International Journal of Emerging Technology & Research
Volume 1, Issue 1, Nov-Dec, 2013 (www.ijetr.org) ISSN (E): 2347-5900 ISSN (P): 2347-6079

Transliteration from Hindi Script to Meetei Mayek: (A Rule Based Approach)

Santoshi Watham[1], V.R.Vimal[2]

[1]M.E, Computer Science and Engineering, Vel Tech Multi Tech Engineering College, Avadi, Chennai

[2]Head, Department of Computer Science and Engineering, Vel Tech Multi Tech Engineering College, Avadi, Chennai

Abstract

This project describes about the transliteration of Manipuri an Eight Schedule Language of Indian Constitution from Hindi script to Meitei Mayek (Meitei script. The transliteration of Manipuri is necessary because Meitei Mayek is an Eight Schedule language of Indian Constitution and transliteration of Hindi will be one of the ways to let understand the language because Hindi is the national language of India and also is one of several languages spoken in different parts of the sub-continent. Manipuri transliteration is hardly found so far but some of transliteration works has been done in such areas as "Manipuri Transliteration from Bengali script to Meitei Mayek" and "Transliteration of Meitei Mayek(Manipuri script) to English (Roman) script" etc. But upto the best of the authors' knowledge no work of transliteration from Hindi to Meitei Mayek Script has done and this is the first work of transliteration for Manipuri. By using simple rule base approach a model and algorithm is being designed for transliterating Manipuri from Hindi script to Meitei script. Even though the model is a simple rule base approached but to our surprise the algorithm proved to come up with an accuracy of 86.28%.

Keywords: *Transliteration, Hindi Script, Meitei Mayek, Iyek.*

1. Introduction

Transliteration is the process of mapping a word of a source language script to another target language script. In another word it is the conversion from one script to target script by preserving the pronunciation of the source script. A simple approach of transliteration is the letter-by-letter approach. Manipuri (or Meiteilon) is a Tibeto-Burman (TB) language which is highly agglutinative in nature, mono-syllabic, influenced and enriched by the Indo-Aryan languages of

Sanskrit origin and English. This language is one of the Eight Scheduled languages of Indian Constitution and spoken in some parts of Bangladesh and Myanmar. Manipuri uses two scripts; the first one is purely of its own origin, *Meitei Mayek* while another one is a borrowed *Bengali Script*.

Hindi is the fourth most widely-spoken language in the world (after Mandarin,spanish and english): an estimated 500-600 million people speak the language. Hindi is spoken in 10 states - Uttarpradesh, Uttaranchal, Haryana, Delhi, Himachal Pradesh, Rajasthan, Madhyapradesh, Chhattisgarh, Jharkand and Bihar. The transliteration of Manipuri is necessary because Meitei Mayek is an Eight Schedule language of Indian Constitution and transliteration of Hindi will be one of the way to let understand the language because Hindi is the national language of India and also is one of several languages spoken in different parts of the sub-continent.Manipuri transliteration is hardly found so far but some of transliteration works has been done in such areas as "Manipuri Transliteration from Bengali script to Meitei Mayek" and "Transliteration of Meitei Mayek(Manipuri script) to English (Roman) script" etc.

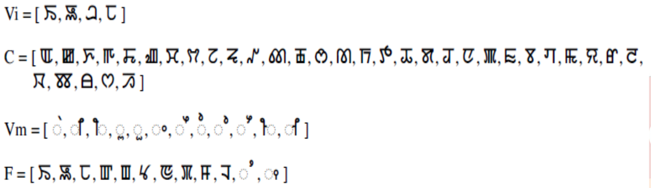
The paper is organized in the following manner: Section 2 gives a brief idea about the related works in transliteration; Section 3 gives the details about the Linguistic Transliterating Scheme of Hindi Script to Meitei Mayek, the model and algorithm is mention in Section 4, the experiment result and evaluation is given in Section 5 and the conclusion is drawn in Section 6.

