The Unicode Standard
Version 8.0 – Core Specification

To learn about the latest version of the Unicode Standard, see http://www.unicode.org/versions/latest/.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

Chapter 18

East Asia

This chapter presents modern-day scripts used in East Asia. This includes major writing systems associated with Chinese, Japanese, and Korean. It also includes several scripts for minority languages spoken in southern China. The scripts discussed are as follows:

Han	*Katakana*	*Lisu*
Bopomofo	*Hangul*	*Miao*
Hiragana	*Yi*	

The characters that are now called East Asian ideographs, and known as Han ideographs in the Unicode Standard, were developed in China in the second millennium BCE. The basic system of writing Chinese using ideographs has not changed since that time, although the set of ideographs used, their specific shapes, and the technologies involved have developed over the centuries. The encoding of Chinese ideographs in the Unicode Standard is described in *Section 18.1, Han*. For more on usage of the term *ideograph*, see "Logosyllabaries" in *Section 6.1, Writing Systems*.

As civilizations developed surrounding China, they frequently adapted China's ideographs for writing their own languages. Japan, Korea, and Vietnam all borrowed and modified Chinese ideographs for their own languages. Chinese is an isolating language, monosyllabic and noninflecting, and ideographic writing suits it well. As Han ideographs were adopted for unrelated languages, however, extensive modifications were required.

Chinese ideographs were originally used to write Japanese, for which they are, in fact, ill suited. As an adaptation, the Japanese developed two syllabaries, *Hiragana* and *Katakana*, whose shapes are simplified or stylized versions of certain ideographs. (See *Section 18.4, Hiragana and Katakana*.) Chinese ideographs are called *kanji* in Japanese and are still used, in combination with *Hiragana* and *Katakana*, in modern Japanese.

In Korea, Chinese ideographs were originally used to write Korean, for which they are also ill suited. The Koreans developed an alphabetic system, *Hangul*, discussed in *Section 18.6, Hangul*. The shapes of Hangul syllables or the letter-like *jamos* from which they are composed are not directly influenced by Chinese ideographs. However, the individual jamos are grouped into syllabic blocks that resemble ideographs both visually and in the relationship they have to the spoken language (one syllable per block). Chinese ideographs are called *hanja* in Korean and are still used together with Hangul in South Korea for modern Korean. The Unicode Standard includes a complete set of Korean Hangul syllables as well as the individual jamos, which can also be used to write Korean. *Section 3.12, Conjoining Jamo Behavior*, describes how to use the conjoining jamos and how to convert between the two methods for representing Korean.

In Vietnam, a set of native ideographs was created for Vietnamese based on the same principles used to create new ideographs for Chinese. These Vietnamese ideographs were used through the beginning of the 20th century and are occasionally used in more recent signage and other limited contexts.

Yi was originally written using a set of ideographs invented in imitation of the Chinese. Modern Yi as encoded in the Unicode Standard is a syllabary derived from these ideographs and is discussed in *Section 18.7, Yi*.

Bopomofo, discussed in *Section 18.3, Bopomofo*, is another recently invented syllabic system, used to represent Chinese phonetics.

In all these East Asian scripts, the characters (Chinese ideographs, Japanese *kana*, Korean Hangul syllables, and Yi syllables) are written within uniformly sized rectangles, usually squares. Traditionally, the basic writing direction followed the conventions of Chinese handwriting, in top-down vertical lines arranged from right to left across the page. Under the influence of Western printing technologies, a horizontal, left-to-right directionality has become common, and proportional fonts are seeing increased use, particularly in Japan. Horizontal, right-to-left text is also found on occasion, usually for shorter texts such as inscriptions or store signs. Diacritical marks are rarely used, although phonetic annotations are not uncommon. Older editions of the Chinese classics sometimes use the ideographic tone marks (U+302A..U+302D) to indicate unusual pronunciations of characters.

Many older character sets include characters intended to simplify the implementation of East Asian scripts, such as variant punctuation forms for text written vertically, halfwidth forms (which occupy only half a rectangle), and fullwidth forms (which allow Latin letters to occupy a full rectangle). These characters are included in the Unicode Standard for compatibility with older standards.

Appendix E, Han Unification History, describes how the diverse typographic traditions of mainland China, Taiwan, Japan, Korea, and Vietnam have been reconciled to provide a common set of ideographs in the Unicode Standard for all these languages and regions.

The Lisu script was developed in the early 20th century by using a combination of Latin letters, rotated Latin letters, and Latin punctuation repurposed as tone letters, to create a writing system for the Lisu language, spoken by large communities, mostly in Yunnan province in China. It sees considerable use in China, where it has been an official script since 1992.

The Miao script was created in 1904 by adapting Latin letter variants, English shorthand characters, Miao pictographs, and Cree syllable forms. The script was originally developed to write the Northeast Yunnan Miao language of southern China. Today it is also used to write other Miao dialects and the languages of the Yi and Lisu nationalities of southern China.

18.1 Han

CJK Unified Ideographs

The Unicode Standard contains a set of unified Han ideographic characters used in the written Chinese, Japanese, and Korean languages. The term *Han*, derived from the Chinese Han Dynasty, refers generally to Chinese traditional culture. The Han ideographic characters make up a coherent script, which was traditionally written vertically, with the vertical lines ordered from right to left. In modern usage, especially in technical works and in computer-rendered text, the Han script is written horizontally from left to right and is freely mixed with Latin or other scripts. When used in writing Japanese or Korean, the Han characters are interspersed with other scripts unique to those languages (Hiragana and Katakana for Japanese; Hangul syllables for Korean).

Although the term "CJK"—Chinese, Japanese, and Korean—is used throughout this text to describe the languages that currently use Han ideographic characters, it should be noted that earlier Vietnamese writing systems were based on Han ideographs. Consequently, the term "CJKV" would be more accurate in a historical sense. Han ideographs are still used for historical, religious, and pedagogical purposes in Vietnam. For more on usage of the term *ideograph*, see "Logosyllabaries" in *Section 6.1, Writing Systems.*

The term "Han ideographic characters" is used within the Unicode Standard as a common term traditionally used in Western texts, although "sinogram" is preferred by professional linguists. Taken literally, the word "ideograph" applies only to some of the ancient original character forms, which indeed arose as ideographic depictions. The vast majority of Han characters were developed later via composition, borrowing, and other non-ideographic principles, but the term "Han ideographs" remains in English usage as a conventional cover term for the script as a whole.

The Han ideographic characters constitute a very large set, numbering in the tens of thousands. They have a long history of use in East Asia. Enormous compendia of Han ideographic characters exist because of a continuous, millennia-long scholarly tradition of collecting all Han character citations, including variant, mistaken, and nonce forms, into annotated character dictionaries.

The Unicode Standard draws its unified Han character repertoire from a number of different character set standards. These standards are grouped into a number of sources listed in tables in *Section E.3, CJK Sources.*

Because of the large size of the Han ideographic character repertoire, and because of the particular problems that the characters pose for standardizing their encoding, this character block description is more extended than that for other scripts and is divided into several subsections. The first subsection, "Blocks Containing Han Ideographs," describes the way in which the Unicode Standard divides Han ideographs into blocks. This subsection is followed by an extended discussion of the characteristics of Han characters, with particular attention being paid to the problem of unification of encoding for characters used for different languages. There is a formal statement of the principles behind the Unified Han

character encoding adopted in the Unicode Standard and the order of its arrangement. For a detailed account of the background and history of development of the Unified Han character encoding, see *Appendix E, Han Unification History*.

Blocks Containing Han Ideographs

Han ideographic characters are found in seven main blocks of the Unicode Standard, as shown in *Table 18-1*.

Table 18-1. Blocks Containing Han Ideographs

Block	Range	Comment
CJK Unified Ideographs	4E00–9FFF	Common
CJK Unified Ideographs Extension A	3400–4DBF	Rare
CJK Unified Ideographs Extension B	20000–2A6DF	Rare, historic
CJK Unified Ideographs Extension C	2A700–2B73F	Rare, historic
CJK Unified Ideographs Extension D	2B740–2B81F	Uncommon, some in current use
CJK Unified Ideographs Extension E	2B820–2CEAF	Rare, historic
CJK Compatibility Ideographs	F900–FAFF	Duplicates, unifiable variants, corporate characters
CJK Compatibility Ideographs Supplement	2F800–2FA1F	Unifiable variants

Characters in the six unified ideograph blocks are defined by the IRG, based on Han unification principles explained later in this section.

The two compatibility ideographs blocks contain various duplicate or unifiable variant characters encoded for round-trip compatibility with various legacy standards. For historic reasons, the CJK Compatibility Ideographs block also contains twelve CJK unified ideographs. Those twelve ideographs are clearly labeled in the code charts for that block.

The initial repertoire of the CJK Unified Ideographs block included characters submitted to the IRG prior to 1992, consisting of commonly used characters. That initial repertoire, also known as the Unified Repertoire and Ordering, or URO, was derived entirely from the G, T, J, and K sources. It has subsequently been extended with small sets of unified ideographs or ideographic components needed for interoperability with various standards, or for other reasons, as shown in *Table 18-2*.

Table 18-2. Small Extensions to the URO

Range	Version	Comment
9FA6–9FB3	4.1	Interoperability with HKSCS standard
9FB4–9FBB	4.1	Interoperability with GB 18030 standard
9FBC–9FC2	5.1	Interoperability with commercial implementations
9FC3	5.1	Correction of mistaken unification
9FC4–9FC6	5.2	Interoperability with ARIB standard
9FC7–9FCB	5.2	Interoperability with HKSCS standard

Table 18-2. Small Extensions to the URO (Continued)

Range	Version	Comment
9FCC	6.1	Interoperability with commercial implementations
9FCD–9FCF	8.0	Interoperability with TGH 2013 standard
9FD0	8.0	Correction of mistaken unification
9FD1–9FD5	8.0	Miscellaneous urgently needed characters

Characters in the CJK Unified Ideographs Extension A block are rare and are not unifiable with characters in the CJK Unified Ideographs block. They were submitted to the IRG during 1992–1998 and are derived entirely from the G, T, J, K, and V sources.

The CJK Unified Ideographs Extension B block contains rare and historic characters that are also not unifiable with characters in the CJK Unified Ideographs block. They were submitted to the IRG during 1998–2002.

