THE DEVELOPMENT OF A SUITABLE SCRIPT—AN INPUT PROBLEM IN THE ANALYSIS OF INDIAN LANGUAGES USING THE IBM 360/44 COMPUTER*

P. C. GANESHSUNDARAM

(*Assistant Professor, Foreign Languages Section, Indian Institute of Science, Bangalore*)

Received on July 15, 1975 and in revised form on November 30, 1975

Key words: Romanized spelling, Indian languages, Computer analysis.

I. INTRODUCTION

In this paper we examine a suitable Romanized script for Indian languages for being processed by the IBM 360|44 computer. The Romanization scheme adopted by us for our study of the Indian languages is given towards the end of this paper.

Although the codification of written characters in the various languages is one of the initial input problems for any study of natural languages through the computer, it would be of interest also to survey the conventional spellings or orthographic systems of different natural languages.

We, therefore, examine also here the state of affairs concerning mesures taken in adopting a common script in different parts of the world, as it obtains today. We shall not concern ourselves with the historical side of the question.

We shall however see how in different areas of the world an attempt to adopt a common script has been in progress. We shall also mention the fact that the Roman script has proved itself to be of advantage in making use of the computer for the analysis of Indian languages.

* This is a revised version of a paper read as a 'Key Note' paper at the Seminar for Formulating a Common Script for the Dravidian Languages on 25th/26th February 1974, at the Bangalore University, under the title: "Attempts to Introduce a Common Script in other Countries—and Suggestions for a Spelling System for Indian Languages based on the Roman Script".

1. MAJOR SCRIPTS IN USE IN THE WORLD FOR MORE THAN ONE LANGUAGE

(1) The Sino-Japanese Ideographs,

(2) The Perso-Arabic Script,

(3) The Cyrillic Script,

(4) The Devanagari Script, and

(5) The Roman Script,

are the major scripts of the world that are used for more than one language today.

We shall examine below how each script has brought together a number of languages into a common group.

2. THE SINO-JAPANESE IDEOGRAPHS

When we see Chinese or Japanese scientific writing, we find that most of the technical terms seem to be the same in the two languages. Still, there are differences:

(1) In Japanese, in addition to the ideographs (*kanji*), there is a syllabic script employed to indicate grammatical endings (*hiragana*) and another syllabic script (*katakana*) to indicate loan words and proper nouns (especially foreign names). All these three scripts are found in the same piece of writing.

It is the ideographic part (*kanji*) that is common to Chinese and Japanese.

However, *kanji* writing is so complicated for memorizing and for modern purposes of typing, teleprinting, composing types, etc., that complex characters of *kanji* are now being simplified. The simplifications done in Japan, however, are different from those done in China. In this respect the *kanji* used in the two languages shows a tendency for divergence

In Japanese, many *kanji* characters are dropped and, wherever such a replacement is possible, *katakana* is used instead.

(2) In Chinese, it is the ideograph that gives the illusion that Chinese is one language. Written Chinese, the visual form of it, is one. But, the moment the written version is read aloud in different parts of China, we

hear entirely different languages or dialects. Some of these dialects are not even mutually intelligible among the Chinese themselves, when spoken.

It has been realised, both in China and Japan, that for the modern technological age the ideographs do not serve as an ideal graphic system. A phonetic or phonemic transcription should be much more advantageous.

In both China and Japan the question of introducing the Roman script has been seriously and actively considered. In dictionaries and grammar books the Roman script along with *kanji* is universally used in Japan.

However, the attempt at introducing the Roman script has met with genuine linguistic difficulties in both the languages. For, in both of them there are scores of *homonyms*. It is only the ideograph that comes to the rescue when the phonetic shape difies an interpretation. These homonyms stand in the way of the Chinese and the Japanese, who want to change over to the Roman script.

3. The Perso-Arabic Script

The spread of this group of scripts in the world is intimately connected with the spread of Islam.

It is a script that gives prominence to the consonants, while the vowels are indicated, if at all, by diacritical marks. In the Semitic languages, the root form of a word is represented by a set of consonants, and the vowels, even if different, could be omitted in writing. The reader supplies the correct vowel from context when he reads.

