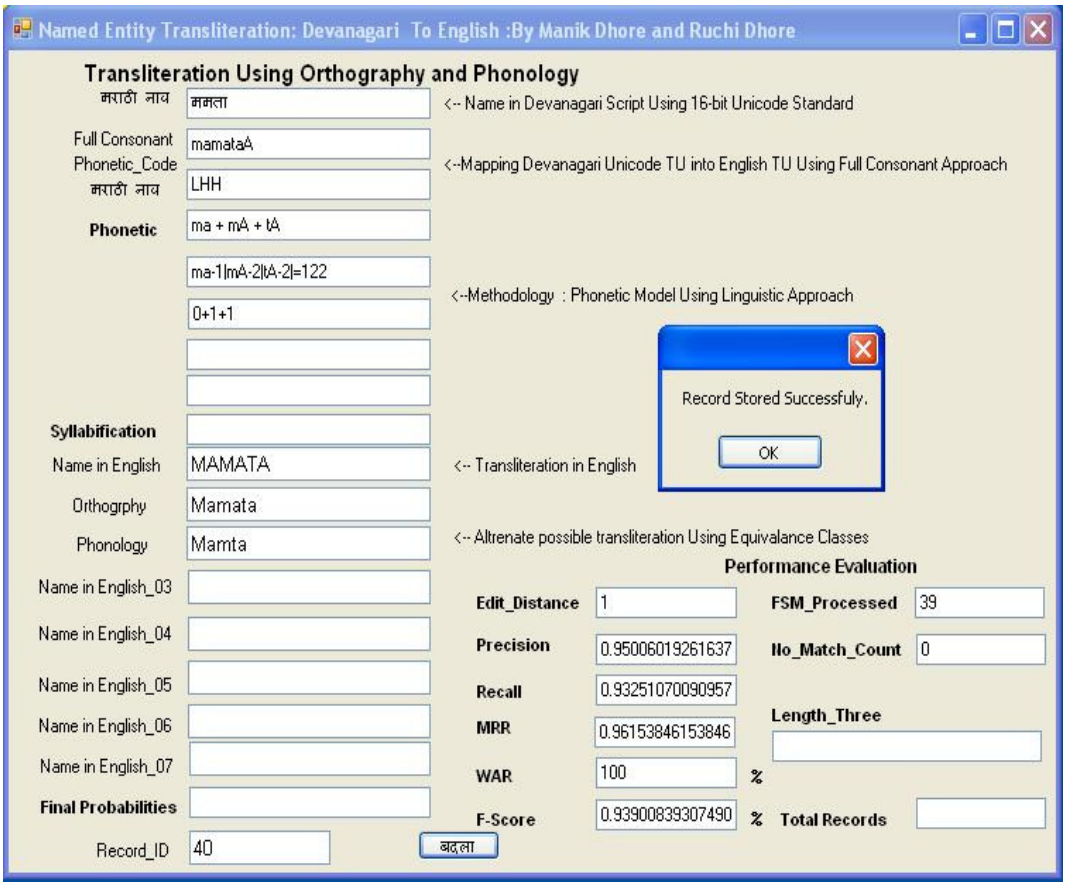

Figure 2. Transliteration Using Orthography and Phonology

In order to find out the ratio between writing the spellings using orthography and phonology we created the database of 4500 records consisting of only three syllable *NEs*. To create the database we used the voter's lists of State of Maharashtra which is available in English, Hindi and Marathi, Census lists and Telephone Directories which are available both in English and Hindi, Census portal of Government of India and few web resources related with the *NEs* [42-51]. Our transliteration engine generates the spellings in English for both the approaches using orthography and phonology. Comparison is made for each NE in the database and frequency for each name entity is calculated in terms of how many people write the same name using orthography and similarly how many people write the same name using phonology.

Table 10 shows the few *NEs* along with their frequencies of writing spellings in English using orthography and phonology.

Table 10. Frequencies of sample *NEs*

NE	#records	#Spellings by Orthography	#Spellings by Phonology
रचना	12	8 (rachanā)	4 (rachnā)
सरला	5	3 (saralā)	2 (sarlā)
ममता	9	7 (mamatā)	2 (mamtā)
मडके	11	8 (madake)	3 (madke)

वनवे	8	6 (vanave)	2 (vanve)
शेळके	15	12 (shelake)	3 (shelke)
मानसी	13	10 (mānasi)	3 (mānsi)
Total	73	54	19

Figure 3 shows the snapshot of database after transliteration.