The CJK Unified Ideographs Extension C, D, and E blocks contain rare, historic, or uncommon characters that are not unifiable with characters in any previously encoded CJK Unified Ideographs block. Some Extension D characters are in current use, particularly for Cantonese special use characters in Hong Kong. Extension C ideographs were submitted to the IRG during 2002–2006. Extension D ideographs were submitted to the IRG during 2006–2009. Extension E ideographs were submitted to the IRG during 2006–2013.

The only principled difference in the unification work done by the IRG on the unified ideograph blocks is that the Source Separation Rule (rule R1) was applied only to the original CJK Unified Ideographs block and not to the extension blocks. The Source Separation Rule states that ideographs that are distinctly encoded in a source must not be unified. (For further discussion, see "Principles of Han Unification" later in this section.)

The six unified ideograph blocks are not closed repertoires. Each contains a small range of reserved code points at the end of the block. Additional unified ideographs may eventually be encoded in those ranges—as has already occurred in the CJK Unified Ideographs block itself. There is no guarantee that any such Han ideographic additions would be of the same types or from the same sources as preexisting characters in the block, and implementations should be careful not to make hard-coded assumptions regarding the range of assignments within the Han ideographic blocks in general.

Several Han characters unique to the U source and which are not unifiable with other characters in the CJK Unified Ideographs block are found in the CJK Compatibility Ideographs block. There are 12 of these characters: U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28, and U+FA29. The remaining characters in the CJK Compatibility Ideographs block and the CJK Compatibility Ideographs Supplement block are either duplicates or unifiable variants of a character in one of the blocks of unified ideographs.

IICore. IICore (International Ideograph Core) is a set of important Han ideographs, incorporating characters from all the defined blocks. This set of nearly 10,000 characters has been developed by the IRG and represents the set of characters in everyday use throughout

East Asia. By covering the characters in IICore, developers guarantee that they can handle all the needs of almost all of their customers. This coverage is of particular use on devices such as cell phones or PDAs, which have relatively stringent resource limitations. Characters in IICore are explicitly tagged as such in the Unihan Database (see Unicode Standard Annex #38, "Unicode Han Database (Unihan)").

General Characteristics of Han Ideographs

The authoritative Japanese dictionary *Koujien* (1983) defines Han characters to be:

> ...characters that originated among the Chinese to write the Chinese language. They are now used in China, Japan, and Korea. They are logographic (each character represents a word, not just a sound) characters that developed from pictographic and ideographic principles. They are also used phonetically. In Japan they are generally called *kanji* (Han, that is, Chinese, characters) including the "national characters" (*kokuji*) such as *touge* (mountain pass), which have been created using the same principles. They are also called *mana* (true names, as opposed to *kana*, false or borrowed names).

For many centuries, written Chinese was the accepted written standard throughout East Asia. The influence of the Chinese language and its written form on the modern East Asian languages is similar to the influence of Latin on the vocabulary and written forms of languages in the West. This influence is immediately visible in the mixture of Han characters and native phonetic scripts (*kana* in Japan, *hangul* in Korea) as now used in the orthographies of Japan and Korea (see *Table 18-3*).

Table 18-3. Common Han Characters

Han Character	Chinese	Japanese	Korean	English Translation
天	tiān	ten, ame	chen	heaven, sky
地	dì	chi, tsuchi	ci	earth, ground
人	rén	jin, hito	in	man, person
山	shān	san, yama	san	mountain
水	shuǐ	sui, mizu	swu	water
上	shàng	jou, ue	sang	above
下	xià	ka, shita	ha	below

The evolution of character shapes and semantic drift over the centuries has resulted in changes to the original forms and meanings. For example, the Chinese character 湯 *tāng*

(Japanese *tou* or *yu*, Korean *thang*), which originally meant "hot water," has come to mean "soup" in Chinese. "Hot water" remains the primary meaning in Japanese and Korean, whereas "soup" appears in more recent borrowings from Chinese, such as "soup noodles" (Japanese *tanmen*; Korean *thangmyen*). Still, the identical appearance and similarities in meaning are dramatic and more than justify the concept of a unified Han script that transcends language.

The "nationality" of the Han characters became an issue only when each country began to create coded character sets (for example, China's GB 2312-80, Japan's JIS X 0208-1978, and Korea's KS C 5601-87) based on purely local needs. This problem appears to have arisen more from the priority placed on local requirements and lack of coordination with other countries, rather than out of conscious design. Nevertheless, the identity of the Han characters is fundamentally independent of language, as shown by dictionary definitions, vocabulary lists, and encoding standards.

Terminology. Several standard romanizations of the term used to refer to East Asian ideographic characters are commonly used. They include *hànzì* (Chinese), *kanzi* (Japanese), *kanji* (colloquial Japanese), *hanja* (Korean), and *Chữ hán* (Vietnamese). The standard English translations for these terms are interchangeable: Han character, Han ideographic character, East Asian ideographic character, or CJK ideographic character. For clarity, the Unicode Standard uses some subset of the English terms when referring to these characters. The term *Kanzi* is used in reference to a specific Japanese government publication. The unrelated term *KangXi* (which is a Chinese reign name, rather than another romanization of "Han character") is used only when referring to the primary dictionary used for determining Han character arrangement in the Unicode Standard. (See *Table 18-7*.)

Distinguishing Han Character Usage Between Languages. There is some concern that unifying the Han characters may lead to confusion because they are sometimes used differently by the various East Asian languages. Computationally, Han character unification presents no more difficulty than employing a single Latin character set that is used to write languages as different as English and French. Programmers do not expect the characters "c", "h", "a", and "t" alone to tell us whether *chat* is a French word for cat or an English word meaning "informal talk." Likewise, we depend on context to identify the American hood (of a car) with the British bonnet. Few computer users are confused by the fact that ASCII can also be used to represent such words as the Welsh word *ynghyd*, which are strange looking to English eyes. Although it would be convenient to identify words by language for programs such as spell-checkers, it is neither practical nor productive to encode a separate Latin character set for every language that uses it.

Similarly, the Han characters are often combined to "spell" words whose meaning may not be evident from the constituent characters. For example, the two characters "to cut" and "hand" mean "postage stamp" in Japanese, but the compound may appear to be nonsense to a speaker of Chinese or Korean (see *Figure 18-1*).

Figure 18-1. Han Spelling

切 + 手 = 1. Japanese "stamp"
to cut　hand　　2. Chinese "cut hand"

Even within one language, a computer requires context to distinguish the meanings of words represented by coded characters. The word *chuugoku* in Japanese, for example, may refer to China or to a district in central west Honshuu (see *Figure 18-2*).

Figure 18-2. Semantic Context for Han Characters

中 + 国 = 1. China
middle　country　　2. Chuugoku district of Honshuu

Coding these two characters as four so as to capture this distinction would probably cause more confusion and still not provide a general solution. The Unicode Standard leaves the issues of language tagging and word recognition up to a higher level of software and does not attempt to encode the language of the Han characters.

Simplified and Traditional Chinese. There are currently two main varieties of written Chinese: "simplified Chinese" (*jiǎntǐzì*), used in most parts of the People's Republic of China (PRC) and Singapore, and "traditional Chinese" (*fántǐzì*), used predominantly in the Hong Kong and Macao SARs, Taiwan, and overseas Chinese communities. The process of interconverting between the two is a complex one. This complexity arises largely because a single simplified form may correspond to multiple traditional forms, such as U+53F0 台, which is a traditional character in its own right and the simplified form for U+6AAF 檯, U+81FA 臺, and U+98B1 颱. Moreover, vocabulary differences have arisen between Mandarin as spoken in Taiwan and Mandarin as spoken in the PRC, the most notable of which is the usual name of the language itself: *guóyǔ* (the National Language) in Taiwan and *pǔtōnghuà* (the Common Speech) in the PRC. Merely converting the character content of a text from simplified Chinese to the appropriate traditional counterpart is insufficient to change a simplified Chinese document to traditional Chinese, or vice versa. (The vast majority of Chinese characters are the same in both simplified and traditional Chinese.)

There are two PRC national standards, GB 2312-80 and GB 12345-90, which are intended to represent simplified and traditional Chinese, respectively. The character repertoires of the two are the same, but the simplified forms occur in GB 2312-80 and the traditional ones in GB 12345-90. These are both part of the IRG G source, with traditional forms and simplified forms separated where they differ. As a result, the Unicode Standard contains a number of distinct simplifications for characters, such as U+8AAC 説 and U+8BF4 说.

While there are lists of official simplifications published by the PRC, most of these are obtained by applying a few general principles to specific areas. In particular, there is a set of radicals (such as U+2F94 言 KANGXI RADICAL SPEECH, U+2F99 貝 KANGXI RADICAL SHELL, U+2FA8 門 KANGXI RADICAL GATE, and U+2FC3 鳥 KANGXI RADICAL BIRD) for which sim-

plifications exist (U+2EC8 ⻈ CJK RADICAL C-SIMPLIFIED SPEECH, U+2EC9 贝 CJK RADICAL C-SIMPLIFIED SHELL, U+2ED4 门 CJK RADICAL C-SIMPLIFIED GATE, and U+2EE6 鸟 CJK RADICAL C-SIMPLIFIED BIRD). The basic technique for simplifying a character containing one of these radicals is to substitute the simplified radical, as in the previous example.

The Unicode Standard does not explicitly encode all simplified forms for traditional Chinese characters. Where the simplified and traditional forms exist as different encoded characters, each should be used as appropriate. The Unicode Standard does not specify how to represent a new simplified form (or, more rarely, a new traditional form) that can be derived algorithmically from an encoded traditional form (simplified form).

Dialects and Early Forms of Chinese. Chinese is not a single language, but a complex of spoken forms that share a single written form. Although these spoken forms are referred to as dialects, they are actually mutually unintelligible and distinct languages. Virtually all modern written Chinese is Mandarin, the dominant language in both the PRC and Taiwan. Speakers of other Chinese languages learn to read and write Mandarin, although they pronounce it using the rules of their own language. (This would be like having Spanish children read and write only French, but pronouncing it as if it were Spanish.) The major non-Mandarin Chinese languages are Cantonese (spoken in the Hong Kong and Macao SARs, in many overseas Chinese communities, and in much of Guangdong province), Wu, Min, Hakka, Gan, and Xiang.