Although this script is specifically suited to the Semitic group of languages, it was adopted, owing to the spread of Islam, by other people, even if their languages were non-semitic and were of a different structure. Persian, Urdu and Malay are the prominent ones among these.

4. The Cyrillic Script

This is one of the scripts said to be derived, like the Roman and Greek scripts, from the old Phoenecian system of writing.

It is now the official script of modern Russian. Variants of this script are found in Bulgarian, Ukrainian and a number of minor languages of the Soviet Union that were never written before.

However, some major languages of the Soviet Union, like Armenian, have their own traditional scripts.

One of the languages of Yugoslavia is written both in the Cyrillic script as well as in the Roman script.

A few languages of the Soviet Union are written in the Roman script. It is said that even for them there is a tendency to switch over to the Cyrillic script as a result of official policy.

5. The Devanagari Script

This is one of the scripts said to be derived from the old Brahmi in the North of India to write Sanskrit and other Indo-Aryan languages.

Although the Gujarathi, Gurumukhi, Bengali, Assamese and Oriya scripts are different from one another, they could all be considered as variants of the Devanagari script.

6. The Roman Script

The most widespread among the scripts of the world today (language-wise, if not population-wise) is the Roman script.

Most languages of the Indo-European family in Europe use the Roman script.

Most languages using the Roman script have adopted a *spelling* system, making use of a combination of letters to indicate particular sounds. Some others have sought to increase the stock of symbols by employing diacritical marks. Still others use a combination of spellings and diacritics.

Thus, most West-European languages and Polish (a Slavic language) use a spelling system with a minimum of diacritical marks, as in French.

Most other languages of Europe, such as the Slavic languages other than Polish, a Romance language like Rumanian and non-Indo-European languages like Finnish, Hungarian and Turkish use the Roman script, making use of a liberal sprinkling of diacritical marks.

The Roman script or a modification of it has been adopted for a few languages of Africa, for Indonesian and for some languages of the Philippines.

Indonesian is especially noteworthy, as many publications have now been brought out, including an encyclopedia, entirely in the Roman script. This has helped to unify the intellectual community, that formerly was partially divided by the use of the Javanese script by one section and the Malay (Arabic) script by another. The language itself is very close to Malay with free borrowings from Javañese, Dutch, French, English, etc.

Even within India, the Roman script has been used for a very long time to write Urdu in the British Indian Army (Roman Urdu) and in Portuguese Goa for a language like Konkani (Concanim). Konkani has been written in the Devanagari, Kannada and Roman scripts.

7. Advantages and Disadvantages of Evolving a Common Script for the Dravidian Languages based on the Dravidian Scripts Themselves

(1) *Advantages*

A common script of *whatever* origin would be an advantage in giving a common look to all the Dravidian languages, which are intrinsically related to one another and are divided mainly through their differing scripts and to a lesser extent by the differences in the percentage of words of Sanskrit and other origin.

(2) *Disadvantages*

(i) A new Dravidian script will have to be evolved afresh, as no linguistic group would willingly accept the script used by any of the other languages of the family. The objection would not be on practical grounds, but basically on emotional grounds.

(ii) If a separate pan-Dravidian script is evolved, it would perpetuate a difference in the writing systems of Dravidian on the one hand, and all other languages, on the other.

8. Advantages and Disadvantages of Adopting the Devanagari Script as a Common Script for the Dravidian Languages

(1) *Advantages*

All the languages of India will have a common 'Indian' script, giving emotional satisfaction to the majority of Indians that are emotionally attached to the Devanagari script or to Sanskrit or both. (Questions of efficiency, economy, etc., wouldn't bother them.)

(2) *Disadvantages*

(i) The adoption of the Devanagari script would cause emotional friction between those who claim equality of antiquity to their own scripts as against the Devanagari script.

(ii) The Devanagari script has been proved, through statistical investigations, to be uneconomical for writing, typing or printing in terms of muscular effort, time and cost.

