Unicode as the Basis of Transcription of Indic Scripts

Sheena Agarwal[#1], Rhythm Agarwal[*2]

#1 Birla Institute of Technology, Mesra-835215, Ranchi, India
**2BMS College of Engineering, Bengaluru-560091, Karnataka, India*

Abstract

The research work done deals with transliteration of Roman characters into various Indian languages. This involved a thorough study of Unicode, a standard that has been implemented all over the world for transliteration of 129 different scripts. This universal standard has without doubt made communication easier and hence, enhanced the exchange of information. The paper briefly explores topics like phonetic transcription and Unicode consortium and proposes means to implement the two to achieve transcription of a few Indic scripts. The main objective of the paper is to provide ample of opportunity to the regional ancient languages of India, to develop a perfect synchronism with the modern information technology.

Keywords-*Unicode Standard, Phonetics, Transcription, Devanagari.*

I. INTRODUCTION

Systematic transliteration is the process of mapping from one system of writing into another, typically grapheme to grapheme. Most transliteration systems are one to one, so that a user who is aware of the system can reconstruct the original spelling. We wish to implement the same for various Indian languages using phones, which are the visual representation of speech sounds. To achieve this goal we need a universal encoding scheme which is Unicode.

II. THE UNICODE STANDARD

The Unicode Standard is the universal character encoding standard used for the representation of text characters for computer processing. Fundamentally, computers just deal with numbers. They store characters by assigning different patterns of numbers to them. Before the advent of Unicode, no single encoding scheme could contain enough characters. Hence, there existed a conflict among the various encoding systems.

Unicode provides a consistent way of encoding multilingual text and thus, provides a solution to the exchange of text files internationally. Unicode standard defines a unique number, known as code point, for every character used in all major languages written today. Incorporating Unicode into client-server applications offers significant cost savings as it enables a single software product to be targeted across multiple platforms, languages and countries without any form of re-engineering.

Traditionally, computers used the American Standard Code for Information Interchange (ASCII) to represent text information since the standard was made for English language. This was a 7-bit code which was able to represent 128 characters. Although, this sufficed the basic need but curiosity for more led to its extension. The 8-bit Extended ASCII, which could represent 256 characters allowed for the inclusion of some other Roman characters in the code. As others, especially non-Latin characters, needed representation in the computer and hence, there was the need of a standardization effort so as to avoid a situation where multiple characters use the same code.

The UNICODE standard is an attempt to avoid such a chaos by assigning a unique code point for every character of every conceivable language independent of computer platform and application for which the textual data is being used. The Unicode Consortium completely endorses the use of any of three encoding forms as a way of implementing the Unicode standard. Each of these encoding schemes are discussed briefly.

III. ENCODING SCHEMES

Character encoding standards represents the identity of each character along with its numeric value, known as the code point. The Unicode standard defines three such mapping methods each of which has its own benefits and shortcomings.

A. UTF-32

Unicode Transformation Format 32 bits is a protocol that encodes characters using exactly 32 bits per Unicode code point. As a result, UTF-32 is a fixed length encoding scheme unlike the others which are variable length formats. Since it directly maps the code points to a sequence of 4 bytes, the code points are directly indexable , which is one of the major advantages of this encoding scheme. The use of 4 bytes per character make the scheme quite space inefficient thus, leading to the evolution of UTF-16.

B.UTF-16

This encoding scheme is used to map the code points of the Unicode characters into stream of bytes for the purpose of communication and storage using a sequence of 2 bytes (16 bits).It is capable of encoding all 1,112,064 possible characters in Unicode. The most significant advantage of UTF-16 over UTF-32 is that it allows for variable-length encoding, as the code points can be encoded with either one or two 16-bit code units. Although, it provides for some of the flaws of UTF-32 but the problem of space-inefficiency still persists as a file encoded in UTF-16 occupies almost twice the size of an ASCII file.

C.UTF-8

It is the most dominant character encoding scheme which is capable of encoding all the valid code points in the Unicode code space using one to four 8-bit code units. Under this mapping scheme the first 128 characters of Unicode are encoded using a single 8-bit unit with the same binary value as ASCII making UTF-8 fully backward compatible with ASCII unlike UTF-16 and -32.For text requiring a single byte for mapping the size of file remains same as that of an ASCII file. But, for certain scripts like Thai and Devanagari , it uses 2 or more blocks of 8-bits for encoding purpose and hence, the size of the file increases greatly. Given to its numerous pros, it is the most preferred encoding technique accounting for a major proportion of all the web pages available on the World Wide Web.