Prior to the 20th century, the standard form of written Chinese was literary Chinese, a form derived from the classical Chinese written, but probably not spoken by Confucius in the sixth century BCE.

The ideographic repertoire of the Unicode Standard is sufficient for all but the most specialized texts of modern Chinese, literary Chinese, and classical Chinese. Preclassical Chinese, written using *seal forms* or *oracle bone forms*, has not been systematically incorporated into the Unicode Standard, because those very early, historic forms differed substantially from the classic and modern forms of Han characters. They require investigation and encoding as distinct historic scripts.

Among modern Chinese languages, Cantonese is occasionally found in printed materials; the others are almost never seen in printed form. There is less standardization for the ideographic repertoires of these languages, and no fully systematic effort has been undertaken to catalog the nonstandard ideographs they use. Because of efforts on the part of the government of the Hong Kong SAR, however, the current ideographic repertoire of the Unicode Standard should be adequate for many—but not all—written Cantonese texts.

Sorting Han Ideographs. The Unicode Standard does not define a method by which ideographic characters are sorted; the requirements for sorting differ by locale and application. Possible collating sequences include phonetic, radical-stroke (*KangXi*, *Xinhua Zidian*, and so on), four-corner, and total stroke count. Raw character codes alone are seldom sufficient to achieve a usable ordering in any of these schemes; ancillary data are usually required. (See *Table 18-7* for a summary of the authoritative sources used to determine the order of Han ideographs in the code charts.)

Character Glyphs. In form, Han characters are monospaced. Every character takes the same vertical and horizontal space, regardless of how simple or complex its particular form is. This practice follows from the long history of printing and typographical practice in China, which traditionally placed each character in a square cell. When written vertically, there are also a number of named cursive styles for Han characters, but the cursive forms of the characters tend to be quite idiosyncratic and are not implemented in general-purpose Han character fonts for computers.

There may be a wide variation in the glyphs used in different countries and for different applications. The most commonly used typefaces in one country may not be used in others.

The types of glyphs used to depict characters in the Han ideographic repertoire of the Unicode Standard have been constrained by available fonts. Users are advised to consult authoritative sources for the appropriate glyphs for individual markets and applications. It is assumed that most Unicode implementations will provide users with the ability to select the font (or mixture of fonts) that is most appropriate for a given locale.

Principles of Han Unification

Three-Dimensional Conceptual Model. To develop the explicit rules for unification, a conceptual framework was developed to model the nature of Han ideographic characters. This model expresses written elements in terms of three primary attributes: semantic (meaning, function), abstract shape (general form), and actual shape (instantiated, typeface form). These attributes are graphically represented in three dimensions according to the X, Y, and Z axes (see *Figure 18-3*).

Figure 18-3. Three-Dimensional Conceptual Model
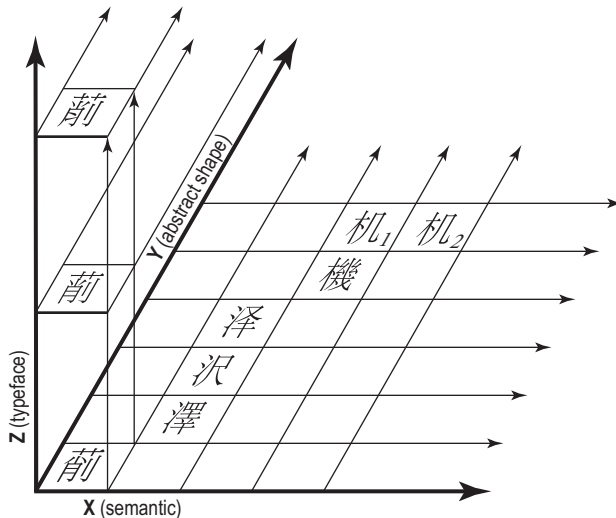
The semantic attribute (represented along the *X* axis) distinguishes characters by meaning and usage. Distinctions are made between entirely unrelated characters such as 澤 (marsh) and 機 (machine) as well as extensions or borrowings beyond the original semantic cluster such as 机₁ (a phonetic borrowing used as a simplified form of 機) and 机₂ (table, the original meaning).

The abstract shape attribute (the *Y* axis) distinguishes the variant forms of a single character with a single semantic attribute (that is, a character with a single position on the *X* axis).

The actual shape (typeface) attribute (the *Z* axis) is for differences of type design (the actual shape used in imaging) of each variant form.

Z-axis typeface and stylistic differences are generally ignored for the purpose of encoding Han ideographs, but can be represented in text by the use of variation sequences; see *Section 23.4, Variation Selectors.*

Unification Rules

The following rules were applied during the process of merging Han characters from the different source character sets.

R1 **Source Separation Rule. If two ideographs are distinct in a primary source standard, then they are not unified.**

- This rule is sometimes called the *round-trip rule* because its goal is to facilitate a round-trip conversion of character data between an IRG source standard and the Unicode Standard without loss of information.

- This rule was applied only for the work on the original CJK Unified Ideographs block [also known as the Unified Repertoire and Ordering (URO)]. The IRG dropped this rule in 1992 and will not use it in future work.

Figure 18-4 illustrates six variants of the CJK ideograph meaning "sword."

Figure 18-4. CJK Source Separation

剣 劍 剱 劔 劎 釖

"sword"

Each of the six variants in *Figure 18-4* is separately encoded in one of the primary source standards—in this case, J0 (JIS X 0208-1990), as shown in *Table 18-4*.

Because the six sword characters are historically related, they are not subject to disunification by the Noncognate Rule (R2) and thus would ordinarily have been considered for possible abstract shape-based unification by R3. Under that rule, the fourth and fifth variants would probably have been unified for encoding. However, the Source Separation Rule required that all six variants be separately encoded, precluding them from any consider-

Table 18-4. Source Encoding for Sword Variants

Unicode	JIS
U+5263	J0-3775
U+528D	J0-5178
U+5271	J0-517B
U+5294	J0-5179
U+5292	J0-517A
U+91FC	J0-6E5F

ation of shape-based unification. Further variants of the "sword" ideograph, U+5251 and U+528E, are also separately encoded because of application of the Source Separation Rule—in that case applied to one or more Chinese primary source standards, rather than to the J0 Japanese primary source standard.

R2   *Noncognate Rule. In general, if two ideographs are unrelated in historical deriva-tion (noncognate characters), then they are not unified.*

For example, the ideographs in *Figure 18-5*, although visually quite similar, are nevertheless not unified because they are historically unrelated and have distinct meanings.

Figure 18-5. Not Cognates, Not Unified

土   ≠   士
earth   warrior, scholar
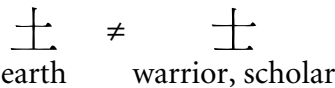
R3   *By means of a two-level classification (described next), the abstract shape of each ideograph is determined. Any two ideographs that possess the same abstract shape are then unified provided that their unification is not disallowed by either the Source Separation Rule or the Noncognate Rule.*

Abstract Shape

Two-Level Classification. Using the three-dimensional model, characters are analyzed in a two-level classification. The two-level classification distinguishes characters by abstract shape (*Y* axis) and actual shape of a particular typeface (*Z* axis). Variant forms are identi-fied based on the difference of abstract shapes.

To determine differences in abstract shape and actual shape, the structure and features of each component of an ideograph are analyzed as follows.

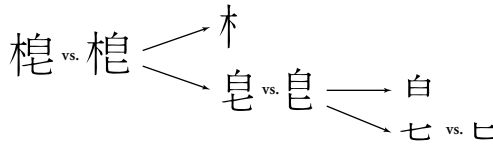***Ideographic Component Structure.*** The component structure of each ideograph is exam-ined. A component is a geometrical combination of primitive elements. Various ideo-graphs can be configured with these components used in conjunction with other components. Some components can be combined to make a component more complicated in its structure. Therefore, an ideograph can be defined as a component tree with the entire

ideograph as the root node and with the bottom nodes consisting of primitive elements (see *Figure 18-6* and *Figure 18-7*).

Figure 18-6. Ideographic Component Structure
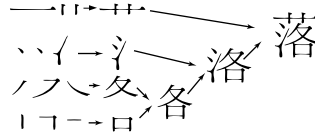
Figure 18-7. The Most Superior Node of an Ideographic Component

Ideograph Features. The following features of each ideograph to be compared are examined:

- Number of components
- Relative positions of components in each complete ideograph
- Structure of a corresponding component
- Treatment in a source character set
- Radical contained in a component

Uniqueness or Unification. If one or more of these features are different between the ideographs compared, the ideographs are considered to have different abstract shapes and, therefore, are considered unique characters and are not unified. If all of these features are identical between the ideographs, the ideographs are considered to have the same abstract shape and are unified.

Spatial Positioning. Ideographs may exist as a unit or may be a component of more complex ideographs. A source standard may describe a requirement for a component with a specific spatial positioning that would be otherwise unified on the principle of having the same abstract shape as an existing full ideograph. Examples of spatial positioning for ideographic components are left half, top half, and so on.

Examples. The examples in *Table 18-5* illustrate the reasons for not unifying characters, including typical differences in abstract character shape.

Table 18-5. Ideographs Not Unified

Characters	Reason
日 ≠ 曰	Non-cognate characters
說 ≠ 説	Characters treated as distinct in a source character set
崖 ≠ 厓	Different number of components
峰 ≠ 峯	Same number of components placed in different relative positions
拡 ≠ 擴	Same number and same relative position of components, corresponding components structured differently
祕 ≠ 秘	Characters with different radical in a component

Differences in the actual shapes of ideographs that *have* been unified are illustrated in *Table 18-6*.

Table 18-6. Ideographs Unified

Characters	Reason
周 ≈ 周	Different writing sequence
雪 ≈ 雪	Differences in overshoot at the stroke termination
酉 ≈ 酉	Differences in contact of strokes
鉅 ≈ 鉅	Differences in protrusion at the folded corner of strokes
堊 ≈ 堊	Differences in bent strokes
朱 ≈ 朱	Differences in stroke termination
父 ≈ 父	Differences in accent at the stroke initiation
八 ≈ 八	Difference in rooftop modification
說 ≈ 説	Difference in rotated strokes/dots[a]

    a.   These ideographs (having the same abstract shape) would have been unified except for the Source Separation Rule.