(iii) Irrespective of the claim that the Devanagari script is 'scientific' and 'phonetic' [which is true only in so far as the grouped consonants (vargas) represented by it in Sanskrit are concerned], it has however the following discrepancies:

(*a*) The same sound is frequently represented by a number of different graphic symbols:

The initial vowel symbol *versus* the final vowel symbol (after a consonant) is one such case.

The symbol for a consonant like *r*, when followed by a vowel, is different from the symbol for the same consonant, when followed by a consonant. It differs again when preceded by a consonant. When different consonants precede it, it has different shapes.

(*b*) All modern ultrarapid communication systems are linear in their operation. Even the computer linearizes two-dimensional pictorial representations (chemical formulae, photographs, etc.). Television pictures are linearly scanned and reassembled. Under these conditions the efficiency of typing, printing, teleprinting, computer processing, etc., of a script with diacritical marks above, below and across a letter would be much lower than desired.

A *linearized* representation would be advantageous instead.

9. ADVANTAGES AND DISADVANTAGES OF ADOPTING THE ROMAN SCRIPT FOR ALL THE INDIAN LANGUAGES

(1) *Disadvantages*

(i) Those of us who are too much attached to our own scripts may have to undergo a lot of emotional readjustment :

(ii) Those of us who feel that we would be ridiculed by foreigners for having adopted a foreign script may have to face their own emotional conflicts.

(Turkey and Indonesia haven't been ridiculed for adopting the Roman script. They enjoy all the advantages of using the universally used Roman alphabet.)

(iii) 'One letter—one sound' equivalence may not be found in the Roman script adapted for Indian languages.

(iv) If a system of diacritical marks is used (as is done in citing Indian language examples in learned articles written in English), all the disadvantages mentioned above about the Devanagari script could be found to be present here too.

(2) *Advantages*

If a *spelling system* (rather than a system of diacritical marks) is adopted for the use of the Roman script in writing the Indian languages, we shall find that:

(i) Even if a 'one letter—one sound' condition may not obtain, one *combination* of letters for one sound (or phoneme) could be established.

(ii) The linearity of the script will not be violated.

(iii) The letters of the Roman script have such simple shapes that details of these letters are visible and the letters legible from a longer distance than the multiflected three tiered letters of most Indian scripts and especially the Devanagari.

(iv) Universally available machinery (international Roman typewriters, teleprinters, computers and linotype machines, etc., with a very simple keyboard) could be directly used for printing books in any Indian language without having to litter printing presses with a variety of types and machines; reference books of bilingual and multilingual character could be compiled with the help of the computer and automatically arranged alphabetically, without having to rewrite them in any other script.

(v) All the Indian languages could be typed on one machine.

(vi) The Roman script, although 'foreign', is emotionally neutral interlinguistically among the Indian languages.

10. ROMAN SPELLINGS FOR THE VOWELS AND CONSONANTS OF INDIAN LANGUAGES*

The chart given below shows the common spellings of the vowels and consonants of the Indian languages:

