World scripts (background paper)

John Clews

The British Library Lending Division, Boston Spa, UK

INTRODUCTION: PRINCIPLES OF SCRIPTS

Computer representation of non-Roman scripts must allow for four different types of writing systems: 'alphabetic', 'diacritical', syllabic and logographic scripts.

<u>Alphabetic scripts</u>

'Alphabetic' scripts give vowels and consonants equal prominence in text, each being allocated one space per letter. In alphabetical order, vowels are usually scattered throughout the alphabet rather than all being together at the beginning. Alphabetic scripts are the only scripts to have upper case and lower case versions of the same character set.

<u>Diacritical scripts</u>

The term 'diacritical scripts' has been coined for scripts where vowels do not occupy an independent character space (except for initial vowels). Vowels following consonants are shown by vowel signs above, below, around, within, or to one side of the consonant. Semitic, South Asian and South East Asian scripts can all be described as diacritical scripts.

<u>Syllabic scripts</u>

In syllabic scripts, consonants and vowels are combined in one symbol. South Asian, South East Asian and Amharic (Ethiopian) scripts are sometimes described as syllabic

scripts, although consonant elements and vowel elements can be easily distinguished, and are reasonably consistent in related sounds. For this reason these have been described as 'diacritical' scripts in this paper.

Korean <u>Hangul</u> characters are syllables, but are composed of easily discernible consonant and vowel elements. They are usually combined with Chinese characters except in North Korea, which has abolished Chinese characters in its script and only uses <u>Hangul</u> characters.

Purely syllabic scripts show no such relationship between related syllables. Japanese <u>Kana</u> is the only purely syllabic script still in current use, but it is nearly always combined with Chinese characters.

<u>Phonetic vs logographic scripts</u>

'Alphabetic', 'diacritical' and syllabic scripts are all phonetic, with no specific meaning attached to each character. Surviving phonetic scripts are all based on the early Phoenician script (Roman, Greek, Cyrillic, Georgian, Armenian, Hebrew and Arabic scripts) or Devanagari script (South Asian and South East Asian scripts).

Most of the earliest scripts were <u>logographic</u>, and expressed concepts rather than sounds, as in Egyptian hieroglyphic script. Some characters gradually lost their original meaning in favour of a standardised phonetic sound, as in Egyptian demotic script. This made the script much more flexible, and later, completely phonetic scripts became dominant. The only surviving logographic script is Chinese script, which is used throughout East Asia, albeit with additional phonetic characters in the Korean and Japanese languages.

<u>Additional letters and diacritics</u>

Scripts which are used for more than one language generally have additional (adapted) letters and a large number of diacritics (signs which can be added to several different letters) in order to represent sounds present only in certain languages. This is particularly true for Roman, Cyrillic and Arabic scripts, and is a major problem in designing character sets for computers.

EUROPEAN SCRIPTS

European scripts are all alphabetical, and have a family resemblance to Greek, which itself was derived from the Phoenician alphabet. The similarity of shapes and alphabetic order of these scripts can be seen in Figure 1.

Roman script

Capital	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Small	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z

Greek script

Capital	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο	Π	Ρ	Σ	Τ	Υ	Φ	Χ	Ψ	Ω
Small	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο	π	ρ	σ,ς	τ	υ	φ	χ	ψ	ω
Phonetic value	a	v	g	d	e	z	i	th	i	k	l	m	n	ks	o	p	r	s,z	t	i	f	x	ps	o

Cyrillic script

Capital	А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
Small	а	б	в	г	д	е	ё	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
Phonetic value	a	b	v	g	d	ye	yo	j	z	i	i	k	l	m	n	o	p	r	s	t	u	f	x	ts	ch	sh	shch	-	ui	-	e	yu	ya

Figure 1.   European scripts

Figure 2.   Georgian and Armenian scripts(l)

Roman script

Historically, 'scripts followed religion' and European scripts are no exception. Roman script spread with Catholic, and later Protestant, Christianity through Western Europe and to the Americas, Africa and Australasia.

During the twentieth century, Roman script replaced Arabic script as a 'modernising' influence in Turkey, Indonesia, Malaysia and Somalia and for a short time in the Central Asian Soviet Socialist Republics. It was even briefly considered as a possible replacement for Cyrillic script in the USSR, although Roman script is now only used for the Estonian, Latvian and Lithuanian languages, which have always used Roman script.

Greek script

Classical Greek script was used in the Eastern Orthodox Church, and is still used for modern Greek. Diacritics are particularly complex, as two (and iota subscript) may be combined. However, there have been recent official moves by Greece to simplify the use of diacritics.