Han Ideograph Arrangement

The arrangement of the Unicode Han characters is based on the positions of characters as they are listed in four major dictionaries. The *KangXi Zidian* was chosen as primary because it contains most of the source characters and because the dictionary itself and the principles of character ordering it employs are commonly used throughout East Asia.

The Han ideograph arrangement follows the index (page and position) of the dictionaries listed in *Table 18-7* with their priorities.

Table 18-7. Han Ideograph Arrangement

Priority	Dictionary	City	Publisher	Version
1	*KangXi Zidian*	Beijing	Zhonghua Bookstore, 1989	Seventh edition
2	*Dai Kan-Wa Jiten*	Tokyo	Taishuukan Shoten, 1986	Revised edition
3	*Hanyu Da Zidian*	Chengdu	Sichuan Cishu Publishing, 1986	First edition
4	*Dae Jaweon*	Seoul	Samseong Publishing Co. Ltd, 1988	First edition

When a character is found in the *KangXi Zidian*, it follows the *KangXi Zidian* order. When it is not found in the *KangXi Zidian* and it is found in *Dai Kan-Wa Jiten*, it is given a position extrapolated from the *KangXi* position of the preceding character in *Dai Kan-Wa Jiten*. When it is not found in either *KangXi* or *Dai Kan-Wa*, then the *Hanyu Da Zidian* and *Dae Jaweon* dictionaries are consulted in a similar manner.

Ideographs with simplified *KangXi* radicals are placed in a group following the traditional *KangXi* radical from which the simplified radical is derived. For example, characters with the simplified radical 讠 corresponding to *KangXi* radical 言 follow the last nonsimplified character having 言 as a radical. The arrangement for these simplified characters is that of the *Hanyu Da Zidian*.

The few characters that are not found in any of the four dictionaries are placed following characters with the same *KangXi* radical and stroke count. The radical-stroke order that results is a culturally neutral order. It does not exactly match the order found in common dictionaries.

Information for sorting all CJK ideographs by the radical-stroke method is found in the Unihan Database (see Unicode Standard Annex #38, "Unicode Han Database (Unihan)"). It should be used if characters from the various blocks containing ideographs (see *Table 18-1*) are to be properly interleaved. Note, however, that there is no standard way of ordering characters with the same radical-stroke count; for most purposes, Unicode code point order would be as acceptable as any other way.

Details regarding the form of the online charts for the CJK unified ideographs are discussed in *Section 24.2, CJK Ideographs*.

Radical-Stroke Indices

To expedite locating specific Han ideographic characters in the code charts, radical-stroke indices are provided on the Unicode web site. An interactive radical-stroke index page enables queries by specific radical numbers and stroke counts. Two fully formatted traditional radical-stroke indices are also posted in PDF format. The larger of those provides a radical-stroke index for all of the Han ideographic characters in the Unicode Standard, including CJK compatibility ideographs. There is also a more compact radical-stroke index limited to the IICore set of 9,810 CJK unified ideographs in common usage. The following text describes how radical-stroke indices work for Han ideographic characters and explains the particular adaptations which have been made for the Unicode radical-stroke indices.

Under the traditional radical-stroke system, each Han ideograph is considered to be written with one of a number of different character elements or radicals and a number of additional strokes. For example, the character 說 has the radical 言 and seven additional strokes. To find the character 說 within a dictionary, one would first locate the section for its radical, 言, and then find the subsection for characters with seven additional strokes.

This method is complicated by the fact that there are occasional ambiguities in the counting of strokes. Even worse, some characters are considered by different authorities to be written with different radicals; there is not, in fact, universal agreement about which set of radicals to use for certain characters, particularly with the increased use of simplified characters.

The most influential authority for radical-stroke information is the eighteenth-century *KangXi* dictionary, which contains 214 radicals. The main problem in using *KangXi* radicals today is that many simplified characters are difficult to classify under any of the 214 *KangXi* radicals. As a result, various modern radical sets have been introduced. None, however, is in general use, and the 214 *KangXi* radicals remain the best known. See "CJK and KangXi Radicals" in the following text.

The Unicode radical-stroke charts are based on the *KangXi* radicals. The Unicode Standard follows a number of different sources for radical-stroke classification. Where two sources are at odds as to radical or stroke count for a given character, the character is shown in *both* positions in the radical-stroke charts.

Simplified characters are, as a rule, considered to have the same radical as their traditional forms and are found under the appropriate radical. For example, the character 倪 is found under the same radical, 人, as its traditional form (倪).

Mappings for Han Ideographs

The mappings defined by the IRG between the ideographs in the Unicode Standard and the IRG sources are specified in the Unihan Database. These mappings are considered to be normative parts of ISO/IEC 10646 and of the Unicode Standard; that is, the characters are *defined* to be the targets for conversion of these characters in these character set standards.

These mappings have been derived from editions of the source standards provided directly to the IRG by its member bodies, and they may not match mappings derived from the published editions of these standards. For this reason, developers may choose to use alternative mappings more directly correlated with published editions.

Specialized conversion systems may also choose more sophisticated mapping mechanisms—for example, semantic conversion, variant normalization, or conversion between simplified and traditional Chinese.

The Unicode Consortium also provides mapping information that extends beyond the normative mappings defined by the IRG. These additional mappings include mappings to character set standards included in the U source, including duplicate characters from KS C

5601-1987, mappings to portions of character set standards omitted from IRG sources, references to standard dictionaries, and suggested character/stroke counts.

CJK Unified Ideographs Extension B: U+20000–U+2A6D6

The ideographs in the CJK Unified Ideographs Extension B block represent an additional set of 42,711 unified ideographs beyond the 27,496 included in *The Unicode Standard, Version 3.0.* The same principles underlying the selection, organization, and unification of Han ideographs apply to these ideographs.

As with other Han ideograph blocks, the ideographs in the CJK Unified Ideographs Extension B block are derived from versions of national standards submitted to the IRG by its members. They may in some instances differ slightly from published versions of these standards.

CJK Unified Ideographs Extension C: U+2A700–U+2B734

The ideographs in the CJK Unified Ideographs Extension C block represent an additional 4,908 unified ideographs beyond the 70,229 included in *The Unicode Standard, Version 5.0.* The same principles underlying the selection, organization, and unification of Han ideographs apply to these ideographs.

CJK Unified Ideographs Extension D: U+2B740–U+2B81D

The ideographs in the CJK Unified Ideographs Extension D block represent an additional 222 unified ideographs beyond the 74,394 included in *The Unicode Standard, Version 5.2.* The same principles underlying the selection, organization, and unification of Han ideographs apply to these ideographs.

CJK Unified Ideographs Extension E: U+2B820–U+2CEA1

The ideographs in the CJK Unified Ideographs Extension E block represent an additional 5,762 unified ideographs beyond the 74,617 included in *The Unicode Standard, Version 7.0.* The same principles underlying the selection, organization, and unification of Han ideographs apply to these ideographs.

CJK Compatibility Ideographs: U+F900–U+FAFF

The Korean national standard KS C 5601-1987 (now known as KS X 1001:1998), which served as one of the primary source sets for the Unified CJK Ideograph Repertoire and Ordering, Version 2.0, contains 268 duplicate encodings of identical ideograph forms to denote alternative pronunciations. That is, in certain cases, the standard encodes a single character multiple times to denote different linguistic uses. This approach is like encoding the letter "a" five times to denote the different pronunciations it has in the words *hat*, *able*, *art*, *father*, and *adrift*. Because they are in all ways identical in shape to their nominal counterparts, they were excluded by the IRG from its sources. For round-trip conversion with

KS C 5601-1987, they are encoded separately from the primary CJK Unified Ideographs block.

Another 34 ideographs from various regional and industry standards were encoded in this block, primarily to achieve round-trip conversion compatibility. Twelve of these ideographs (U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28, and U+FA29) are not encoded in the CJK Unified Ideographs Areas. These 12 characters are not duplicates and should be treated as a small extension to the set of unified ideographs.

Except for the 12 unified ideographs just enumerated, CJK compatibility ideographs from this block are not used in Ideographic Description Sequences.

An additional 59 compatibility ideographs are found from U+FA30 to U+FA6A. They are included in the Unicode Standard to provide full round-trip compatibility with the ideographic repertoire of JIS X 0213:2000 and should not be used for any other purpose.

An additional three compatibility ideographs are encoded at the range U+FA6B to U+FA6D. They are included in the Unicode Standard to provide full round-trip compatibility with the ideographic repertoire of the Japanese television standard, ARIB STD-B24, and should not be used for any other purpose.

An additional 106 compatibility ideographs are encoded at the range U+FA70 to U+FAD9. They are included in the Unicode Standard to provide full round-trip compatibility with the ideographic repertoire of KPS 10721-2000. They should not be used for any other purpose.

The names for the compatibility ideographs are also algorithmically derived. Thus the name for the compatibility ideograph U+F900 is cjk compatibility ideograph-f900.

All of the compatibility ideographs in this block, except for the 12 unified ideographs, have standardized variation sequences defined in StandardizedVariants.txt. See the discussion in *Section 23.4, Variation Selectors* for more details.

CJK Compatibility Supplement: U+2F800–U+2FA1D

The CJK Compatibility Ideographs Supplement block consists of additional compatibility ideographs required for round-trip compatibility with CNS 11643-1992, planes 3, 4, 5, 6, 7, and 15. They should not be used for any other purpose and, in particular, may not be used in Ideographic Description Sequences.

All of the additional compatibility ideographs in this block have standardized variation sequences defined in StandardizedVariants.txt. See the discussion in *Section 23.4, Variation Selectors* for more details.

Kanbun: U+3190–U+319F

This block contains a set of Kanbun marks used in Japanese texts to indicate the Japanese reading order of classical Chinese texts. These marks are not encoded in any other current

character encoding standards but are widely used in literature. They are typically written in an annotation style to the left of each line of vertically rendered Chinese text. For more details, see JIS X 4051.

Symbols Derived from Han Ideographs

A number of symbols derived from Han ideographs can be found in other blocks. See "Enclosed CJK Letters and Months: U+3200**..**U+32FF," "CJK Compatibility: U+3300**..** U+33FF," and "Enclosed Ideographic Supplement" in *Section 22.10, Enclosed and Square*.