Roman spelling	Sanskrit	Hindi	Malayalam	Tamil	Telugu	Kannada	Remarks
A	अ	अ	അ	அ	అ	ಅ	
AA	आ	आ	ആ	ஆ	ఆ	ಆ	
I	इ	इ	ഇ	இ	ఇ	ಇ	
II	ई	ई	ഈ	ஈ	ఈ	ಈ	
U	उ	उ	ഉ	உ	ఉ	ಉ	
UU	ऊ	ऊ	ഊ	ஊ	ఊ	ಊ	
W	—	—	ഁ	—	—	—	(1)
R	ऋ	ऋ	ഋ	—	ఋ	ಋ	(2)
RR	ॠ	—	ൠ	—	ౠ	ೠ	(3)
L	ऌ	—	ഌ	—	ఌ	—	(4)
LL	ॡ	—	ൡ	—	ౡ	—	(5)
E	ए	ए	എ	எ	ఎ	ಎ	
EE	—	—	ഏ	ஏ	ఏ	ಏ	
AI	ऐ	ऐ	ഐ	ஐ	ఐ	ಐ	
O	ओ	ओ	ഒ	ஒ	ఒ	ಒ	
OO	—	—	ഓ	ஓ	ఓ	ಓ	
AU	औ	औ	ഔ	ஔ	ఔ	ಔ	
AW	अं	अं	അം	அம	అం	ಅಂ	(6)
AH	अः	अः	അഃ	அஃ	అః	ಅಃ	(7)
K	क	क	ക	க	క	ಕ	
KH	ख	ख	ഖ	—	ఖ	ಖ	(8)
G	ग	ग	ഗ	—	గ	ಗ	
GH	घ	घ	ഘ	—	ఘ	ಘ	
NG	ङ	ङ	ങ	ங	ఙ	ಙ	(9)
C	च	च	ച	ச	చ	ಚ	
CH	छ	छ	ഛ	—	ఛ	ಛ	
J	ज	ज	ജ	ஜ	జ	ಜ	
JH	झ	झ	ഝ	—	ఝ	ಝ	
NJ	ञ	ञ	ഞ	ஞ	ఞ	ಞ	(10)
TX	ट	ट	ട	ட	ట	ಟ	(11)
TXH	ठ	ठ	ഠ	—	ఠ	ಠ	
DX	ड	ड	ഡ	—	డ	ಡ	
DXH	ढ	ढ	ഢ	—	ఢ	ಢ	
NX	ण	ण	ണ	ண	ణ	ಣ	

Roman spelling	Sanskrit	Hindi	Malayalam	Tamil	Telugu	Kannada	Remarks
T	त	त	ത	த	త	ತ	
TH	थ	थ	ഥ	—	థ	ಥ	
D	द	द	ദ	—	ద	ದ	
DH	ध	ध	ധ	—	ధ	ಧ	
N	न	न	ന	ந	న	ನ	
P	प	प	പ	ப	ప	ಪ	
PH	फ	फ	ഫ	—	ఫ	ಫ	
B	ब	ब	ബ	—	బ	ಬ	
BH	भ	भ	ഭ	—	భ	ಭ	
M	म	म	മ	ம	మ	ಮ	
Y	य	य	യ	ய	య	ಯ	
R	र	र	ര	ர	ర	ರ	
L	ल	ल	ല	ல	ల	ಲ	
V	व	व	വ	வ	వ	ವ	
LZH	—	—	ഴ	ழ	—	—	(12)
LX	ळ	—	ള	ள	ళ	ಳ	(13)
RH	—	—	റ	ற	—	—	(14)
NH	—	—	ണ	ண	—	—	(15)
SH	श	श	ശ	—	శ	ಶ	(16)
SX	ष	ष	ഷ	ஷ	ష	ಷ	(17)
S	स	स	സ	ஸ	స	ಸ	
H	ह	ह	ഹ	ஹ	హ	ಹ	(18)
KSX	क्ष	क्ष	ക്ഷ	க்ஷ	క్ష	ಕ್ಷ	(19)
JNJ	ज्ञ	ज्ञ	ജ്ഞ	—	జ్ఞ+	ಜ್ಞ	(20)
F	—	फ़	—	—	—	—	
Z	—	ज़	—	—	—	—	
RX	—	ड़	—	—	—	—	(21)
RXH	—	ढ़	—	—	—	—	(22)
Q	—	क़	—	—	—	—	(23)
CS	—	—	—	—	ౘ	—	(24)
JZ	—	—	—	—	ౙ	—	(25)

Remarks

(*) This entire Roman Spelling System for Indian languages is based on sound phonetic principles.

Long vowels are represented by writing the symbol for the corresponding short vowel twice, as : *aa*, *ii*, *uu*, etc.

Geminate (doubled) consonants are indicated by writing the symbol for the single consonant twice, as: *kk*, *cc*, *pp*, *mm*, *ll*, *vv*, etc.

Single consonants which are represented by two or more letters (as: *kh*, *ng*, *nj*, *tx*, *txh*, *rh* *nh*, *lx*, etc.), when doubled, are represented by writing the first letter twice, as: *kkh* *nng*, *nnj*, *ttx* *ttxh*, *rrh*, *nnh*, *llx*, etc.

