Chapter 1

Scripts and Languages

Writing is speech put in visible form.
— Michael D. Coe

Writing Systems,[1] Scripts, and Languages

Michael Coe began his book on the decipherment of the Maya script with the memorable sentence above.[2] Over the centuries, a hundred or so different scripts have been developed to record human languages. Many of them, such as Egyptian hieroglyphs or the Cuneiform script went out of existence; others, such as Chinese characters, are still in use and have gone through an amazingly small amount of change during their evolution. Where, when, how, and why writing originated, we shall not decide in our context. However, it looks like writing developed independently in a number of places. The art of writing was used for recording many different languages.

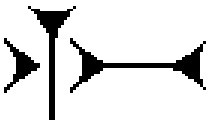
About 2000 years passed between the earliest writing systems and the remarkable system that the ancient Greeks developed based on the Phoenician invention of the alphabet. The alphabetic principle essentially reflects the basic insight that each word of the spoken language consists of a limited group of individual sounds (phonemes that can be represented by a limited group of individual letters (graphemes).[3] This apparently simple-sounding principle was revolutionary when it arose, because with it came the possibility of recording each spoken word of each individual language in writing. The Latin script evolved from the Western variety of the Greek alphabet and was originally the writing system of the ancient Roman Empire. Currently, it is one of the most widely used scripts in the world.

In another part of the world, the Japanese adapted Chinese characters to their own language. Modern Japanese is written with a complex mixture of Chinese characters (kanji) and the kana syllabaries (hiragana and katakana). Occasionally it also uses the Latin alphabet (rōmaji) for writing foreign words or abbreviations. In fact, three different scripts coexist in Japan. Theoretically, Japanese could be written all phonetically by using kana alone. However, the Japanese writing system continues to be of *multiscriptal nature*. Probably the most important reason is the great number of homophones that the Japanese language uses. Homophones are words that sound alike but have a different meaning—such as *to*, *too*, or *two* in English, or *viel* (much) and *fiel* (fell) in German. All languages have homophones, but Japanese (and Chinese) is particularly rich in these. If all kanji were to be replaced by kana spellings, there might be many ambiguities that could be a serious handicap to communication.

Experts estimate that there are between 4,000 and 5,000 languages spoken in the world today.[4] Not all of them are recorded in writing, and many more languages might have disappeared without leaving a trace. Compared to the huge number of languages, the number of scripts in use is relatively small. This is because scripts have been frequently borrowed by languages throughout history. In any case, essential to the development of full writing was the discovery of the *rebus principle*.[5] In a nutshell, this means representing a word by means of another word that is phonetically the same (a homophone) or at least similar. For example, in English, you can use the combination of

a picture of a bee and the picture of a leaf to write the word *belief*.[6] The English language with its complicated syllable structure is not well suited for being written like this. However, the rebus principle was used in all ancient scripts because it allowed depicting abstract concepts with concrete objects. For example, in early Sumerian writing the same symbol was used to write the words *ti* (arrow) and *ti(l)* (life), as illustrated in Figure 1.[7] Similar examples can be found in other ancient scripts such as Egyptian hieroglyphs or Maya writing, and so on.[8]

Figure 1: Cuneiform Sign "arrow," "life" (Sumerian TI); U+122FE



Eventually, all scripts operate on the same principle; that is, they use symbols to represent sounds, and each type of script entails about the same amount of effort to record the same amount of information. Nevertheless, there are considerable differences, which are explained below.

Categories of Writing Systems

Writing systems fall roughly into one of the following categories: alphabets, syllabaries, and logographic (or ideographic) writing systems, which are explained below.

Alphabets

An alphabet consists of a set of characters that represent the phonemes of a language in writing. A true alphabet contains separate letters for both consonants and vowels. Examples are the Latin and the Cyrillic alphabets. An *abjad* is a consonantal alphabet in which each character usually stands for a consonant; the reader must supply the appropriate vowel. Examples are the Arabic and the Hebrew alphabets. Alphabets usually consist of about 20 to 40 different characters.

Syllabaries

A syllabary (katakana and hiragana, for example) is a writing system in which the characters stand for entire syllables. Syllabaries usually have between 50 and 200 different characters. An *abugida* is a kind of intermediate between syllabary and alphabet; this means that consonant signs are inherently associated with a

following vowel. Examples are the Indic scripts (Bengali, Devanagari, and so on), and Ethiopic.