1.1 Meitei Mayek Scripts

The Meitei Mayek script was originally of the Brahmic type [1]. Consonants bear the inherent vowel, and vowel matras modify it. Unlike most other Brahmic scripts, Meitei Mayek makes use of explicit final consonants which have no inherent vowel. Consonant conjuncts are not formed productively in the

Volume 1, Issue 1, Nov-Dec, 2013 (www.ijetr.org) ISSN (E): 2347-5900 ISSN (P): 2347-6079

modern script, although some conjuncts are known in earlier texts (see “Conjunct consonants” below). The MEITEI MAYEK KILLER does not cause conjunct formation, and is always visible when used. Its use is an optional feature of spelling. The use of the KILLER with letters (like t TA) which have an explicit final consonant(. T) is not attested, and would not be expected because of the existence of explicit finals. In other contexts, the KILLER helps to show the absence of an inherent vowel—while kr may be read either *kara* or *kra*, krB must be read *kra*. When word internal, the glyph of the KILLER typically extends beneath the killed letter and the letter following. A syllable is structured (and represented in the backing store) as follows:



(Vi | (C Vm? F?)), where “Vi” is an independent vowel, “C” is a consonant (including the independent vowel ?A), “Vm” is a vowel matra, “F” is an independent vowel used in final position or a final consonant or ANUSVARA or VISARGA. In the unusual and historic abbreviations described below, the syntax is (Vi | (C Vm* F?)).

Character names:

The name of the script itself has a number of different names and spellings: *Meitei Mayek*, is found alongside *Methei* and *Meetei* as well as the older *Manipuri*. In the modern version of the script, each letter but the 18th letter and other characters with dashes in gloss column are not related to our body parts, is named after a part of the our body.

Table 1. 27 modern letters of Meitei Mayek scripts

Letter	Name	Transliterate	Gloss
K	koK	Kok	Head
S	sM	Sam	Hair
L	laI	Lai	Forehead
M	miT	Mit	Eye
P	P	Pa	Eyelash
N	N	Na	Ear
C	ciL	Cil	Lips
T	tiL	Til	Saliva
S	SO	Khou	Throat
Z	zO	Ngou	Pharynx
H	HO	Thou	Chest
W	waI	Wai	Navel
Y	Yx	Yang	Backbone
H	huK	Huk	lower spine
U	UN	Un	Skin
I	I	I	Blood
F	fM	Pham	Placenta
A	Atiy	Atiya	Sky
G	goK	Gok	----
J	JM	Jham	----
R	raI	Raai	----
B	Ba	Baa	----
J	jiL	Jil	----
D	diL	Dil	----
G	GO	Ghou	----
D	DO	Dhou	----
V	vM	Bham	----

Table : 27 modern letters of Meitei Mayek scrips

Matra of Meitei Mayek :

Meitei Mayek Matra (Cheitap Mayek)				
a–a	i–i	u– u	o– o	e– e
E– ai	O– au	x– am		

Table : Matra of Meitei Mayek script

Digits and punctuation:

Digits have distinctive forms in Meitei Mayek. Five punctuation marks are attested for Meitei Mayek: the | DANDA and ‖ DOUBLE DANDA are used as the sign of full stop but frequently only the DOUBLE DANDA is used. Meitei Mayek has question mark ꫱ but it is not found to be written in school books, newspaper etc. the ꫳ SYLLABLE REPETITION MARK and ꫴ WORD REPETITION MARK seem to have fallen out of use. Generic ASCII punctuation is also expected in Meitei Mayek fonts: ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ ` [\ ] ^ _ { | } ~.

The digits practiced currently are listed as follow:-

Cheising Iyek with English Numeral figure			
1(ama)->1	2(ani)->2	3(ahum)->3	4(mari)->4
5(manga)->5	6(taruk)->6	7(taret)->7	8(nipal)->8
9(mapal)->9	0(phoon)->0		

Table: Numerical figure (Cheising Iyek) of Meitei Mayek script with English Numerical figure