(1) The letter *w* is a vowel (the neutral vowel phonemically contrasting with *u*, *o*, and *a*, as in Malayalam), when it occurs after a consonant. In Malayalam, *vannu*, *vannw* and *vanna* are three different words [(*cf.* (6)].

(2 and 3). The letters *r* and *rr* are vowels, when not preceded or followed by any other vowel, as in Sanskrit: *Krsxnxa*, *dhaatr* or *kartrrn*. (If the letter *r* or *rr* is either preceded or followed by a vowel, it is a consonant, as in Sanskrit : *Raamah*, *Harih*, etc.).

Thus in the Sanskrit words *ka*r*trrn*, *bh*r*aatrbhih* or *vrt*r*uhan*, etc., the non-italicised *r* is a consonant.

(4 and 5) The letters *l* and *ll* under the same conditions as *r* and *rr* are either vowels or consonants.

(6) The letter *w* indicates anusvara (or nasalization of the preceding vowel), when it occurs after a vowel, as in Hindi : *Maiw huuw*.

(7) The letter *h* is the visarga in words of Sanskrit origin when it follows a vowel and occurs before pause or a stop consonant (*cf.* 8, 14, 15, 16 and 18), as in *Punah*, *duhkh*, etc.

(8) The letter *h* stands for aspiration when it follows a stop consonant as in: *kh*, *gh*, *ch*, *jh* *txh*, *dxh*, *th*, *dh*, *ph*, or *bh*, or when it follows a retroflex flap : *rxh* (*cf.* 7, 14, 15, 16 and 18)

(9 and 10). The combination *ng* is a pure velar nasal and the combination *nj* is a pure palatal nasal. In these combinations the *g* and *j* do not stand for the stop consonant or the affricate. They are merely signs of velarity or palatality. If these nasals are followed by the respective stops *k*, *g* and affricates *c*, *j*, they are written as *ngk*, *ngg* and *njc*, *njj* respectively.

(11, 13, 17, 21 and 22). The letter *x* is a mark of retroflection. Thus, *tx*, *txh*, *dx*, *dxh*, *nx* *lx*, *sx*, *rx*, *rxh* are all retroflex consonants, as in Hindi : *betxaa*, *motxhaa*, *barxaa*, *parxhnaa*, and in Marathi : *vedxaa*, *kevdxhaa*, *Punxe*, *halxuu*, and Sanskrit : *visxam*, *paasxaanxam*, etc. (The last letter *h* in all this is aspiration).

(12) The combination *lzh* indicates a voiced alveolar lateralized weak groove spirant, in which *l* indicates lateralization, *z* indicates spirantization and *h* alveolarization. This sound *lzh* occurs in Tamil and Malayalam. The name for Tamil in Tamil is *Tamilzh* (*cf.* 14, 15 and 16).

(14, 15 and 16) The letter *h* when used after non-stop, non-flap consonants indicates an alveolar variant of a dental variety. Thus *s*, *n*, *r*, are dental and *sh*, *nh*, *rh* are alveolar. The *r* vs. *rh* and *n* vs. *nh* opposition is found only in Tamil and Malayalam,

(18) The letter *h* is an independent consonant (glottal fricative), when it is not preceded by a stop or flap consonant or when it is preceded by a vowel and followed by any consonant other than stop consonants, especially in words of Sanskrit origin. In words of non-Sanskrit origin, *h* after a vowel is an independent consonant even when followed by a stop consonant, as in: *behtariin* (*cf.* 7, 8, 14, 15 and 16).

(19) Ksx — k + sx, as in *ksxamaa.*

(20) Jnj — j + nj, as in *jnjaanam,* or Hindi: *vijnjaan* (if the Sanskrit origin of the word is preserved, but *vigyaan,* if the common Hindi pronunciation is to be reflected in writing this word).

(23) The letter *q* denotes the post-velar or phryngeal stop, as in Hindi/Urdu : *sabaq, qilaa*

(24 and 25) The combinations *cs* and *jz* are respectively the voiceless and voiced dental affricates (as opposed to the alveo-palatal affricates, *c* and *j*). These are found in Telugu and Marathi as in Marathi: *csaawglaa, jzavalx.*

II. Sample Texts in Different Languages

(1) *Sanskrit* (*Sawskrtam*):

Laksxmii vasati jihvaagre jihvaagre mitrabaandhavaah,
Bandhanaw caiva jihvaagre jihvaagre maranxaw dhruvam.

