Romanization of Kannada Text

Karthik Kashyap
Dept. of Information Science and Engineering
RV College of engineering
Bangalore,India
kashyapkarthik3@gmail.com

Kuche Bhavani Priya
Dept. of Information Science and Engineering
RV College of engineering
Bangalore,India
bhavani.priya156@gmail.com

Abstract*—In this paper we analyze the translation of kannada text to english. We discuss the method which is in use today. *We believe the idea we have presented here will also be of interest to people in many other countries where the language situation is similar.

INTRODUCTION

Language consists of the maintainance, development and complex system of communication. Either written or spoken, used by particular country or community. Script is a form of written document where people use it to store or communicate. Romanization is a conversion of one writing script to other writing system.

In translation, text from one language are mapped to texts in a different language. The script used for obtaining these texts is not given a lot of importance. The source language text as well as the target language text can be in any suitable script.

Why Romanization of kannada is necessary?

we write texts in Indian languages using the Roman alphabet. This process is known as Romanization. This can be bidirectional and if so, we can actually map from any script to any other script via Roman. The idea presented in this paper are generic and applicable to other language and script scenarios anywhere in the world.

ಸಿಂಹ ಮತ್ತು ಮುದಿ ಮೊಲ. (1)

To translate this to romanized english script, we need to first split the words into single root characters.Each root character will have their own specific unicode which can be obtained by using hex(ord(char)) function.

When we try to split the first word of statement(1), number of unicodes obtained might be more than the characters displayed. Such as: ಸಿಂಹ - ಸ+◌ಿ+ ◌ಂ+ಹ will be splited in this way. When translated to english script it returns sa+i+M+ha making the word saiMha. But the target word is siMha.

For some letters in kannada which needs some effort in pronounciation for example: Kha, M, Tta,etc. We use h to indicate its heavy pronounciation and M for indicating anusvara.

Advantages.

1) To obtain english script of kannada language using suitable rendering algorithm.
2) It can help readers who does not know kannada script but understands kannada.
3) Processing of texts rendered in different scripts may require different techniques for dealing. It can be simple, direct, natural and efficient.

In order to simplify the characters of kannada language we make use of python dictionaries. We create two dictionaries in which one contains all the consonants and other contains vowels. Consonants dictionary contains numericals and special symbols. These dictionaries does not contain the kannada character themselves instead they unicode values of kannada characters in string form as keys. These keys will have their respective english character in string form as their values.

With the help of these dictionaries we are able to transform their kannada characters to their english characters. When a kannada character is taken as input their respective unicode value is obtained and is matched with the one of the key in the dictionary. If there is a match then their value will be returned as the output,if not then the loop terminates.

If the unicode of the current character is present in consonants and if the next character unicode is not present in vowels then the key of the current unicode is written to the file. If the unicode of the current character is present in consonants and if the next character unicode is present in vowels then the last letter of the key element of the current character unicode is removed and appended with the key value of the next character unicode.

Even the special symbols like Full-stop(.),Comma(,) etc have their respective unicode as that also has to be printed in the target file. When we encounter hex(ord(char))= ‘0xccd’ which is ◌್ we just continue the code as usual. Because this pronounciation matters but is not considered in written text.

Consonants table:

	U+0C9x	
5	ಕ	Ka
6	ಖ	Kha
7	ಗ	ga
8	ಘ	gha
9	ಙ	nga
a	ಚ	ca
b	ಛ	cha
c	ಜ	ja
d	ಝ	jha
e	ಞ	nya
f	ಟ	TTa

Vowels table:

	U+0C8x	
2	◌ಂ	anusvara
3	◌ಃ	visarga
	Reserved	
5	ಅ	a
6	ಆ	aa
7	ಇ	i
8	ಈ	ii
9	ಉ	u
a	ಊ	uu
b	ಋ	r
c	ಌ	l
d	Reserved	
e	ಎ	e
f	ಏ	ee

Conditions to be taken care of:-

1)A single input consisting of a consonant that is followed by another consonant will be displayed without any modification to the string value of that input key.

ಕಗ will be translated as kaga.

2)An input consisting of a consonant and a vowel will be displayed by discarding the last character of the value string for the consonant key and appending the prefix to the vowel value.

ಸಿ will be split into ಸ and ◌ಿ.

This will be displayes by removing a from sa and appending I to it. The output will be si.

3)When a ottakshara is given as input, it consists of three to four unicode,where the first and the third will be consonants and the second and the fourth will be vowel. The final output will be obtained by removing the last character of the first and the third(if the fourth character is given) consonant and appending to the value of the key for the second and the fourth(if given) vowel.

ಪ್ರ has 3 characters :ಪ,◌್ and ರ. Their equivalent pa and ra. Thus the ‘a’ in pa will be removed and the final word pra will be generated.

M is pronounced as the nasal sound corresponding to the row of the consonant that follows it in the given word. Thus,

Mk is pronounced as nGk,

Mc is pronounced as nYc,

MT is pronounced as NT,

Md is pronounced as nd,

Mb is pronounced as mb.

For the unclassified consonants, M is pronounced as m. when preceded by the vowel a or aa, as hi when preceded by i, ii or ai, as hu when preceded by u, uu or au, as as he when preceded by e or ee.

References:

1)Roman Transliteration of Indic Scripts,Kavi Narayana Murthy, Srinivasu Badugu,*Department of Computer and Info.Sciences, University of Hyderabad*

2) https://en.wikipedia.org/wiki/Kannada_(Unicode_block)

3)https://en.wikipedia.org/wiki/Language