1.2 Hindi Script

The Hindi Script is the writing system for the Hindi language. Hindi inherited its writing system from Sanskrit. The script, Devanagari, is extremely logical and therefore straightforward and easy to learn. Pronunciation is easy because, unlike English, letters are always pronounced exactly the same way. Like other scripts of India, Devanagari also developed from Brahmi script. Through centuries Brahmi developed into different branches. The middle branch of Brahmi came to be known as 'Kutil' script. It again developed into different branches, one of which, began to called as Nagari. The modern form of Devanagari developed from the western from of old Nagari script. Nagari developed in the 10th century. On the basis of the writing of inscriptions of Bhimdev I (1029), Bhimdev II (1200) and Udayavarman (1200) It can be said that the present Devanagari in nearest to the Nagari of the inscriptions. Thus the beginning of Nagari script can be said to be 1000 to 1200 A.D. Later on many changes and amendments also took place. In the 18th century Nagari developed fully and this from is nearer to the present day Devanagari with the exception of some mātrās.

Type of Script:

Devanagari is known as the syllabic script, because its consonant letters includes the neutral vowel (ə), i.e. all the letters are syllabic in character.

Letters:

The letter order of Devanāgarī, like nearly all Brahmi scripts, is based on phonetic principles that consider both the manner and place of articulation of the consonants and vowels they represent. This arrangement is usually referred to as the varṇamālā "garland of letters". The format of Devanāgarī for Sanskrit serves as the prototype for its application, with minor variations or additions, to other languages.

Vowels:

The vowels and their arrangement are:

Independent form

Romanized

As diacritic with प

d

As diacritic with प

***kaṇṭhya*(Guttural)**

अ

A

प

आ

Ā

पा

***tālavya*(Palatal)**

I

पि

ई

Ī

पी

***oṣṭhya*(Labial)**

उ

U

पु

ऊ

Ū

पू

***mūrdhanya*(Retroflex)**

ऋ

ṛ

पृ

ॠ

ṝ
पॄ

dantya(Dental)

ऌ
ḷ

पॢ
ॡ
ḹ
पॣ

kaṇṭhatālavya(**Palato-Guttural**)

ए
E

पे
ऐ
Ai
पै

kaṇṭhoṣṭhya(**Labio-Guttural**)

ओ
O

पो
औ
Au
पौ

kaṇṭhya(Guttural)

आ
Ā
पा

tālavya(Palatal)

ई
Ī
पी

oṣṭhya(Labial)

ऊ
Ū
पू

mūrdhanya(retro)

ऋ

ṝ
पॄ

dantya(Dental)

ॡ
ḹ
पॣ

kaṇṭhatālavya(**Palato-Guttural**)

ऐ
Ai
पै

kaṇṭhoṣṭhya(**Labio-Guttural**)

औ
Au
पौ

Consonants:

The consonants and their arrangement are:

sparśa (Plosive) *anunāsika*(Nasal)

Voicing → *aghoṣa*

Aspiration → *alpaprāṇa* *mahāprāṇa* *alpaprāṇa* *mahāprāṇa*

aṇṭhya(G क ka /k/ ख kha /kh ग

uttural)								
īlavya(Palatal)	च	ca /c	छ	cha /cʰ	ज	ja /ɟ, dʒ/	झ	jha /dʒʰ/
ūrdhanya (Retroflex)	ट	ṭa /ʈ/	ठ	ṭha /ʈʰ	ड	ḍa /ɖ/	ढ	ḍha /ɖʱ/
'antya(Dental)	त	ta /t̪/	थ	tha /t̪ʰ	द	da /d̪/	ध	dha /d̪ʱ/
ṣṭhya(Labial)	प	pa /p	फ	pha /pʰ	ब	ba /b/	भ	bha /bʱ/

tastha(Approximant)				ṣma/saṃghashrī(Fricative)		
Voicing →	aghoṣa	ghoṣa				
Aspiration →	alpaprāṇa	mahāprāṇa				
ṇṭhya(Guttural)	ङ	ṅa /ŋ/		/ɦ/	ह	ha/h
ālavya(Palatal)	ञ	ña /ɲ/	य	ya /j/	श	śa
ūrdhanya(Retroflex)	ण	ṇa /ɳ/	र	ra /r/	ष	ṣa
antya(Dental)	न	na /n/	ल	la /l/	स	sa
ṣṭhya(Labial)	म	ma /m/	व	va /ʋ/		

Punctuation:

The end of a sentence or half-verse may be marked with a dot known as a pūrṇa virām or a vertical line danda: ।. The end of a full verse may be marked with a two vertical lines: ॥. A comma, or alpa virām, is used to denote a natural pause in speech.