IV. PHONETIC TRANSCRIPTION

The branch of linguistics that deals the sounds of speech and their production, combinations and description is known as phonetics and the visual representation of these sounds is known as phonetic transcription. It is the process which displays a one-to-one relationship between character symbols and their sounds.

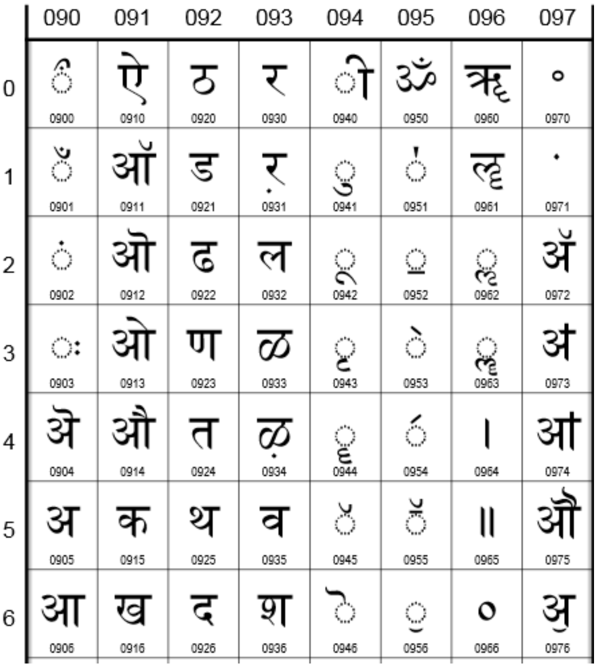
For some specific pair of source ad target script, a system of transliteration maps the letters of the source script to letters pronounced similarly in the target script. Although transcription and transliteration are two different processes, but if the relations between sounds and letters are similar in two languages, transcription may be very close to transliteration.

The phonetic transcription of a particular script involves producing the pronunciation corresponding to the text in the entry of the list of phonemes. The phonetization of a text is a tedious task as there is not always a direct correspondence between the graphemes and the phonemes.

V. TEXT PROCESSING

In an attempt to make communication easier and enhance exchange of information, we wish to realize the phonetic transcription of a few Indian languages like Hindi, Marathi, Nepali, Sanskrit and many more all of which are included in the Devanagari script. Each character represents a unique sound hence, assisting the process of transcription.

The characters of the Devanagari script do not belong to the ASCII code therefore, we take into account the Unicode standard for the process of transcription. This allows to code all characters used by the Devanagari script and to exchange data of text between different platforms and systems.

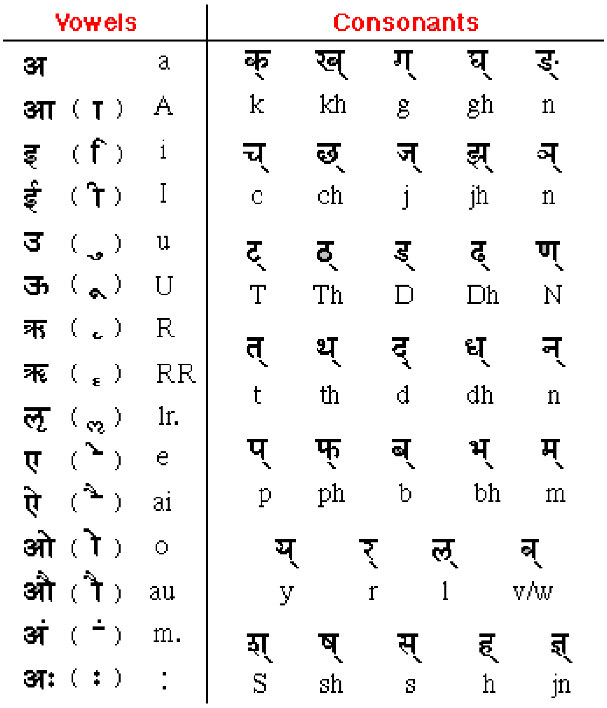


	090	091	092	093	094	095	096	097
0	◌ऀ 0900	ऐ 0910	ठ 0920	र 0930	◌ी 0940	ॐ 0950	ॠ 0960	॰ 0970
1	◌ँ 0901	ऑ 0911	ड 0921	ऱ 0931	◌ु 0941	◌॑ 0951	ॡ 0961	ॱ 0971
2	◌ं 0902	ऒ 0912	ढ 0922	ल 0932	◌ू 0942	◌॒ 0952	◌ॢ 0962	ॲ 0972
3	◌ः 0903	ओ 0913	ण 0923	ळ 0933	◌ृ 0943	◌॓ 0953	◌ॣ 0963	ॳ 0973
4	ऄ 0904	औ 0914	त 0924	ऴ 0934	◌ॄ 0944	◌॔ 0954	। 0964	ॴ 0974
5	अ 0905	क 0915	थ 0925	व 0935	◌ॅ 0945	◌ॕ 0955	॥ 0965	ॵ 0975
6	आ 0906	ख 0916	द 0926	श 0936	◌ॆ 0946	◌ॖ 0956	० 0966	ॶ 0976