CJK and KangXi Radicals: U+2E80–U+2FD5

East Asian ideographic *radicals* are ideographs or fragments of ideographs used to index dictionaries and word lists, and as the basis for creating new ideographs. The term *radical* comes from the Latin *radix*, meaning "root," and refers to the part of the character under which the character is classified in dictionaries. See "Radical-Stroke Indices" earlier in this section for a more detailed discussion of how ideographic radicals are used in indices.

There is no single radical set in general use throughout East Asia. However, the set of 214 radicals used in the eighteenth-century *KangXi* dictionary is universally recognized.

The visual appearance of radicals is often very different when they are used as radicals than their appearance when they are stand-alone ideographs. Indeed, many radicals have multiple graphic forms when used as parts of characters. A standard example is the water radical, which is written 水 when an ideograph and generally 氵 when part of an ideograph.

The Unicode Standard includes two blocks of encoded radicals: the KangXi Radicals block (U+2F00..U+2FD5), which contains the base forms for the 214 radicals, and the CJK Radicals Supplement block (U+2E80..U+2EF3), which contains a set of variant shapes taken by the radicals either when they occur as parts of characters or when they are used for simplified Chinese. These variant shapes are commonly found as independent and distinct characters in dictionary indices—such as for the radical-stroke charts in the Unicode Standard. As such, they have not been subject to the usual unification rules used for other characters in the standard.

Most of the characters in the CJK and KangXi Radicals blocks are equivalents of characters in the CJK Unified Ideographs block of the Unicode Standard. Radicals that have one graphic form as an ideograph and another as part of an ideograph are generally encoded in both forms in the CJK Unified Ideographs block (such as U+6C34 and U+6C35 for the water radical).

Standards. CNS 11643-1992 includes a block of radicals separate from its ideograph block. This block includes 212 of the 214 KangXi radicals. These characters are included in the KangXi Radicals block.

Those radicals that are ideographs in their own right have a definite meaning and are usually referred to by that meaning. Accordingly, most of the characters in the KangXi Radicals

block have been assigned names reflecting their meaning. The other radicals have been given names based on their shape.

Semantics. Characters in the CJK and KangXi Radicals blocks should never be used as ideographs. They have different properties and meanings. U+2F00 KANGXI RADICAL ONE is not equivalent to U+4E00 CJK UNIFIED IDEOGRAPH-4E00, for example. The former is to be treated as a symbol, the latter as a word or part of a word.

The characters in the CJK and KangXi Radicals blocks are compatibility characters. Except in cases where it is necessary to make a semantic distinction between a Chinese character in its role as a radical and the same Chinese character in its role as an ideograph, the characters from the Unified Ideographs blocks should be used instead of the compatibility radicals. To emphasize this difference, radicals may be given a distinct font style from their ideographic counterparts.

CJK Additions from HKSCS and GB 18030

Several characters have been encoded because of developments in HKSCS-2001 (the Hong Kong Supplementary Character Set) and GB 18030-2000 (the PRC National Standard). Both of these encoding standards were published with mappings to Unicode Private Use Area code points. PUA ideographic characters that could not be remapped to non-PUA CJK ideographs were added to the existing block of CJK Unified Ideographs. Fourteen new ideographs (U+9FA6..U+9FB3) were added from HKSCS, and eight multistroke ideographic components (U+9FB4..U+9FBB) were added from GB 18030.

To complete the mapping to these two Chinese standards, a number of non-ideographic characters were encoded elsewhere in the standard. In particular, two symbol characters from HKSCS were added to the existing Miscellaneous Technical block: U+23DA EARTH GROUND and U+23DB FUSE. A new block, CJK Strokes (U+31C0..U+31EF), was created and populated with a number of stroke symbols from HKSCS. Another block, Vertical Forms (U+FE10..U+FE1F), was created for vertical punctuation compatibility characters from GB 18030.

CJK Strokes: U+31C0–U+31EF

Characters in the CJK Strokes block are single-stroke components of CJK ideographs. The first characters assigned to this block were 16 HKSCS–2001 PUA characters that had been excluded from CJK Unified Ideograph Extension B on the grounds that they were not true ideographs. Further additions consist of traditionally defined stroke types attested in the representative forms appearing in the Unicode CJK ideograph code charts or occurring in pre-unification source glyphs.

CJK strokes are used with highly specific semantics (primarily to index ideographs), but they may lack the monosyllabic pronunciations and logographic functions typically associated with independent ideographs. The strokes in this block are single strokes of well-defined types. For more information about these strokes, see *Appendix F, Documentation of CJK Strokes.*

18.2 Ideographic Description Characters

Ideographic Description: U+2FF0–U+2FFB

Although the Unicode Standard includes more than 75,000 CJK unified ideographs, thousands of extremely rare CJK ideographs have nevertheless been left unencoded. Research into cataloging additional ideographs for encoding continues, but it is anticipated that at no point will the entire set of potential, encodable ideographs be completely exhausted. In particular, ideographs continue to be coined and such new coinages will invariably be unencoded.

The 12 characters in the Ideographic Description block provide a mechanism for the standard interchange of text that must reference unencoded ideographs. Unencoded ideographs can be described using these characters and encoded ideographs; the reader can then create a mental picture of the ideographs from the description.

This process is different from a formal *encoding* of an ideograph. There is no canonical description of unencoded ideographs; there is no semantic assigned to described ideographs; there is no equivalence defined for described ideographs. Conceptually, ideographic descriptions are more akin to the English phrase "an 'e' with an acute accent on it" than to the character sequence <U+0065, U+0301>.

In particular, support for the characters in the Ideographic Description block does *not* require the rendering engine to recreate the graphic appearance of the described character.

Note also that many of the ideographs that users might represent using the Ideographic Description characters will be formally encoded in future versions of the Unicode Standard.

The Ideographic Description Algorithm depends on the fact that virtually all CJK ideographs can be broken down into smaller pieces that are themselves ideographs. The broad coverage of the ideographs already encoded in the Unicode Standard implies that the vast majority of unencoded ideographs can be represented using the Ideographic Description characters.

Although Ideographic Description Sequences are intended primarily to represent unencoded ideographs and should not be used in data interchange to represent encoded ideographs, they also have pedagogical and analytic uses. A researcher, for example, may choose to represent the character U+86D9 蛙 as "□虫圭" in a database to provide a link between it and other characters sharing its phonetic, such as U+5A03 娃. The IRG is using Ideographic Description Sequences in this fashion to help provide a first-approximation, machine-generated set of unifications for its current work.

Applicability to Other Scripts. The characters in the Ideographic Description block are derived from a Chinese standard and were encoded for use specifically in describing CJK ideographs. As a result, the following detailed description of Ideographic Description Sequences is specified entirely in terms of CJK unified ideographs and CJK radicals. However, there are several large, historic East Asian scripts whose writing systems were heavily

influenced by the Han script. Like the Han script, those siniform historic scripts, which include Tangut, Jurchen, and Khitan, are logographic in nature. Furthermore, they built up characters using radicals and components, and with side-by-side and top-to-bottom stacking very similar in structure to the way CJK ideographs are composed. These historic scripts are not yet encoded in Version 7.0 of the Unicode Standard, but it is quite likely that one or more of them will be encoded eventually.

The general usefulness of Ideographic Description Sequences for describing unencoded characters and the applicability of the characters in the Ideographic Description block to description of siniform logographs mean that the syntax for Ideographic Description Sequences can be generalized to extend to additional East Asian logographic scripts.

Ideographic Description Sequences. Ideographic Description Sequences are defined by the following grammar. The list of characters associated with the *Ideographic* and *Radical* properties can be found in the Unicode Character Database.

IDS := *Ideographic* | *Radical* | *CJK_Stroke* | *Private Use* | U+FF1F
      | *IDS_BinaryOperator IDS IDS*
      | *IDS_TrinaryOperator IDS IDS IDS*

CJK_Stroke := U+31C0 | U+31C1 | ... | U+31E3

IDS_BinaryOperator := U+2FF0 | U+2FF1 | U+2FF4 | U+2FF5 | U+2FF6 | U+2FF7 |
                              U+2FF8 | U+2FF9 | U+2FFA | U+2FFB

IDS_TrinaryOperator:= U+2FF2 | U+2FF3

Previous versions of the Unicode standard imposed various limits on the length of a sequence or parts of it, and restricted the use of IDS to Unified CJK Ideographs. Those limits and restrictions are no longer imposed by the standard.

The operators indicate the relative graphic positions of the operands running from left to right and from top to bottom. A user wishing to represent an unencoded ideograph will need to analyze its structure to determine how to describe it using an Ideographic Description Sequence. As a rule, it is best to use the natural radical-phonetic division for an ideograph if it has one and to use as short a description sequence as possible; however, there is no requirement that these rules be followed. Beyond that, the shortest possible Ideographic Description Sequence is preferred.

Figure 18-8 illustrates the use of this grammar to provide descriptions of unencoded ideographs. Examples 9–13 illustrate more complex Ideographic Description Sequences showing the use of some of the less common operators.

Equivalence. Many unencoded ideographs can be described in more than one way using this algorithm, either because the pieces of a description can themselves be broken down further (examples 1–3 in *Figure 18-8*) or because duplications appear within the Unicode Standard (examples 5 and 6 in *Figure 18-8*).

The Unicode Standard does not define equivalence for two Ideographic Description Sequences that are not identical. *Figure 18-8* contains numerous examples illustrating how

Figure 18-8. Using the Ideographic Description Characters

different Ideographic Description Sequences might be used to describe the same ideograph.

In particular, Ideographic Description Sequences should not be used to provide alternative graphic representations of encoded ideographs in data interchange. Searching, collation, and other content-based text operations would then fail.

Interaction with the Ideographic Variation Mark. As with ideographs proper, the Ideographic Variation Mark (U+303E) may be placed before an Ideographic Description Sequence to indicate that the description is merely an approximation of the original ideograph desired. A sequence of characters that includes an Ideographic Variation Mark is not an Ideographic Description Sequence.