LTE3

ID	Local_Name	Phonetic_Cc	Local_From_	English_Nan	English_Nan
107	वंदना	vaMdanaA	HLH	Vandana	Vandna
108	मुरली	mauralaI	HHH	Murali	Murli
109	सायली	saAyalaI	HHH	Sayali	Sayli
110	पायली	paAyalaI	HHH	Payali	Payli
111	प्रेरणा	pa`raeraNaA	HLH	Prerana	no
112	रचना	rachanaA	LLH	Rachana	Rachna
113	साधना	saAdhanaA	HLH	Sadhana	Sadhna
114	जानकी	jaAnakaI	HHH	Janaki	Janki
115	अनिल	anaila	HHL	Anil	Aneel
116	शंकर	shaMkara	HHL	Shankar	Sankar
117	दत्तात्रय	da`ta`taAta`ray	HHL	Dttatray	no
118	अमेय	amaeya	HHL	Amey	no
119	शर्वरी	shara`varaI	LHH	Sharwari	Sharvari
120	करन	karana	LHL	Karan	no
121	अथर्व	athara`va	HLH	Atharva	Atharwa
122	चिन्मय	chaina`maya	HHL	Chinmay	Cheenmay
123	प्राजक्ता	pa`raAjaka`taA	HLH	Prajakta	Prajacta
124	सलोनी	salaonaI	LHH	Saloni	Salonee
125	तरुण	tarauNa	LHL	Tarun	Taroon
126	कौशल	kaaushala	HLL	Kaushal	Kausal
127	साहिल	saAhaila	HHL	Sahil	Saheel
128	शार्दुल	shaAra`daula	HHL	Shardul	Shardool
129	आशिष	AAshaiSha	HHL	Ashish	Aashish
130	वरद	varada	LHL	Varad	no
131	मानसी	maAnasaI	HHH	Manasi	Mansi

Figure 3. Database after Transliteration

Figure 4 shows the results of experimentation.

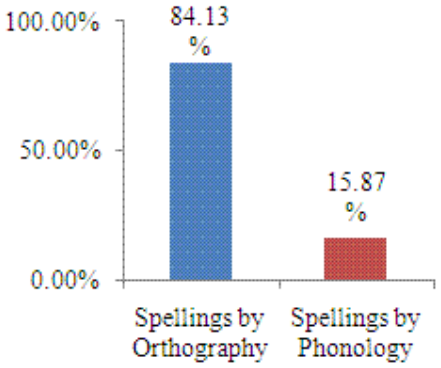

Figure 4. Results

Our case study shows that, 84.13% people write their names using the orthography approach and remaining writes it using phonology. From this outcome, it is clear that the transliteration engine must generate every three *aksharas NE* using orthography as well as phonology approach.

6. CONCLUSIONS

We have presented our results with the help of experimentation. A result shows that most of the people write the spelling in English by using the orthography rather than phonology. The reason behind the ambiguity of not writing the spellings using phonology especially for three syllable NEs is the deletion of schwa of second syllable leads to consonant cluster in Hindi and Marathi during back transliteration. If the spellings are written using phonology then the back transliteration may result in consonant cluster in Hindi and Marathi. For example, *NE* 'Bansi' in English may be back transliterated as 'बन्सी' rather than 'बनसी'. The issue of writing the spelling by using mixed approach certainly affects the top-1 accuracy of Hindi and Marathi to English transliteration.

REFERENCES

[1] Saha Sujan Kumar, Ghosh P. S, Sarkar Sudeshna and Mitra Pabitra (2008) "Named entity recognition in Hindi using maximum entropy and transliteration."

[2] Joshi R K, Shroff Keyur and Mudur S P (2003) "A Phonemic code based scheme for effective processing of Indian languages", *National Centre for Software Technology, Mumbai, 23rd Internationalization and Unicode Conference, Prague, Czech Republic*, pp 1-17.

[3] Arbabi M, Fischthal S M, Cheng V C and Bart E (1994) "Algorithms for Arabic name transliteration", *IBM Journal of Research and Development*, pp 183-194.

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

[5] Wan Stephen and Verspoor Cornelia Maria (1998), "Automatic English-Chinese name transliteration for development of multilingual resources", *Microsoft Research Institute, Macquarie University, Sydney, Australia,* pp. 1352-1356.