Priyavaakya pradaanena sarve tusxyanti jantavah,
Tasmaat tadeva vaktavyaw vacane kaa daridraata.

Hastasya bhuusxanxaw daanaw satyaw kanxtxhasya bhuusxanxam,
Shrotrasya bhuusxanxaw shaastraw bhuusxanxaih kiw prayojanam.

Divaa pashyati noluukah kaako naktaw na pashyati,
Vidyaavihiino muudxhas tu divaa naktaw na pashyati.

(2) *Hindi* (*Hindii*):

Nagar ke duusre kinaare par, Puraanii Dillii mew, Laal Qile ke saamne Caawdnii Cauk naam kaa ek barxaa baazaar hai. Caawdnii Cauk Dillii kaa sab se barxaa baazaar hai. Vahaaw din bhar bhiidx rahtii hai. Saaikilew aur gaarxiyaaw sab samay caltii haiw. Vahaaw bhii aap sab kuch khariid sakte haiw.

(3) *Malayalam* (*Malayaalxam*)

Marrhudraavidxa bhaasxakalx aaya tamilzhw, karnxaatxakam, telunggw enn ivayep poole daksxinxentyay ile bhaasxakalxil onn aanxw malayaalxam. Malayaalxattinnw valxare saamiipyam tamilzh inootx aanxw,

(4) *Tamil* (*Tamilzh*)

Appolzhutu niirnilaiyaic currhi irunta marangkalxukk itxaiyee ulamta cullxikalx 'matxa-matxav' enrhu murhiyum oocai keettxatu.

(5) *Telugu* (*Telugu*)

Naa illu oka mawci illu. Aa iwtxiloo padi gadulu unnavi. Konni gadulu pedda gadulu. Konni gadulu cinna gadulu. Neenu csaduvu gadiloo csaalaa pustakamulu unnavi. Naaku samayamu dorikin appudxella aa gadiki vellxi kuurcsuwdxi csaduvutaanu.

(6) *Kannada* (*Kannadxa*)

Yuddha bhumiyinda tamma tamma manegalixige hindirugida yoodharige, tamage tilxidasxtxu naitika haaguu saundaryada maulyagalxuu apuurnxavaagi kandxuvu. Idxii vishvave ivarannu nirdayigalx endu, daityar endu nindisutt ittu.

III. REMARKS ON THE ROMANIZATION OF INDIAN LANGUAGES

(i) The system of Romanization given here is a spelling system, which since it is linear is convenient for use with the computer for proces ing Indian languages in our research.

(ii) It calls for no diacritical marks or any other special devices that are not found on any standard typewriter keyboard.

(iii) It takes advantage of the letters H, X, W and Z (as well as G and J) to represent, in conjunction with other letters, certain special sounds peculiar to particular languages.

(iv) It is based on sound phonetic principles.

(v) The system is common to all Indian languages and so we could use this system with the existing International machines designed for the Roman script with ease and advantage for typing, punching or printing any of the Indian languages.

(vi) This Roman script is not intended to be adopted for general use. It is meant for the technical handling of these languages, as an extremely handy measure in teleprinting, computerized linguistic processing, indexing in a library, note taking in speech therapy, etc.

IV. Speed in the Production of Technical Reference Material in the Indian Languages by using the Roman Script and the Computer

Work on the compilation of specialised glossaries with minimum human effort and with a modest amount of computer time has been done by us using a linearized Roman spelling.

It is now possible to accomplish more voluminous work in producing technical glossaries using the computer for as many Indian languages at a time as we choose.

(Of course, for this purpose, Indian languages have to be written *necessarily* in the Roman script, as that is the only script the input devices of the computer can take.)

Acknowledgement

The author wishes to thank Mr. V. N. K. Menon, Artist, I.I.Sc., for preparing the Indian Ink drawing of the chart.