Rendering. Ideographic Description characters are visible characters and are not to be treated as control characters. Thus the sequence U+2FF1 U+4E95 U+86D9 must have a distinct appearance from U+4E95 U+86D9.

An implementation may render a valid Ideographic Description Sequence either by rendering the individual characters separately or by parsing the Ideographic Description Sequence and drawing the ideograph so described. In the latter case, the Ideographic Description Sequence should be treated as a ligature of the individual characters for purposes of hit testing, cursor movement, and other user interface operations. (See *Section 5.11, Editing and Selection*.)

Character Boundaries. Ideographic Description characters are not combining characters, and there is no requirement that they affect character or word boundaries. Thus U+2FF1 U+4E95 U+86D9 may be treated as a sequence of three characters or even three words.

Implementations of the Unicode Standard may choose to parse Ideographic Description Sequences when calculating word and character boundaries. Note that such a decision will make the algorithms involved significantly more complicated and slower.

Standards. The Ideographic Description characters are found in GBK—an extension to GB 2312-80 that adds all Unicode ideographs not already in GB 2312-80. GBK is defined as a normative annex of GB 13000.1-93.

18.3 Bopomofo

Bopomofo: U+3100–U+312F

Bopomofo constitute a set of characters used to annotate or teach the phonetics of Chinese, primarily the standard Mandarin language. These characters are used in dictionaries and teaching materials, but not in the actual writing of Chinese text. The formal Chinese names for this alphabet are *Zhuyin-Zimu* ("phonetic alphabet") and *Zhuyin-Fuhao* ("phonetic symbols"), but the informal term "Bopomofo" (analogous to "ABCs") provides a more serviceable English name and is also used in China. The Bopomofo were developed as part of a populist literacy campaign following the 1911 revolution; thus they are acceptable to all branches of modern Chinese culture, although in the People's Republic of China their function has been largely taken over by the Pinyin romanization system.

Bopomofo is a hybrid writing system—part alphabet and part syllabary. The letters of Bopomofo are used to represent either the initial parts or the final parts of a Chinese syllable. The initials are just consonants, as for an alphabet. The finals constitute either simple vowels, vocalic diphthongs, or vowels plus nasal consonant combinations. Because a number of Chinese syllables have no initial consonant, the Bopomofo letters for finals may constitute an entire syllable by themselves. More typically, a Chinese syllable is represented by one initial consonant letter, followed by one final letter. In some instances, a third letter is used to indicate a complex vowel nucleus for the syllable. For example, the syllable that would be written *luan* in Pinyin is segmented l-u-an in Bopomofo—that is, <U+310C, U+3128, U+3122>.

Standards. The standard Mandarin set of Bopomofo is included in the People's Republic of China standards GB 2312 and GB 18030, and in the Republic of China (Taiwan) standard CNS 11643.

Mandarin Tone Marks. Small modifier letters used to indicate the five Mandarin tones are part of the Bopomofo system. In the Unicode Standard they have been unified into the Modifier Letter range, as shown in *Table 18-8*.

Table 18-8. Mandarin Tone Marks

first tone	U+02C9 MODIFIER LETTER MACRON
second tone	U+02CA MODIFIER LETTER ACUTE ACCENT
third tone	U+02C7 CARON
fourth tone	U+02CB MODIFIER LETTER GRAVE ACCENT
light tone	U+02D9 DOT ABOVE

Standard Mandarin Bopomofo. The order of the Mandarin Bopomofo letters U+3105.. U+3129 is standard worldwide. The code offset of the first letter U+3105 BOPOMOFO LETTER B from a multiple of 16 is included to match the offset in the ISO-registered standard GB 2312.

Extended Bopomofo. To represent the sounds of Chinese dialects other than Mandarin, the basic Bopomofo set U+3105..U+3129 has been augmented by additional phonetic characters. These extensions are much less broadly recognized than the basic Mandarin set. The three extended Bopomofo characters U+312A..U+312C are cited in some standard reference works, such as the encyclopedia *Xin Ci Hai*. Another set of 24 extended Bopomofo, encoded at U+31A0..U+31B7, was designed in 1948 to cover additional sounds of the Minnan and Hakka dialects. The extensions are used together with the main set of Bopomofo characters to provide a complete phonetic orthography for those dialects. There are no standard Bopomofo letters for the phonetics of Cantonese or several other Southern Chinese dialects.

The small characters encoded at U+31B4..U+31B7 represent syllable-final consonants not present in standard Mandarin or in Mandarin dialects. They have the same shapes as Bopomofo "b", "d", "k", and "h", respectively, but are rendered in a smaller form than the initial consonants; they are also generally shown close to the syllable medial vowel character. These final letters are encoded separately so that the Minnan and Hakka dialects can be represented unambiguously in plain text without having to resort to subscripting or other fancy text mechanisms to represent the final consonants.

Three Bopomofo letters for sounds found in non-Chinese languages are encoded in the range U+31B8..U+31BA. These characters are used in the Hmu and Ge languages, members of the Hmong-Mien (or Miao-Yao) language family, spoken primarily in southeastern Guizhou. The characters are part of an obsolete orthography for Hmu and Ge devised by the missionary Maurice Hutton in the 1920s and 1930s. A small group of Hmu Christians are still using a hymnal text written by Hutton that contains these characters.

Extended Bopomofo Tone Marks. In addition to the Mandarin tone marks enumerated in *Table 18-8*, other tone marks appropriate for use with the extended Bopomofo transcriptions of Minnan and Hakka can be found in the Modifier Letter range, as shown in *Table 18-9*. The "departing tone" refers to the *qusheng* in traditional Chinese tonal analysis, with the *yin* variant historically derived from voiceless initials and the *yang* variant from voiced initials. Southern Chinese dialects in general maintain more tonal distinctions than Mandarin does.

Table 18-9. Minnan and Hakka Tone Marks

yin departing tone	U+02EA MODIFIER LETTER YIN DEPARTING TONE MARK
yang departing tone	U+02EB MODIFIER LETTER YANG DEPARTING TONE MARK

Rendering of Bopomofo. Bopomofo is rendered from left to right in horizontal text, but also commonly appears in vertical text. It may be used by itself in either orientation, but typically appears in interlinear annotation of Chinese (Han character) text. Children's books are often completely annotated with Bopomofo pronunciations for every character. This interlinear annotation is structurally quite similar to the system of Japanese *ruby* annotation, but it has additional complications that result from the explicit usage of tone marks with the Bopomofo letters.

U+3127 ʙᴏᴘᴏᴍᴏꜰᴏ ʟᴇᴛᴛᴇʀ ɪ has notable variation in rendering in horizontal and vertical layout contexts. In traditional typesetting, the stroke of the glyph was chosen to stand perpendicular to the writing direction. In that practice, the glyph is shown as a horizontal stroke in vertically set text, and as a vertical stroke in horizontally set text. However, modern digital typography has changed this practice. All modern fonts use a horizontal stroke glyph for U+3127, and that form is generally used in both horizontal and vertical layout contexts. In the Unicode Standard, the form in the charts follows the modern practice, showing a horizontal stroke for the glyph; the vertical stroke form is considered to be an occasionally occurring variant. Earlier versions of the standard followed traditional typographic practice, and showed a vertical stroke glyph in the charts.

In horizontal interlineation, the Bopomofo is generally placed above the corresponding Han character(s); tone marks, if present, appear at the end of each syllabic group of Bopomofo letters. In vertical interlineation, the Bopomofo is generally placed on the right side of the corresponding Han character(s); tone marks, if present, appear in a separate interlinear row to the right side of the vowel letter. When using extended Bopomofo for Minnan and Hakka, the tone marks may also be mixed with European digits 0–9 to express changes in actual tonetic values resulting from juxtaposition of basic tones.

18.4 Hiragana and Katakana

Hiragana: U+3040–U+309F

Hiragana is the cursive syllabary used to write Japanese words phonetically and to write sentence particles and inflectional endings. It is also commonly used to indicate the pronunciation of Japanese words. Hiragana syllables are phonetically equivalent to the corresponding Katakana syllables.

Standards. The Hiragana block is based on the JIS X 0208-1990 standard, extended by the nonstandard syllable U+3094 HIRAGANA LETTER VU, which is included in some Japanese corporate standards. Some additions are based on the JIS X 0213:2000 standard.

Combining Marks. Hiragana and the related script Katakana use U+3099 COMBINING KATAKANA-HIRAGANA VOICED SOUND MARK and U+309A COMBINING KATAKANA-HIRAGANA SEMI-VOICED SOUND MARK to generate voiced and semivoiced syllables from the base syllables, respectively. All common precomposed combinations of base syllable forms using these marks are already encoded as characters, and use of these precomposed forms is the predominant JIS usage. These combining marks must follow the base character to which they apply. Because most implementations and JIS standards treat these marks as spacing characters, the Unicode Standard contains two corresponding noncombining (spacing) marks at U+309B and U+309C.

Iteration Marks. The two characters U+309D HIRAGANA ITERATION MARK and U+309E HIRAGANA VOICED ITERATION MARK are punctuation-like characters that denote the iteration (repetition) of a previous syllable according to whether the repeated syllable has an unvoiced or voiced consonant, respectively.

Vertical Text Digraph. U+309F HIRAGANA DIGRAPH YORI is a digraph form which was historically used in vertical display contexts, but which is now also found in horizontal layout.

Katakana: U+30A0–U+30FF

Katakana is the noncursive syllabary used to write non-Japanese (usually Western) words phonetically in Japanese. It is also used to write Japanese words with visual emphasis. Katakana syllables are phonetically equivalent to corresponding Hiragana syllables. Katakana contains two characters, U+30F5 KATAKANA LETTER SMALL KA and U+30F6 KATAKANA LETTER SMALL KE, that are used in special Japanese spelling conventions (for example, the spelling of place names that include archaic Japanese connective particles).

Standards. The Katakana block is based on the JIS X 0208-1990 standard. Some additions are based on the JIS X 0213:2000 standard.

Punctuation-like Characters. U+30FB KATAKANA MIDDLE DOT is used to separate words when writing non-Japanese phrases. U+30A0 KATAKANA-HIRAGANA DOUBLE HYPHEN is a delimiter occasionally used in analyzed Katakana or Hiragana textual material.