Cyrillic script

Orthodox Christianity spread widely through Eastern Europe, but Greek script proved inadequate for the Slavonic languages of Eastern Europe. Cyrillic script added new characters to Greek to form a new script which was used throughout Eastern Europe.

In this century, Cyrillic script has been used with additional letters to replace Arabic script for the Central Asian languages of the USSR, and also for Mongolian in Outer Mongolia.

OTHER ALPHABETIC SCRIPTS

Georgian and Armenian also derive from Greek script, although the family resemblance in terms of shape and word order is less clear than that of Roman, Greek and Cyrillic (see Figure 2). They are little used outside the Soviet Republics of Georgia and Armenia.

The cursive Mongolian script is alphabetic, but resembles neighbouring East Asian scripts in being written from top to bottom. It also resembles neighbouring Semitic scripts in having different initial, medial, and final forms. Hebrew script is essentially a 'diacritical' script but for Yiddish some consonants have been used alphabetically to express vowels. In the Jewish Autonomous Region of the USSR, Yiddish now

has a completely alphabetical Hebrew script. Similarly, Arabic has been made alphabetic for the Uighur and Kazakh languages in the Sinkiang Uighur Autonomous Region of China, by modifying consonants to represent vowels.

Hebrew script

Hebrew script has 22 basic consonants, derived, like Arabic and Greek, from the 22-letter Phoenician script (see Figure 3). When they occur at the end of a word, 5 of these consonants have a different shape. Vowels are not usually represented in Hebrew. When they are required, for the sake of clarity, it is by points above, below, or within the character. Where the text lacks vowels, the reader mentally supplies the vowels from past experience, according to the context. The reader is helped in this by the existence of Matres Lectionis (semivowel consonants) such as w or y, which are often associated with particular vowels, such as u or i.

FORM	PHONETIC VALUE		FORM			PHONETIC VALUE	
א	a, (')		כ כ		ך	kh	k
ב ב	bh	b	ל			l	
ג ג	gh	g	מ	ם		m	
ד ד	dh	d	נ		ן	n	
ה	h		ס			s	
ו	w		ע			('), e	
ז	z		פ פ		ף	ph	p
ח	x		צ		ץ	s	
ט	t		ק			q	
י	y		ר			r	
			שׁ שׂ			s	sh
			ת ת			th	t

SHORT VOWELS							
	-a	-a	-e	-e	-i	-o	-u

LONG VOWELS							
	-a, -o	-ē	-ē	-ī	-ō	ō	ū

Figure 3.   Hebrew scripts(l)

Arabic script

Arabic script added a further 6 letters to the 22 Phoenician letters, so that Arabic script has 28 consonants (see Figure 4). Other languages using Arabic script have added extra consonants to the original 28. Persian has 32, Urdu has 36 and Sindhi has 52 consonants. There are 10 extra consonants in both Pushto and Uighur, used respectively in Afghanistan and China.

Arabic script is cursive, and consonants have different forms, depending on their position in the word. Vowels are usually omitted, but where they are included they are shown by subscript or superscript dashes.

Arabic consonants

Similar consonants are distinguished by adding 1, 2 or 3 dots to the 15 basic shapes. Numbers indicate alphabetical order and the next three columns show final, medial and initial forms of the basic letter shapes.

Figure 4.   Arabic script (2)

Maldivian consonants

∕	*h*
⅄	*sh*
⅄	*n*
⅛	*r*
⅀	*b*
⅄	*l*
⅃	*k*
⅂	*a, [']*
⅄	*w, v*
⅄	*m*
⅄	*f, ph*
⅄	*d*
⅄	*th*
⅄	*l*
⅄	*g*
⅄	*ny*
⅄	*s*
⅄	*d*
⅄	*z*
⅄	*t*
⅄	*y*
⅄	*p*
⅄	*j*
⅄	*ch*

Maldivian vowels

○	○	○	○	○
-a	*-i*	*-u*	*-e*	*-o*
○	○	○	○	○
-ā	*-ī*	*-ū*	*-ē*	*-ō*

Figure 5.   Maldivian script

Maldivian script

Maldivian script derives from Arabic script, although the letters are very different in form and arrangement (see Figure 5). The first nine letters of Maldivian script resemble Arabic (Persian) numerals. Maldivian vowels are mandatory, and extend the basic 3 vowel signs of Arabic to 10.

SOUTH ASIAN SCRIPTS

Indian scripts also represent vowels by vowel signs attached to consonants, but here their use is mandatory. Where no diacritic is shown, the vowel defaults to an inherent 'a'. There are additional diacritics for aspiration, nasalisation and negation of vowels (not shown here). The similarity between these scripts reflects the influence of Hinduism.