Logographic (Ideographic) Writing

In a logographic (ideographic) writing system, a character (or logogram) represents a word or a morpheme (the smallest meaningful unit of language). Logograms are often called *ideograms*. Strictly speaking, however, ideograms represent ideas directly rather than words and morphemes, and none of the writing systems that we deal with are truly ideographic.[9] Logographic scripts usually contain many thousands of characters.

The International Organization for Standardization (ISO) has developed more than 18,000 international standards on a variety of subjects; one of them is dedicated to codes for the representation of names of scripts. The ISO 15924 standard defines two sets of codes (a four-letter code and a numeric code) for writing systems (scripts). The table below lists a number of scripts, their ISO 15924 codes, the types of script, and the main languages that they are used with.

Table 1: Scripts and Language

Name of Script	ISO 15924 Code	Type of Script	Languages (Main)
Arabic	Arab 160	alphabet (abjad)	Arabic, Persian (Farsi), Kurdish, Urdu
Bengali	Beng 325	abugida	Bengali
Cyrillic	Cyrl 220	alphabet	Belarusian, Bulgarian, Kazakh, Macedonian, Mongolian, Russian, Serbian, Ukrainian, Tatar
Devanagari	Deva 315	abugida	Hindi, Konkani, Marathi, Nepali
Greek	Grek 200	alphabet	Greek
Hanzi (Simplified variant)	Hans 501	logographic	Chinese (China, Singapore, Malaysia)
Hanzi (Traditional variant)	Hant 502	logographic	Chinese (Taiwan, Hong Kong, Macau)
Hangul	Hang 286	alphabet	Korean
Hanja	Hani 500	logographic	Korean

(*continued*)

Table 1: (*continued*)

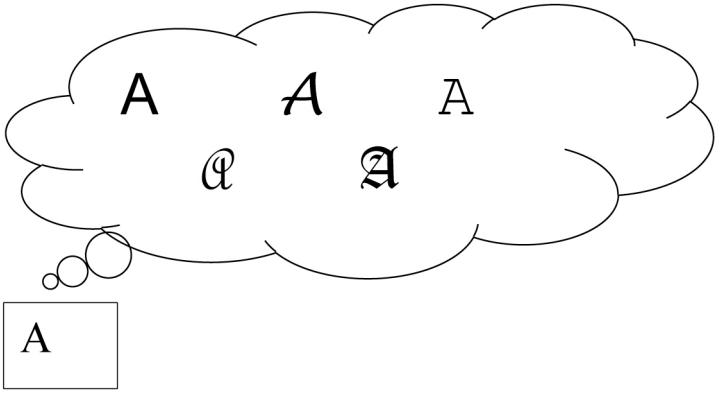
Name of Script	ISO 15924 Code	Type of Script	Languages (Main)
Hebrew	Hebr 125	alphabet (abjad)	Hebrew, Yiddish
Hiragana	Hira 410	syllabary	Japanese
Japanese (alias for Kanji + Hiragana + Katakana)	Jpan 413	mixed	Japanese
Kanji	Hani 500	logographic	Japanese
Katakana	Kana 411	syllabary	Japanese
Korean (alias for Hangul + Hanja)	Kore 287	mixed	Korean
Latin	Latn 215	alphabet	Afrikaans, Albanian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Hungarian, Icelandic, Indonesian, Italian, Latvian, Lithuanian, Malay, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swahili, Swedish, Tagalog, Turkish, Vietnamese, Welsh, Wolof, Xhosa, Yoruba, Zulu
Tamil	Taml 346	abugida	Tamil
Telugu	Telu 340	abugida	Telugu
Thai	Thai 352	abugida	Thai

Characters and Glyphs

What is a character? A character can be defined pragmatically as the smallest unit of a script conveying a meaning. It is an abstract concept that does not imply a specific shape or concrete visual representation. The Unicode glossary definition of *character* "refers to the abstract meaning and/or shape, rather than a specific shape …, though in code tables some form of visual representation is essential for the reader's understanding."[10]

A *glyph* is the concrete graphical representation of an abstract character. The Latin capital letter *A*, for example, can be displayed or printed using any of a number of glyphs or fonts. A *font* is a collection of glyphs—a means of visually representing writing. The font type determines how the characters actually look on a computer screen or when printed on paper. Hence, an uppercase Latin A will look very different depending on the font in use. Figure 2 shows how the character looks when rendered using Arial, Lucida Handwriting, Courier New, French Script, and Old English Text fonts.