U+30FC KATAKANA-HIRAGANA PROLONGED SOUND MARK is used predominantly with Katakana and occasionally with Hiragana to denote a lengthened vowel of the previously written syllable. The two iteration marks, U+30FD KATAKANA ITERATION MARK and U+30FE KATAKANA VOICED ITERATION MARK, serve the same function in Katakana writing that the two Hiragana iteration marks serve in Hiragana writing.

Vertical Text Digraph. U+30FF KATAKANA DIGRAPH KOTO is a digraph form which was historically used in vertical display contexts, but which is now also found in horizontal layout.

Katakana Phonetic Extensions: U+31F0–U+31FF

These extensions to the Katakana syllabary are all "small" variants. They are used in Japan for phonetic transcription of Ainu and other languages. They may be used in combination with U+3099 COMBINING KATAKANA-HIRAGANA VOICED SOUND MARK and U+309A COMBINING KATAKANA-HIRAGANA SEMI-VOICED SOUND MARK to indicate modification of the sounds represented.

Standards. The Katakana Phonetic Extensions block is based on the JIS X 0213:2000 standard.

Kana Supplement U+1B000–U+1B0FF

The Kana Supplement block is intended for the encoding of historic and variant forms of Japanese kana characters, including those variants collectively known as *hentaigana* in Japanese.

Currently this block contains two kana which are of historical use only. These have not been used in Japanese orthography for about one thousand years. The character U+1B000 KATAKANA LETTER ARCHAIC E is an obsolete form of the letter U+30A8 KATAKANA LETTER E. In its pre-10th century use, this letter represented the sound "e", and U+30A8 KATAKANA LETTER E represented the sound "ye". The character U+1B001 HIRAGANA LETTER ARCHAIC YE represents a long-obsolete syllable that would have come between U+3086 HIRAGANA LETTER YU and U+3088 HIRAGANA LETTER YO. This sound merged with "e", which is now represented by U+3048 HIRAGANA LETTER E. These relationships are illustrated in *Figure 18-9.*

Figure 18-9. Japanese Historic Kana for *e* and *ye*

18.5 Halfwidth and Fullwidth Forms

Halfwidth and Fullwidth Forms: U+FF00–U+FFEF

In the context of East Asian coding systems, a double-byte character set (DBCS), such as JIS X 0208-1990 or KS X 1001:1998, is generally used together with a single-byte character set (SBCS), such as ASCII or a variant of ASCII. Text that is encoded with both a DBCS and SBCS is typically displayed such that the glyphs representing DBCS characters occupy two display cells—where a display cell is defined in terms of the glyphs used to display the SBCS (ASCII) characters. In these systems, the two-display-cell width is known as the *fullwidth* or *zenkaku* form, and the one-display-cell width is known as the *halfwidth* or *hankaku* form. While *zenkaku* and *hankaku* are Japanese terms, the display-width concepts apply equally to Korean and Chinese implementations.

Because of this mixture of display widths, certain characters often appear twice—once in fullwidth form in the DBCS repertoire and once in halfwidth form in the SBCS repertoire. To achieve round-trip conversion compatibility with such mixed-width encoding systems, it is necessary to encode both fullwidth and halfwidth forms of certain characters. This block consists of the additional forms needed to support conversion for existing texts that employ both forms.

In the context of conversion to and from such mixed-width encodings, all characters in the General Scripts Area should be construed as halfwidth (*hankaku*) characters if they have a fullwidth equivalent elsewhere in the standard or if they do not occur in the mixed-width encoding; otherwise, they should be construed as fullwidth (*zenkaku*). Specifically, most characters in the CJK Miscellaneous Area and the CJKV Ideograph Area, along with the characters in the CJK Compatibility Ideographs, CJK Compatibility Forms, and Small Form Variants blocks, should be construed as fullwidth (*zenkaku*) characters. For a complete description of the East Asian Width property, see Unicode Standard Annex #11, "East Asian Width."

The characters in this block consist of fullwidth forms of the ASCII block (except SPACE), certain characters of the Latin-1 Supplement, and some currency symbols. In addition, this block contains halfwidth forms of the Katakana and Hangul Compatibility Jamo characters. Finally, a number of symbol characters are replicated here (U+FFE8..U+FFEE) with explicit halfwidth semantics.

Unifications. The fullwidth form of U+0020 SPACE is unified with U+3000 IDEOGRAPHIC SPACE.

18.6 Hangul

Korean Hangul may be considered a featural syllabic script. As opposed to many other syllabic scripts, the syllables are formed from a set of alphabetic components in a regular fashion. These alphabetic components are called *jamo*.

The name *Hangul* itself is just one of several terms that may be used to refer to the script. In some contexts, the preferred term is simply the generic *Korean characters*. *Hangul* is used more frequently in South Korea, whereas a basically synonymous term *Choseongul* is preferred in North Korea. A politically neutral term, *Jeongum*, may also be used.

The Unicode Standard contains both the complete set of precomposed modern Hangul syllable blocks and a set of conjoining Hangul jamo. The conjoining Hangul jamo can be used to to represent all of the modern Hangul syllable blocks, as well as the ancient syllable blocks used in Old Korean. For a description of conjoining jamo behavior and precomposed Hangul syllables, see *Section 3.12, Conjoining Jamo Behavior*. For a discussion of the interaction of combining marks with jamo and Hangul syllables, see "Combining Marks and Korean Syllables" in *Section 3.6, Combination*.

For other blocks containing characters related to Hangul, see "Enclosed CJK Letters and Months: U+3200–U+32FF" and "CJK Compatibility: U+3300–U+33FF" in *Section 22.10, Enclosed and Square*, as well as *Section 18.5, Halfwidth and Fullwidth Forms*.

Hangul Jamo: U+1100–U+11FF

The Hangul Jamo block contains the most frequently used conjoining jamo. These include all of the jamo used in modern Hangul syllable blocks, as well as many of the jamo for Old Korean.

The Hangul jamo are divided into three classes: *choseong* (leading consonants, or syllable-initial characters), *jungseong* (vowels, or syllable-peak characters), and *jongseong* (trailing consonants, or syllable-final characters). Each class may, in turn, consist of one to three subunits. For example, a *choseong* syllable-initial character may either represent a single consonant sound, or a consonant cluster consisting of two or three consonant sounds. Likewise, a *jungseong* syllable-peak character may represent a simple vowel sound, or a complex diphthong or triphthong with onglide or offglide sounds. Each of these complex sequences of two or three sounds is encoded as a single conjoining jamo character. Therefore, a complete Hangul syllable can always be conceived of as a single *choseong* followed by a single *jungseong* and (optionally) a single *jongseong*.

This block also contains two invisible filler characters which act as placeholders for a missing *choseong* or *jungseong* in an incomplete syllable. These filler characters are U+115F HANGUL CHOSEONG FILLER and U+1160 HANGUL JUNGSEONG FILLER.

Hangul Jamo Extended-A: U+A960–U+A97F

This block is an extension of the conjoining jamo. It contains additional complex leading consonants (*choseong*) needed to complete the set of conjoining jamo for the representation of Old Korean.

Hangul Jamo Extended-B: U+D7B0–U+D7FF

This block is an extension of the conjoining jamo. It contains additional complex vowels (*jungseong*) and trailing consonants (*jongseong*) needed to complete the set of conjoining jamo for the representation of Old Korean.

Hangul Compatibility Jamo: U+3130–U+318F

This block consists of spacing, nonconjoining Hangul consonant and vowel (jamo) elements. These characters are provided solely for compatibility with the KS X 1001:1998 standard. Unlike the characters found in the Hangul Jamo block (U+1100..U+11FF), the jamo characters in this block have no conjoining semantics.

The characters of this block are considered to be fullwidth forms in contrast with the halfwidth Hangul compatibility jamo found at U+FFA0..U+FFDF.

Standards. The Unicode Standard follows KS X 1001:1998 for Hangul Jamo elements.

Normalization. When Hangul compatibility jamo are transformed with a compatibility normalization form, NFKD or NFKC, the characters are converted to the corresponding conjoining jamo characters. Where the characters are intended to remain in separate syllables after such transformation, they may require separation from adjacent characters. This separation can be achieved by inserting any non-Korean character.

- U+200B ᴢᴇʀᴏ ᴡɪᴅᴛʜ ѕᴘᴀᴄᴇ is recommended where the characters are to allow a line break.

- U+2060 ᴡᴏʀᴅ ᴊᴏɪɴᴇʀ can be used where the characters are not to break across lines.

Table 18-10 illustrates how two Hangul compatibility jamo can be separated in display, even after transforming them with NFKD or NFKC.

Table 18-10. Separating Jamo Characters

Original	NFKD	NFKC	Display
ㄱ ㅏ 3131 314F	ㄱ ㅏ 1100 1161	가 AC00	가
ㄱ [ZW SP] ㅏ 3131 200B 314F	ㄱ [ZW SP] ㅏ 1100 200B 1161	ㄱ [ZW SP] ㅏ 1100 200B 1161	ㄱㅏ

Hangul Syllables: U+AC00–U+D7A3

The Hangul script used in the Korean writing system consists of individual consonant and vowel letters (jamo) that are visually combined into square display cells to form entire syllable blocks. Hangul syllables may be encoded directly as precomposed combinations of individual jamo or as decomposed sequences of conjoining jamo.

Modern Hangul syllable blocks can be expressed with either two or three jamo, either in the form *consonant + vowel* or in the form *consonant + vowel + consonant*. There are 19 possible leading (initial) consonants (*choseong*), 21 vowels (*jungseong*), and 27 trailing (final) consonants (*jongseong*). Thus there are 399 possible two-jamo syllable blocks and 10,773 possible three-jamo syllable blocks, giving a total of 11,172 modern Hangul syllable blocks. This collection of 11,172 modern Hangul syllables encoded in this block is known as the *Johab* set.

Standards. The Hangul syllables are taken from KS C 5601-1992, representing the full Johab set. This group represents a superset of the Hangul syllables encoded in earlier versions of Korean standards (KS C 5601-1987 and KS C 5657-1991).

Equivalence. Each of the Hangul syllables encoded in this block may be represented by an equivalent sequence of conjoining jamo. The converse is not true because thousands of archaic Hangul syllables may be represented only as a sequence of conjoining jamo.