Numerals:

Devanāgarī numerals:

०	१	२	३	४	५	६	७	८	९
0	1	2	3	4	5	6	7	8	9

2 Related works

So far works on Transliteration of Indian language using different technique are found. Works of IT3 is a transliteration scheme developed by IISc Bangalore, India and Carnegie Mellon University with the primary focus on user readability of the transliteration scheme [1]. Other transliteration works for Indian languages can be seen in [2], [3], [4], [5] etc. For other foreign languages works of European languages in [6], works on some Asian in [7], English to Korean in [8] etc. Although Manipuri is a eight schedule language of India very few works of transliteration is being recorded to the best of the authors knowledge. May be it's because of the Tibeto-Burman origin or of its agglutinative in nature.

3 Linguistic Transliteration Scheme

Transliteration from Hindi Script to Meitei Mayek needs to map its corresponding alphabets from one script to another. Hindi Script altogether there is 29 non-aspirated, and 15 aspirated consonants, 11 vowels and 10 numeric symbols. On the other side that is Meitei Mayek we have 27 (Twenty seven) alphabets and its supplement (use of Lonsum, Cheitap-Cheikhei, Khudam, Cheising etc) as per annexures as recommended by the committee in the educational institute in Manipur.

Table 1: Iyek Ipee characters in Meitei Mayek.

Iyek Ipee	
k->k(Kok)	ˆ,x,z,s>s(Sam)
l->l(Lai)	m->m(Mit)
p->p(Paa)	Å,n,,->n(Naa)
c->c(Cheen)	t,$->t(Til)
% -> S(Khou)	; -> z(Ngou)
q,# ->H(Thou)	v -> w(Wai)
y,yý –> y(Yang)	h ->h(Huk)
£,¤ -> U(Uoon)	š,i -> I(Ee)
f -> f(Pham)	A -> A(Atiya)
g -> g(Gok)	& ->J(Jham)
r->r(Rai)	b -> b(Baa)
j -> j(Jil)	D,@ -> d(Dil)
` -> G(Ghou)	!,/ -> D(Dhou)

Table 1 shows the Ipee Iyek which sounds like the consonant of the Hindi Script, table 2 gives details of the vowel letters used in Meitei Mayek which is followed by Cheitap Iyek of Meitei Mayek in table 3.

Table 2. Vowels of Meitei Mayek.

vowel letter
Aa->Aa(Aa)
E>Ae(Ae)
Ee-> AE(Ei)
A'-> Aq(Ang)
Ao-> Ao(Oo)
AO-> AO(Ou)

Table 3. Cheitap Iyek of Meitei Mayek

Cheitap Iyek	
a -> a(atap)	i , I -> i (iinap)
U, U ->u(unap)	e-> e(yenap)
W-> E (cheinap)	o-> o(otnap)
O->O(sounap)	'-> q(nung)
	-

The numerical figures are known as Cheising Iyek and Table 4 is shown with their respective Hindi Script numerical figure.

Table 4. Cheising Iyek or numerical figures of Meitei Mayek

Cheising Iyek(Numeral figure)	
1->1(ama)	2->2(ani)
3>3(ahum)	4->4(mari)
5>5(mang)	6->6(taruk)
7>7(taret)	8->8(nipal)
9>9(mapa)	10->10(tara)

Alphabets of Meitei Mayek are less in number and the use of it against Hindi Script sometimes shows a repeated uses of the same alphabet for different Hindi alphabet. For example ˆ, *s, x, z* in Hindi is transliterated to *s* in Meitei Mayek and so on.

Table 5:lonsum iyek of meetei mayek

Lonsum Iyek
K-> K (kok lonsum)
L-> L (lai lonsum)
P-> P(pa lonsum)
M->M (mit lonsum)
N-> N (na lonsum)
T-> T (til lonsum)
;->Z(ngou lonsum)
i,š ->I(ee lonsum)

In Meitei Mayek, Lonsum Iyek (in Table 5) is used when *K* is transliterate to *K*, *;* transliterate to Z, *T* transliterate to *T* etc. So as an example when *k* is used in a word it uses *k* but use as *K* when *K* is used. Apart from the above character set Meitei Mayek has symbols which are different from Hindi Script like '-' (Cheikhie) for '|' (full stop in Hindi Script).