Figure 2: Different Glyphs of the Letter A



Each character looks different; nevertheless, it is the same character. If you understand that a character can come in many different shapes, you can identify the abstract character in different handwriting and fonts. Of course, this depends on an individual's familiarity with the particular culture and language. For the untrained eye, or before you have learned the particular language and script, all those letters and characters may mean nothing and look almost the same to you.

Arabic script does not distinguish between uppercase and lowercase characters. However, all characters have a different shape, depending on whether they are at the beginning,

middle, or end of a word. So they can appear in four distinct forms (initial, medial, final, or isolated). Thus, for example, the Arabic letter *hā* can have four different shapes. See Table 2.

Table 2: Different Forms of the Arabic Letter hā

Position	Isolated	Initial	Medial	Final
Shape	ه	هـ	ـهـ	ـه

Let us regard another aspect. The Latin capital letter A, the Greek capital letter Alpha, or the Latin capital letter B and the Greek capital letter Beta might look the same—which is no surprise considering their origin, and they might be represented with identical glyphs. Nevertheless, they are different characters. Note that the similarities between the lowercase letters are not that great. These characters are part of different character repertoires.[11]

The two Japanese kana syllabaries consist of 48 syllable signs each;[12] ka, ki, ku, ke, ko, for example. Katakana and hiragana both render the same syllables. Katakana characters are angular and mostly used to spell foreign borrowings, whereas hiragana have a more rounded and flowing shape and are used to write native words for which there are no kanji. See Table 3.

Table 3: Examples of Hiragana and Katakana Characters

Rōmaji	Hiragana Character	Katakana Character
ka	か	カ
ki	き	キ
ku	く	ク
ke	け	ケ
ko	こ	コ

Some Chinese characters reveal their origin as pictographs (pictures of what they represent), but over the years, they have become more and more abstract. A good example is 人, the character for *man* or *person* (rén in *pinyin*, the official system for transliterating Chinese characters into the Latin alphabet). It "basically shows a straight back above two strong legs firmly planted on the ground."[13] Other examples are 木 (mù in pinyin), which means *tree*, or 日 (rì in pinyin), which means *sun*. In these cases, a character consists of one component only. In most other cases, a character consists of two (or more) components.

The traditional classification of Chinese characters distinguishes six classes: pictographic characters, simple ideographic characters, compound ideographic characters, phonetic loans or rebus characters, semantic-phonetic compound characters, and mutually

interpretative characters.[14] The vast majority of Chinese characters (about 90% or more) are semantic-phonetic compounds. That is, they consist of a semantic component which stands for the meaning and a phonetic component which denotes the pronunciation. For example, the character for *mother* (媽 mā in pinyin) consists of the component *female* (女) and the component *horse* (馬 mǎ in pinyin). The first component carries the semantic meaning (female), whereas the second indicates the sound by referring to the word for *horse*, which is pronounced similarly. Despite being called *compounds*, these characters are single entities, and they are written so that they take up the same amount of space as any other character.

At this point, we need to re-emphasize that there are two varieties of the Chinese script. Traditional characters are the form of Chinese writing as it was used for many hundreds of years. Simplified characters were introduced in the 1950s in an effort to make them easier to learn and to increase literacy. Since 1956, about 2,300 hanzi have been simplified, and more than 1,000 variants were ruled out.[15] The above-mentioned character that means *horse* (馬) is in traditional form; its simplified equivalent is 马. The traditional characters are used in Taiwan and Hong Kong, Macau, and in older literature; simplified characters are used in the People's Republic of China, Singapore, and Malaysia.

Korean is mainly written in hangul (or hangeul, pronounced han'gŭl), which is an alphabet that consists of 14 consonants and 10 vowels; but also mixes in Chinese characters (hanja). Jamo, the letters of hangul, are built into syllabic blocks that can be written horizontally from left to right as well as vertically from top to bottom in columns from right to left. For example, the word hangul consists of the following jamo:

ㅎ	ㅏ	ㄴ	ㄱ	ㅡ	ㄹ
H	a	n	g	eu	l

It is written like this: 한글.

As we have seen, the Chinese script has been borrowed by and adapted for other languages, most significantly Japanese (kanji), Korean (hanja), and (historically) Vietnamese (hán tư). This is why many characters look the same or similar in these languages, but they may have slightly different shapes. Currently, there are four traditions for character shapes in East Asia: Simplified Chinese, Traditional Chinese, Japanese, and Korean. Rendered with a single font for all four locales, the characters are legible, but some may look odd to speakers of a particular language. This can be illustrated by the example in Table 4, which uses a different font for each locale.