[6] Lee J S and Choi K S (1998), "English to Korean statistical transliteration for information retrieval", *Computer Processing of Oriental Languages*

[7] Jeong K S, Myaeng S H, Lee J S and Choi K S (1999), "Automatic identification and back-transliteration of foreign words for information retrieval", *Information Processing and Management*, pp. 523-540.

[8] Kang I H and Kim G (2000), English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks, In Proceedings of the 18th Conference on Computational Linguistics, pp. 418–424.

[9] Kang B J, and Choi K S (2000), "Automatic transliteration and back-transliteration by decision tree learning", *In Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp.1135-1411.

[10] Jung S Y, Hong S S and Paek E (2000), "English to Korean transliteration model of extended Markov window", *In Proceedings of the 18th Conference on Computational Linguistics*, pp.383–389.

[11] Oh J H, and Choi K S (2002), "An English-Korean transliteration model using pronunciation and contextual rules", *In Proceedings of COLING 2002*, pp.758-764.

[12] Lin W H and Chen H H (2002), "Backward machine transliteration by learning phonetic similarity", *In Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–7.

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

[14] Paola Virga and Khudanpur Sanjeev (2003), "Transliteration of proper names in cross-lingual information retrieval", *Johns Hopkins University, USA, Proceedings of the ACL Workshop on Multilingual and Mixed-language Named Entity Recognition*, pp. 57-64.

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

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

[17] Gao W, Wong K F and Lam W (2004), "Phoneme-based transliteration of foreign names for OOV problem", *In Proceedings of the First International Joint Conference on Natural Language Processing, Lecture Notes in Computer Science, vol. 3248, Springer, Berlin*, pp. 110–119.

[18] Li Haizhou, Zhang Min and Su Jian (2004), "A joint source-channel model for machine transliteration", *In Proceedings of ACL*, pp. 160-167.

[19] Bilac S and Tanaka H (2004), "Improving back-transliteration by combining information sources", *In Proceedings of IJCNLP2004*, pp. 542-547.

[20] Malik M G A (2006) “Punjabi Machine Transliteration”, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp 1137–1144.

[21] Ekbal A, Naskar S and Bandyopadhyay S (2006), "A modified joint source channel model for transliteration", *In Proceedings of the COLING-ACL, Australia*. pp. 191-198.

[22] Oh J H and Choi K S (2006), "An ensemble of transliteration models for information retrieval", *Information Processing and Management, 42, 4,* pp. 980–1002.

[23] Oh J H and Ishara H (2007), "Machine transliteration using multiple transliteration engines and hypothesis re-ranking", *In Proceedings of the 11th Machine Translation Summit,* pp. 353–360.

[24] Ganesh S, Harsha S, Pingali P, and Verma V (2008) “Statistical transliteration for cross language information retrieval using HMM alignment and CRF”, *In Proceedings of the Workshop on CLIA, Addressing the Needs of Multilingual Societies.*

[25] Surana Harshit and Singh Anil Kumar (2008), "A more discerning and adaptable multilingual transliteration mechanism for Indian languages", *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP), Asian Federation of Natural Language Processing, Hyderabad, India,* pp. 64-71.

[26] Karimi Sarvnaz (2008),"Machine Transliteration of proper names between English and Persian", *Thesis, RMIT University, Melbourne, Victoria, Australia.*

[27] Martin Jansche and Sproat Richard (2009), "Named Entity Transcription with Pair n-Gram Models", *Machine Transliteration Shared Task, ACL-IJCNLP*, pp. 32-35

[28] Oh Jong-Hoon, Kiyotaka Uchimoto, and Kentaro Torisawa (2009) "Machine transliteration using target-language grapheme and phoneme: Multi-engine transliteration approach", *Proceedings of the Named Entities Workshop ACL-IJCNLP Suntec, Singapore, AFNLP*, pp 36–39

[29] Sittichai Jiampojamarn, Bhargava Aditya, Dou, Qing Dwyer Kenneth and Kondrak Grzegorz (2009), "DirecTL: a Language independent approach to transliteration", *Proceedings of the 2009 Named Entities Workshop, Singapore,* pp. 28–31.

[30] Vijayanand Kommaluri (2009), "Testing and performance evaluation of machine transliteration system for Tamil language", *Proceedings of the 2009 Named Entities Workshop, Singapore,* pp. 48–51.