The most widespread script in India is Devanagari, which in modern Hindi has 33 consonants and 11 vowels (see Figure 6). Ten of these consonants are aspirated versions of other consonants. Initial vowels occupy one character space; subsequent vowels are shown as vowel signs above consonants' diacritical signs. Certain conjunct consonants are also formed by adding part of one letter to part of another.

Devanagari is used for Hindi and several other Indian languages. Other North Indian scripts (Gujerati, Gurmukhi [Punjabi], Oriya and Bengali) and the more rounded South Indian (Dravidian) scripts (Kannada, Malayalam, Tamil and Telegu) follow the same basic pattern, even when the character shapes appear totally unrelated (see Figure 7).

Sinhalese script is similar to South Indian scripts, although the language is more closely related to North Indian languages. Tibetan script resembles North Indian scripts, although Tibetan spelling appears more idiosyncratic, by comparison, partly to distinguish homophones in the monosyllabic Tibetan language. Examples of North Indian and South Indian scripts are given in Figures 6 and 7.

Vowels

अ (अ) आ इ ई उ ऊ
a   ā   i ī u ū

ऋ ए ऐ ओ औ
ri ē ai ō au

Consonants

GUTTURALS	क	ख	ग	घ	ङ
	ka	kha	ga	gha	nga
PALATALS	च	छ	ज	झ (झ)	ञ
	cha	chha	ja	jha	nya
CEREBRALS	ट	ठ	ड	ढ	ण (ण)
	ṭa	ṭha	ḍa	ḍha	ṇa
DENTALS	त	थ	द	ध	न
	ta	tha	da	dha	na
LABIALS	प	फ	(फ)	ब	भ	म
	pa	pha	fa	ba	bha	ma

SEMI-VOWELS	य	र	ल	व
	ya	ra	la	va, wa
SIBILANTS	श	ष	स	
	sa, sha	sha	sa	
ASPIRATE	ह			
	ha			

Vowel signs

○ -a   ○ा -ā   ○ि -i   ○ी -ī   ○ु -u   ○ू -ū

○ृ -ri   ○े -ē   ○ै -ai   ○ो -ō   ○ौ -au

Thus, कि *ki*, के *kē*.

Figure 6.   Devanagari script (l)

Vowels

അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ
a	*ā*	*i*	*ī*	*u*	*ū*	*ru*

എ	ഏ	ഐ	ഒ	ഓ	ഔ
e	*ē*	*ai*	*o*	*ō*	*au*

Consonants

ക	ഖ	ഗ	ഘ	ങ	ച	ഛ
ka	*kha*	*ga*	*gha*	*nga*	*cha*	*chha*

ജ	ഝ	ഞ	ട	ഠ	ഡ	ഢ
a	*jha*	*nya*	*ṭa*	*ṭha*	*ḍa*	*ḍha*

ണ	ത	ഥ	ദ	ധ	ന	പ
ṇa	*ta*	*tha*	*da*	*dha*	*na*	*pa*

ഫ	ബ	ഭ	മ	യ	ര	ല	വ
pha	*ba*	*bha*	*ma*	*ya*	*ra*	*la*	*va*

ശ	ഷ	സ	ഹ	ള	ഴ	റ
sa	*sha*	*sa*	*ha*	*la*	*ra*	*ra*

Vowel signs

◌ാ -*ā*	◌ി -*i*	◌ീ -*ī*	◌ു -*u*	◌ൂ -*ū*
◌ൃ -*ru*	െ◌ -*e*	േ◌ -*ē*	ൈ◌ -*ai*	
ൊ◌ാ -*o*	ോ◌ാ -*ō*	ൌ◌ൗ -*au*		

Figure 7.   Malayan script (l)

SOUTH EAST ASIAN SCRIPTS

Burmese, Khmer, Thai and Lao scripts derive from Indian scripts, but reflect the spread of Buddhism in Asia. Compared to Indian scripts, they have additional consonants, vowels and diphthongs, and pronunciation differs considerably from the original Indian models. Most South East Asian languages are tonal, so additional signs are used for tones, as well as for aspiration, nasalisation and negation of vowels (not shown here).

Burmese is closest to Indian scripts, with few extra consonants, vowels, compound vowels or tone signs (see Figure 8).

Figure 8.   Burmese script (l)

Khmer and Lao scripts have a similar consonant system to Burmese, although they share with Thai script a particularly complex vowel system, with about a dozen vowel elements which may be combined to represent two dozen or more vowels and diphthongs. Thai script also has a more complex system of consonants than Khmer or Lao (see Figure 9).