Table 4: Examples of Different Forms of Han Character

Code Number in Unicode	Simplified Chinese	Traditional Chinese	Japanese	Korean
U+6B21	次	次	次	次

To summarize: In any given language there is a set of definable characters. This set is called an *Abstract Character Repertoire*. In other words, a *character repertoire* is a set of abstract characters that are needed for writing a language or a group of languages without assuming a specific internal representation in computers.

The character repertoire of English includes the (uppercase and lowercase) letters A–Z, the numbers 0–9, as well as special characters such as punctuation marks, and so on:

ABCDEFGHIHKLMNOPQURSTUVWXYZ
abcdefghijklomopqurstuvwxyz0123456789-=!
@#$%^&*()_+{}[];':"<>?,./

Actually, this is not quite true since "there are many Englishes, not just one."[16] So the character repertoire shown previously would, strictly speaking, be true only for American English. British English uses at least one more special character: the pound sign (£).

The French character repertoire also has a number of accented characters:

ÂÀÆÇÉÊÈËÎÏÔŒÛÙÜŸâàæçéêèëîïôœûùüÿ

French also has special characters such as single and double quotation marks (‹,›, «,»), the section sign (§), and the Euro symbol (€).

What about the character repertoire of Chinese? No one can give an exact figure since the repertoire of Chinese characters is theoretically an open set.[17] Estimates go from 40,000[18] to more than 100,000[19]. In any case, a 3,000-character vocabulary seems to be sufficient for day-to-day communication.

For systematic purposes, we can go a step further and group languages according to (mainly) geographical principles.

The *Latin 1 character repertoire*, for example, contains accented characters and other letters needed for writing the languages of Western Europe; the *Latin 2 character repertoire* has the letters needed for writing the languages of Central and Eastern Europe, and so on. These language (or rather script) groups are explained in more detail in the following chapters.

Chapter Summary

Compared to the huge number of languages, the number of scripts in use is relatively small. Writing systems all operate on the same principle. That is, they use symbols to represent sounds, and each type of system entails about the same amount of effort to record the same amount of information.

Writing systems fall roughly into one of the following categories: alphabets, syllabaries, and logographic (or ideographic) systems.

A character can be defined as the smallest unit of a script that conveys meaning. A glyph is the concrete graphical representation of an abstract character. A font is a collection of glyphs.

In any given language, there is a set of definable characters that can occur. This is called an *abstract character repertoire*.

---

[1] The terms 'writing system' and 'script' are often used interchangeably; both are conceptually independent of each other, however. A writing system needs a script for its physical representation. Generally, all important technical terms are explained in the glossary at the back of the book.

[2] Coe, 1993, p. 13. For more complete discussions of the history of writing and language, see also Comrie (1990), Cook et al. (1998), Coulmas (1996), Daniels et al. (1996), Robinson (2001), and Ruhlen (1994).

[3] Wolf, 2006, p. 7.

[4] Comrie, 1990, p. 2; Ruhlen (1994), p. vii.

[5] Robinson, 2001, p. 12; Coulmas, 1996, p. 433 f.

[6] Robinson, 2001, ibid.

[7] TI (cuneiform). 2010, September 21. Wikipedia. Retrieved 11:47, October 26, 2011, from http://en.wikipedia.org/w/index.php?title=TI_(cuneiform)&oldid=386098085.

[8] Cook et al., 1998, p. 103; Coe, 1993, p. 118; Coe, 2001, p. 219; Robinson, 2001, p. 42.

[9] See Coe, 2001, p. 218; Coulmas, 1996, p. 225.

[10] See the "Glossary of Unicode Terms" at http://unicode.org/glossary/ .

[11] There are a number of reasons why the Latin, Greek, and Cyrillic scripts have been separately encoded, rather than being encoded as a single script. See "Unicode Technical Note #26" at http://unicode.org/notes/tn26/.

[12] Coulmas, 1991, p. 129.

[13] Fazzioli, 2005, front flap.

[14] Cf. Coulmas, 1991, p. 98 ff.

[15] Coulmas, 1991, p. 245.

[16] Andy Kirkpatrick and David Deterding, World Englishes, in: James Simpson (ed.), *The Routledge Handbook of Applied Linguistics*, Abingdon: Routledge, p. 373.

[17] Chinese characters. (2011, October 24). In Wikipedia, The Free Encyclopedia. Retrieved 11:49, October 26, 2011, from http://en.wikipedia.org/w/index.php?title=Chinese_characters&oldid=378387214.

[18] Fazzioli, 2005, p. 14.

[19] Gillam, 2003, p. 45.