Table 1. Unicode Standard for a few Devanagari characters.

Thus, it is quite clear that Unicode consortium assigns a unique code point to each of the character of the script and that too in the same sequence in which the characters were traditionally arranged.

The difference between identifying a code point and rendering it on the screen is quite crucial in understanding the role of Unicode in text processing. The visual representation of the character-called a glyph-is not defined by the Unicode standard, it only decides how the characters are to be interpreted. To resolve this issue for various Indic scripts, phonetic transcription comes to rescue. The text elements are encoded as sequences of one or more characters. Certain characters of the script being processed are represented as a single character while others are represented using a combination of characters. Following is the table that represents such sequences for Devanagari script.



Vowels		
अ		a
आ	(ा)	A
इ	(ि)	i
ई	(ी)	I
उ	(ु)	u
ऊ	(ू)	U
ऋ	(ृ)	R
ॠ	(ॄ)	RR
ऌ	(ॢ)	lr.
ए	(े)	e
ऐ	(ै)	ai
ओ	(ो)	o
औ	(ौ)	au
अं	(ं)	m.
अः	(ः)	:

Consonants				
क्	ख्	ग्	घ्	ङ्
k	kh	g	gh	n
च्	छ्	ज्	झ्	ञ्
c	ch	j	jh	n
ट्	ठ्	ड्	ढ्	ण्
T	Th	D	Dh	N
त्	थ्	द्	ध्	न्
t	th	d	dh	n
प्	फ्	ब्	भ्	म्
p	ph	b	bh	m
य्	र्	ल्	व्	
y	r	l	v/w	
श्	ष्	स्	ह्	ज्ञ्
S	sh	s	h	jn

***Table 2. Phonetic equivalents for Devanagari characters*.**

Hence, we observe that due to the phonetic sounds of various characters and the presence of a unique code point for each character, test processing of various Indic scripts like Hindi, Bengali, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil and Telugu becomes quite convenient and efficient.

VI. CONCLUSION AND FUTURE WORK

This paper is an attempt to demonstrate that with the current implementation of Unicode, it is possible to provide widely globalized products for the Indic market. However, there are certain deficiencies in the Unicode Standard which act as barrier to it being used in the production of culturally and linguistically appropriate products for the Indian market. One such concern has been that the Unicode for Indian languages are too Devanagari based. As such, changes to better represent non-Devanagari languages like Tamil or Kannada need to be proposed. Like other script repertoires in Unicode, it might take some time to refine the set of these characters and make the changes necessary to fully satisfy the linguistic community. In conclusion, this paper shows that Unicode as an encoding scheme when used in tandem with other processes like phonetic transcription is more than sufficient to support Indic scripts and languages.

REFERENCES

1. Marcus Otlowski, Pronunciation:What are the expectations?,The Internet TESL Journal,Vol.IV,No.1.
2. Notice of Retraction A kind of Chinese language Phonetic input output system code,Lu Qiao;Wan Pu;Zhang Li;Zhu Daoyong,IEEE, Vol.9,Pages 508-511,2010.
3. Audio Visual based pronunciation dictionary for Indian languages. Palanisamy,K,Technology for Education(T4E),pages 82-84,2010.
4. Hamad , M.Hussain , ArabicText-to-Speech Synthesizer,IEEE Student conference on Research and Development,pages409-414(2011).
5. “Issues in Corpus creation and Distribution:The Evolution of Linguistic Data Consortium”,Cristopher Cieri,Mark Liberman,University of Pennsylvania and Linguistic Data consortium Philadelphia,Pennsylvania,USA.
6. Ramani.G,Menakambal.S,“Advanced RISC machine based Data Acquisition Development and Control”,SSRG-IJEEE,,vol. 1,issue 8,2014.
7. Ruhlen Marritt,A Guide to World’s Languages,Vol.1, Page 463,1991.