Consonants

ก	ข	ฃ	ค	ฅ	ฆ	ง	
ko	*kho*	*kho*	*kho*	*kho*	*kho*	*ngo*	

จ	ฉ	ช	ซ	ฌ	ญ		
cho	*chho*	*chho*	*so*	*chho*	*yo*		

ฎ	ฏ	ฐ	ฑ	ฒ	ณ		
do	*to*	*tho*	*tho*	*tho*	*no*		

ด	ต	ถ	ท	ธ	น		
do	*to*	*tho*	*tho*	*tho*	*no*		

บ	ป	ผ	ฝ	พ	ฟ	ภ	ม
bo	*po*	*pho*	*fo*	*pho*	*fo*	*pho*	*mo*

ย	ร	ล	ว	ศ	ษ	ส	
yo	*ro*	*lo*	*wo*	*so*	*so*	*so*	

ห	ฬ	อ	ฮ				
ho	*lo*	*'o*	*ho*				

Vowel signs

Figure 9.   Thai script (l)

EAST ASIAN SCRIPTS

East Asian languages (Chinese, Japanese and Korean) have a large number of homonyms, where similar sounding words have a wide variety of meanings, so a completely phonetic script could be very ambiguous. Because of this, all three languages retain the use of ancient Chinese characters (called respectively <u>Hanzi</u> and <u>Kanji</u> in China and Japan) in order to avoid ambiguous meanings. One <u>character</u> represents one word, and the <u>meaning</u> is usually similar in Chinese, Japanese and Korean, although pronunciation differs considerably.

Chinese script contains well over 40,000 characters, each of which has a unique meaning and is composed of from 1 to around 40 strokes. The larger ones can be decomposed into subordinate elements, most of which have conceptual rather than purely phonetic connotations. Because the large number of homophones in Chinese, Japanese and Korean requires an effective means of distinguishing different words which sound the same, romanisation schemes such as Chinese <u>pinyin</u> and Japanese <u>Kunrei</u> are unlikely to replace Chinese characters in East Asia, although they form a useful supplement. The shapes of characters have been simplified in Japan and, in a more radical form, in the People's Republic of China (shown as JAP. and SIMP, respectively in Figure 10).

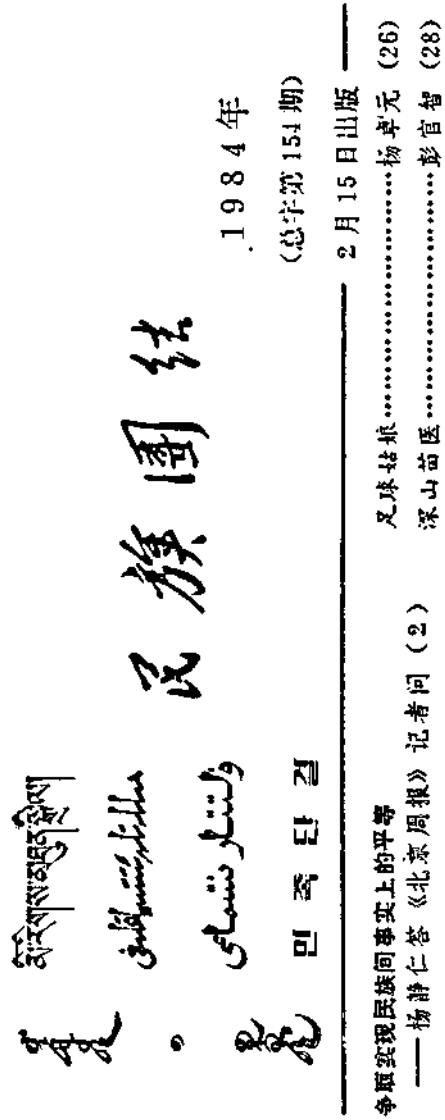| FORMS			SOUNDS			DEF.
ORIG.	JAP.	SIMP.	CHINESE	JAPANESE	KOREAN	
廣	広	广	*kuang*	*ko* (*hiro-i*)	*kwang*	wide
與	与	与	*yu*	*yo* (*ata-eru*)	*yo*	give
習	習	习	*shi*	*shu* (*nara-u*)	*sup*	learn
鬪	闘	斗	*tou*	*to* (*tataka-u*)	*thu*	fight
專	專	专	*chuan*	*sen* (*moppa-ra*)	*chon*	solely
勸	勧	劝	*chhuan*	*kan* (*susu-meru*)	*kwon*	advise
讀	読	读	*tou*	*doku* (*yo-mu*)	*tok*	read

Figure 10.   Chinese characters(1)

Figure 11. Title page of book, showing Mongolian, Tibetan, Kazakh, Uighur and Korean versions of the title written in cursive Chinese script (3)

Chinese script

Written Chinese includes only Chinese characters, as flexional word endings do not occur in the Chinese language. Pinyin romanised forms are sometimes used to supplement the characters, e.g. in language teaching, and in book titles. Sometimes minority language equivalents are also given. Figure 11 shows Mongolian, Tibetan, Kazakh, Uighur and Korean versions of the title written in cursive Chinese script.