4 Model and Algorithm

In our model (Fig. 1) we used two mapping file, one which consist of Hindi Characters List and the other which consists of corresponding Meitei Mayek Character List. They are read and stored in the *Hin_arr* and *Mm_arr* arrays respectively. *Hin_Char_count* is the total number of character in the list. A test file is used so that it can compare its *index* of mapping in the Hindi Characters List file which later on used to find the corresponding Meitei Mayek Characters Combination which is after all the *target Meitei Mayek Characters Combination.* The output transliterated Meitei Mayek Character Combination is written on the output file *OutPutFile.*

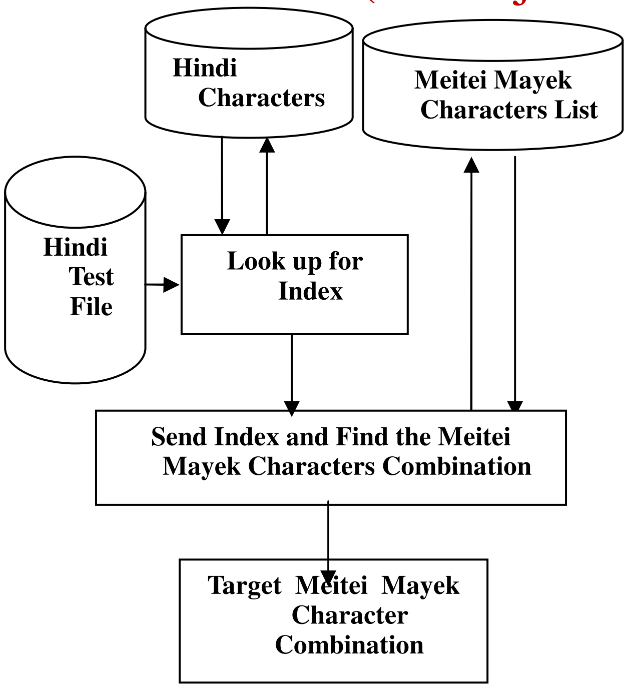

Hindi Characters
Meitei Mayek Characters List
Hindi Test File
Look up for Index
Send Index and Find the Meitei Mayek Characters Combination
Target Meitei Mayek Character Combination

Fig 1. Model of Transliteration Scheme used in transliteration from Hindi Script to Meitei Mayek.

Algorithm use for our model is as follows:

```
Algorithm:trans(Str    line,    Hin_Char_count
   Mm_arr[],Hin_arr[])
{
Try
{Str result="";
FileOutputStream       fos       =        new
   FileOutputStream("OutPutFile.txt",true);
BufferedWriter bw = new BufferedWriter(new
   OutputStreamWriter(fos,"UTF-8"));
int len=line.length();
int m=0;
int pos=0;
Str tline="";
while(m<len){
tline=line.substring(m,m+1);
if(tline.equals(" ")){
bw.write(" ");
}
else{
for(int
   index=0;index<Hin_Char_count;index++)
{
if(tline.equals(Hin_arr[index]))
{      pos = index;
break;
}
}
bw.write(Mm_arr[pos]);
}
m++;
}
bw.newLine();
bw.close();
fos.close();
}catch(IOException f)
{System.out.println(f);}
```

5 Experiment and Evaluation

Hindi is a less computerize language and collecting corpus is a hard task. The experiments of the systems are done with the corpus collected from a hindi newspaper, hindi books etc. A corpus of 20,687 words is collected for testing of the system. Filtering of spelling mistakes and improper syntax are checked manually by a linguist so that maximum output is yielded.

In Evaluation of the result, the system shows an accuracy of 86.28%. Due to use of same character set of the Meitei Mayek relative to Hindi Script as mention in Section 3 we found a lower accuracy.