[31] Chai Wutiwiwatchai and Thangthai Ausdang (2010), "Syllable-based Thai-English machine transliteration", *National Electronics and Computer Technology Center Pathumthani, Thailand, NEW-Sweden* pp. 66-70.

[32] Chinnakotla Manoj K., Damani Om P., and Satoskar Avijit (2010) "Transliteration for Resource-Scarce Languages", *ACM Trans. Asian Lang. Inform,*Article 14, pp 1-30.

[33] Yu-Chun Wang and Richard Tzong-Han Tsai (2011), "English-Korean Named Entity Transliteration Using Statistical Substring-based and Rule-based Approaches", *IJCNLP, Proceedings of NEWS 2011, Named Entities Workshop, Thailand,* pp 32-35

[34] Ying Qin and GuoHua Chen (2011), "Forward-backward Machine Transliteration between English and Chinese Based on Combined CRFs", *IJCNLP, Proceedings of NEWS 2011, Named Entities Workshop, Thailand,* pp 82-85

[35] Najmeh Mousavi Nejad, Shahram Khadivi and Kaveh Taghipour (2011), "The Amirkabir Machine Transliteration System for NEWS 2011: Farsi-to-English Task",", *IJCNLP, Proceedings of NEWS 2011, Named Entities Workshop, Thailand,* pp 91-95

[36] Kishorjit Nongmeikapam (2012), "Transliterated SVM Based Manipuri POS Tagging", *Advances in Computer Science and Engineering and Applications*, pp 989-999

[37] Dhore M L, Dixit S K and Sonwalkar T D (2012), "Hindi to English Machine Transliteration of Named Entities using Conditional Random Fields", *International Journal of Computer Applications, Vol 48, No. 23,* pp 31-37

[38] Rathod P H, Dhore M L and Dhore R M (2013), "Hindi and Marathi To English Machine Transliteration Using SVM", *International Journal on Natural Language Computing, (IJNLC) Vol. 2, No.4,* pp 55-71

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

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

[41] Pandey Pramod Kumar (1990), "Hindi schwa deletion", *Lingua 82,* pp. 277-31

[42] ceo.maharashtra.gov.in/ Voter Lists of Government of Maharashtra in English and Marathi

[43] http://www.censusindia.gov.in/

[44] http://www.indianchild.com/

[45] http://en.wiktionary.org/

[46] http://www.whereincity.com/babynames

[47] http://en.wikipedia.org/wiki/List_of_cities_and_towns_in_India

[48] http://www.gaminggeeks.org/Resources

[49] http://encyclopedia.thefreedictionary.com/

[50] http://www.dte.org/

[51] http://www.vit.net/ Intranet Portal of Vishwakarma Institute of Technology, Pune

Authors

M. L. Dhore (manikrao.dhore@vit.edu) has completed ME in Computer Science and Engineering from NITTR, Chandigarh, India in 1998. Currently he is working as Associate Professor in Department of Computer Engineering at Vishwakarma Institute of Technology, Pune. He is on the verge of getting his Ph.D. from University of Solapur, Maharashtra, India, in the area of Computational Linguistics. He has his interest in Machine Translation and Machine Transliteration specifically in Marathi-English and Hindi- English Language Pairs. He has developed the tools for Devanagari to English Machine Transliteration for the online web based commercial applications. His current areas of research are Internet Routing Algorithms, Computer Networking, Machine Translation and Transliteration

Ruchi M Dhore (ruchidhore93@gmail.com) is the student of Third Year Computer Engineering at Pune Vidyarthi Grih's College of Engineering and Technology, Pune, Maharashtra, India. She is scholar student of her college and securing distinction every year in the University of Pune examinations. She is very good in programming and won the prizes in state level and national level competitions. Her area of research interest includes Text Processing and Pattern Searching. She likes to build her carrier in the development of language processing tools for Marathi language

P. H. Rathod (pravin.rathod@vit.edu) has completed BE in Information Technology, from Gov ernment College of Engineering, Karad, Maharashtra, India, in 2008. Recently he has completed ME in Computer Science and Engineering from Vishwakarma Institute of Technology, Pune, India in 2013. Currently he is working as Assistant Professor in Department of Computer Engineering at Vishwakarma Institute of Technology, Pune. He has his interest in Machine Translation and Machine Transliteration specifically in Devanagari-English Language Pairs. His current areas of research are Mobile Ad hoc Networks, Internet Routing Algorithms, Computer Networking, Machine Translation and Transliteration.