Korean script

Korean uses Hangul characters, which are completely phonetic characters, being built up from several alphabetic elements. These are usually mixed in text with Chinese characters, except in North Korea where Chinese characters are no longer used. Hangul characters are built up as illustrated in Figure 12.

consonants   vowels

ㄱ   ㄴ   ㄷ   ㅏ   ㅑ   ㅓ   ㅕ
k   n   t   *a   ya   o   yo*

consonant + vowel

가   갸   거   겨
ka   kya   ko   kyo

consonant + vowel + consonant

각   간   갈
kak   kan   kat

Figure 12.   Korean script (l)

	a	i	u	e	o
	あ ア a	い イ i	う ウ u	え エ e	お オ o
k	か カ ka	き キ ki	く ク ku	け ケ ke	こ コ ko
s	さ サ sa	し シ shi/si	す ス su	せ セ se	そ ソ so
t	た タ ta	ち チ chi/ti	つ ツ tsu/tu	て テ te	と ト to
n	な ナ na	に ニ ni	ぬ ヌ nu	ね ネ ne	の ノ no
h	は ハ ha(wa)	ひ ヒ hi	ふ フ fu/hu	へ ヘ he(e)	ほ ホ ho
m	ま マ ma	み ミ mi	む ム mu	め メ me	も モ mo
y	や ヤ ya	—	ゆ ユ yu	—	よ ヨ yo
r	ら ラ ra	り リ ri	る ル ru	れ レ re	ろ ロ ro
w	わ ワ wa	—	—	—	を ヲ o
					ん ン n
g	が ガ ga	ぎ ギ gi	ぐ グ gu	げ ゲ ge	ご ゴ go
z	ざ ザ za	じ ジ ji/zi	ず ズ zu	ぜ ゼ ze	ぞ ゾ zo
d	だ ダ da	ぢ ヂ ji/zi	づ ヅ zu	で デ de	ど ド do
b	ば バ ba	び ビ bi	ぶ ブ bu	べ ベ be	ぼ ボ bo
p	ぱ パ pa	ぴ ピ pi	ぷ プ pu	ぺ ペ pe	ぽ ポ po

Figure 13.   Japanese script (5)

<u>Japanese script</u>

Japanese writing uses Chinese characters combined with two types of <u>Kana</u>. The Chinese characters have only one form, but two possible pronunciations or readings: <u>On</u> readings (Chinese derived pronunciation) and <u>Kun</u> readings (Japanese derived pronunciation). Both are illustrated in Figure 10 with <u>Kun</u> readings in brackets. Japanese text including Kanji, Katakana and Hiragana is illustrated below. (4)

パーソナルコンピュータを用いた虚血性
心疾患手術症例の登録システム及びその
臨床研究への応用

```
Patient Registry System for the Coronary Artery Surgery
       with the Aid of Personal Computer and It's
       Application for the Clinical Research
```

The 46 <u>Kana</u> symbols appear in two forms: <u>katakana</u>, frequently used to indicate borrowed words, and the more cursive <u>hiragana</u>, which are those <u>kana</u> usually used to indicate flexional endings in Japanese words (see Figure 13).

SOURCES OF ILLUSTRATIONS

1. NAKANISHI, Akira   <u>Writing systems of the world:
   alphabets, syllabaries, pictograms.</u>
   Rutland, Vt: Tuttle, 1980.

2. TRACY, Walter.   The flourishing reed: Arabic script.
   In <u>Alphabet; international annual of letter forms</u>,
   HUTCHINGS, R.S.   (ed.) James Moran Ltd, 1964.
   pp. 139-146

3. <u>Minzu Tuanjie.</u>   Published in China, 1984.

4. <u>Juntendo Medical Journal</u>, 29 (3), p. 374-380.   Published
   by Juntendo Medical Society, 2-1-1, Mongo, Tokyo 113
   Japan.

5. HADAMITZKY, Wolfgang, and SPAHN, Mark.   <u>Kanji and
   Kana; a handbook and dictionary of the Japanese writing
   system</u>.   Rutland, Vt: Tuttle 1981.

AUTHOR

John Clews,   The British Library Lending Division,
Boston Spa, Wetherby, West Yorkshire LS23 7BQ