6 Conclusion

The project aims to develop a model and design an algorithm for transliterating from Hindi language scripts to the Manipuri's language scripts. So we have studied about Manipuri language, its scripts i.e Meitei Mayek (Manipuri own script) and Hindi script.More of the above, we studied about related works to transliteration for different authors in different language' scripts with various algorithms, models, approaches. After surveying related works, such as Meitei Mayek to English (Roman) Script (A Rule Based Approach) and English (Roman) Script to Manipuri Script (A Dictionary Spell Based Approach) , Bengali Manipuri Script to Meitei Mayek (A Rule Base Approach) in different chapters. we are able to design our own models, algorithms for transliteration from Hindi Script to Meitei Mayek (A Rule Base Approches). The project will enable the one who do not know Meitei Mayek (Manipuri script) for reading and writing by using the algorithm and models. It also enables to know how a Hindi word will pronounce or sound. In addition,it will overcome the difficulty of typing Meitei Mayek for the corresponding Hindi script for newspaper publications as the algorithm can produce the trans-scripted output of Meitei Mayek for the input Hindi script file in a very short time.The models and algorithms result in less accuracy in the output of our system. Although it might be considered as a best models and algorithms for the transliteration since these are of first attempts. So we keep our objective to improve the accuracy of our system output and designing new models and algorithm is our future work.

References

[1]Ganapathiraju, M., Balakrishnan, M., Balakrishnan, N., Reddy, R., Om.: One Tool for Many (Indian) Languages, In: International conference on Universal Digital Library (ICUDL), Hangzhou, China. Journal of Zhejiang University SCIENCE, **6A**(11):1348-1353 (2005)

[2]Surana, Harshit, Singh, Anil Kumar.: A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages. In: 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), pp. 64-71, India (2008).

[3]Das, A., Ekbal A., Tapabrata, Mondal, Sivaji, B.: English to Hindi Machine Transliteration at NEWS 2009. In: ACL-IJCNLP 2009, Singapore (2009).

[4]Ekbal, A., Naskar, S. K., Bandyopadhyay, S. :A Modified Joint Source-Channel Model for Transliteration. In: COLING/ACL on Main Conference Poster Sessions. Association for Computational Linguistics, pp. 191–198, Morristown, NJ, USA (2006)

[5]Ekbal, A., Naskar, S., Bandyopadhyay, S..: Named Entity Transliteration. International Journal of Computer Processing of Oriental Languages (IJCPOL), Volume (20:4), pp.289-310, World Scientific Publishing Company, Singapore (2007)

[6]Marino, J., B., Banchs R., Crego J., M., de Gispert, A., Lambert, P., Fonollosa J., A., Ruiz., M.: Bilingual n-gram Statistical Machine Translation. In: MT-Summit X, pp. 275–282 (2005)

[7]Vigra, P., Khudanpur, S.: Transliteration of Proper Names in Cross-Lingual Information Retrieval. In: ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition, pp. 57–60, (2003)

[8]Kangjia Mangang, Ng.: Revival of a closed account. Sanamahi Laining Amasung Punsiron Khupham, Imphal (2003)

[9]Michael Everson,"Proposal for encoding the Meitei Mayek script in the BMP of the UCS"

[10]lekha Mishra, Ananthakrishnan R, Sasikumar M, "Automatic Derivation of Rules for Transliteration from English to Hindi: a Genetic Algorithm Approach".

[11]Hindi Language-free wikipedia , the encyclopedia

[12]http://www.karr.net/hindi_script/encyclopedia.htm

[13]Kishorjit, N. Herojit, N. Sonia, Th. Shinghajit, Kh. Sivaji, B:Transliteration of Manipuri : Meitei Mayek to English Script. In: International Conference on language Development and Computing Methods (ICLDCM 2010) Department of English & Department of Information Technology, Karunya University , Coimbatore, Dec-2010, pp.240-243.

[14]Kishorjit, N. Herojit, N. Sonia, Th. Shinghajit, Kh. Sivaji, B: Manipuri Transliteration from Bengali Script to Meitei Mayek : A Rule Based Approach. In: Information Systems for Indian Languages International Conference, ICISIL 2011, pp. Patiala, India, March 9-11,2011.