Hangul Syllable Composition. The Hangul syllables can be derived from conjoining jamo by a regular process of composition. The algorithm that maps a sequence of conjoining jamo to the encoding point for a Hangul syllable in the Johab set is detailed in *Section 3.12, Conjoining Jamo Behavior*.

Hangul Syllable Decomposition. Any Hangul syllable from the Johab set can be decomposed into a sequence of conjoining jamo characters. The algorithm that details the formula for decomposition is also provided in *Section 3.12, Conjoining Jamo Behavior*.

Hangul Syllable Name. The character names for Hangul syllables are derived algorithmically from the decomposition. (For full details, see *Section 3.12, Conjoining Jamo Behavior*.)

Hangul Syllable Representative Glyph. The representative glyph for a Hangul syllable can be formed from its decomposition based on the categorization of vowels shown in *Table 18-11*.

If the vowel of the syllable is based on a vertical line, place the preceding consonant to its left. If the vowel is based on a horizontal line, place the preceding consonant above it. If the vowel is based on a combination of vertical and horizontal lines, place the preceding consonant above the horizontal line and to the left of the vertical line. In either case, place a following consonant, if any, below the middle of the resulting group.

In any particular font, the exact placement, shape, and size of the components will vary according to the shapes of the other characters and the overall design of the font.

Collation. The unit of collation in Korean text is normally the Hangul syllable block. Because the order of the syllables in the Hangul Syllables block reflects the preferred order-

Table 18-11. Line-Based Placement of Jungseong

Vertical		Horizontal		Both	
1161	A	1169	O	116A	WA
1162	AE	116D	YO	116B	WAE
1163	YA	116E	U	116C	OE
1164	YAE	1172	YU	116F	WEO
1165	EO	1173	EU	1170	WE
1166	E			1171	WI
1167	YEO			1174	YI
1168	YE				
1175	I				

ing, sequences of Hangul syllables for modern Korean may be collated with a simple binary comparison.

When Korean text includes sequences of conjoining jamo, as for Old Korean, or mixtures of precomposed syllable blocks and conjoining jamo, the easiest approach for collation is to decompose the precomposed syllable blocks into conjoining jamo before comparing. Additional steps must be taken to ensure that comparison is then done for sequences of conjoining jamo that comprise complete syllables. See Unicode Technical Report #10, "Unicode Collation Algorithm," for more discussion about the collation of Korean.

18.7 Yi

Yi: U+A000–U+A4CF

The Yi syllabary encoded in Unicode is used to write the Liangshan dialect of the Yi language, a member of the Sino-Tibetan language family.

Yi is the Chinese name for one of the largest ethnic minorities in the People's Republic of China. The Yi, also known historically and in English as the Lolo, do not have a single ethnonym, but refer to themselves variously as Nuosu, Sani, Axi or Misapo. According to the 1990 census, more than 6.5 million Yi live in southwestern China in the provinces of Sichuan, Guizhou, Yunnan, and Guangxi. Smaller populations of Yi are also to be found in Myanmar, Laos, and Vietnam. Yi is one of the official languages of the PRC, with between 4 and 5 million speakers.

The Yi language is divided into six major dialects. The Northern dialect, which is also known as the Liangshan dialect because it is spoken throughout the region of the Greater and Lesser Liangshan Mountains, is the largest and linguistically most coherent of these dialects. In 1991, there were about 1.6 million speakers of the Liangshan Yi dialect. The ethnonym of speakers of the Liangshan dialect is Nuosu.

Traditional Yi Script. The traditional Yi script, historically known as Cuan or Wei, is an ideographic script. Unlike in other Chinese-influenced siniform scripts, however, the ideographs of Yi appear not to be derived from Han ideographs. One of the more widespread traditions relates that the script, comprising about 1,840 ideographs, was devised by someone named Aki during the Tang dynasty (618–907 CE). The earliest surviving examples of the Yi script are monumental inscriptions dating from about 500 years ago; the earliest example is an inscription on a bronze bell dated 1485.

There is no single unified Yi script, but rather many local script traditions that vary considerably with regard to the repertoire, shapes, and orientations of individual glyphs and the overall writing direction. The profusion of local script variants occurred largely because until modern times the Yi script was mainly used for writing religious, magical, medical, or genealogical texts that were handed down from generation to generation by the priests of individual villages, and not as a means of communication between different communities or for the general dissemination of knowledge. Although a vast number of manuscripts written in the traditional Yi script have survived to the present day, the Yi script was not widely used in printing before the 20th century.

Because the traditional Yi script is not standardized, a considerable number of glyphs are used in the various script traditions. According to one authority, there are more than 14,200 glyphs used in Yunnan, more than 8,000 in Sichuan, more than 7,000 in Guizhou, and more than 600 in Guangxi. However, these figures are misleading—most of the glyphs are simple variants of the same abstract character. For example, a 1989 dictionary of the Guizhou Yi script contains about 8,000 individual glyphs, but excluding glyph variants reduces this count to about 1,700 basic characters, which is quite close to the figure of 1,840 characters that Aki is reputed to have devised.

Standardized Yi Script. There has never been a high level of literacy in the traditional Yi script. Usage of the traditional script has remained limited even in modern times because the traditional script does not accurately reflect the phonetic characteristics of the modern Yi language, and because it has numerous variant glyphs and differences from locality to locality.

To improve literacy in Yi, a scheme for representing the Liangshan dialect using the Latin alphabet was introduced in 1956. A standardized form of the traditional script used for writing the Liangshan Yi dialect was devised in 1974 and officially promulgated in 1980. The standardized Liangshan Yi script encoded in Unicode is suitable for writing only the Liangshan Yi dialect; it is not intended as a unified script for writing all Yi dialects. Standardized versions of other local variants of traditional Yi scripts do not yet exist.

The standardized Yi syllabary comprises 1,164 signs representing each of the allowable syllables in the Liangshan Yi dialect. There are 819 unique signs representing syllables pronounced in the high level, low falling, and midlevel tones, and 345 composite signs representing syllables pronounced in the secondary high tone. The signs for syllables in the secondary high tone consist of the sign for the corresponding syllable in the midlevel tone (or in three cases the low falling tone), plus a diacritical mark shaped like an inverted breve. For example, U+A001 YI SYLLABLE IX is the same as U+A002 YI SYLLABLE I plus a diacritical mark. In addition to the 1,164 signs representing specific syllables, a syllable iteration mark is used to indicate reduplication of the preceding syllable, which is frequently used in interrogative constructs.

Standards. In 1991, a national standard for Yi was adopted by China as GB 13134-91. This encoding includes all 1,164 Yi syllables as well as the syllable iteration mark, and is the basis for the encoding in the Unicode Standard. The syllables in the secondary high tone, which are differentiated from the corresponding syllable in the midlevel tone or the low falling tone by a diacritical mark, are not decomposable.

Naming Conventions and Order. The Yi syllables are named on the basis of the spelling of the syllable in the standard Liangshan Yi romanization introduced in 1956. The tone of the syllable is indicated by the final letter: "t" indicates the high level tone, "p" indicates the low falling tone, "x" indicates the secondary high tone, and an absence of final "t", "p", or "x" indicates the midlevel tone.

With the exception of U+A015, the Yi syllables are ordered according to their phonetic order in the Liangshan Yi romanization—that is, by initial consonant, then by vowel, and finally by tone (t, x, unmarked, and p). This is the order used in dictionaries of Liangshan Yi that are ordered phonetically.

Yi Syllable Iteration Mark. U+A015 YI SYLLABLE WU does not represent a specific syllable in the Yi language, but rather is used as a syllable iteration mark. Its character properties therefore differ from those for the rest of the Yi syllable characters. The misnomer of U+A015 as YI SYLLABLE WU derives from the fact that it is represented by the letter *w* in the romanized Yi alphabet, and from some confusion about the meaning of the gap in traditional Yi syllable charts for the hypothetical syllable "wu".

The Yi syllable iteration mark is used to replace the second occurrence of a reduplicated syllable under all circumstances. It is very common in both formal and informal Yi texts.

Punctuation. The standardized Yi script does not have any special punctuation marks, but relies on the same set of punctuation marks used for writing modern Chinese in the PRC, including U+3001 IDEOGRAPHIC COMMA and U+3002 IDEOGRAPHIC FULL STOP.

Rendering. The traditional Yi script used a variety of writing directions—for example, right to left in the Liangshan region of Sichuan, and top to bottom in columns running from left to right in Guizhou and Yunnan. The standardized Yi script follows the writing rules for Han ideographs, so characters are generally written from left to right or occasionally from top to bottom. There is no typographic interaction between individual characters of the Yi script.

Yi Radicals. To facilitate the lookup of Yi characters in dictionaries, sets of radicals modeled on Han radicals have been devised for the various Yi scripts. (For information on Han radicals, see "CJK and KangXi Radicals" in *Section 18.1, Han*). The traditional Guizhou Yi script has 119 radicals; the traditional Liangshan Yi script has 170 radicals; and the traditional Yunnan Sani Yi script has 25 radicals. The standardized Liangshan Yi script encoded in Unicode has a set of 55 radical characters, which are encoded in the Yi Radicals block (U+A490..U+A4C5). Each radical represents a distinctive stroke element that is common to a subset of the characters encoded in the Yi Syllables block. The name used for each radical character is that of the corresponding Yi syllable closest to it in shape.

18.8 Lisu

Lisu: U+A4D0–U+A4FF

Somewhere between 1908 and 1914 a Karen evangelist from Myanmar by the name of Ba Thaw modified the shapes of Latin characters and created the Lisu script. Afterwards, British missionary James Outram Fraser and some Lisu pastors revised and improved the script. The script is commonly known in the West as the Fraser script. It is also sometimes called the Old Lisu script, to distinguish it from newer, Latin-based orthographies for the Lisu language.

There are 630,000 Lisu people in China, mainly in the regions of Nujiang, Diqing, Lijiang, Dehong, Baoshan, Kunming and Chuxiong in the Yunnan Province. Another 350,000 Lisu live in Myanmar, Thailand and India. Other user communities are mostly Christians from the Dulong, the Nu and the Bai nationalities in China.

At present, about 200,000 Lisu in China use the Lisu script and about 160,000 in the other countries are literate in it. The Lisu script is widely used in China in education, publishing, the media and religion. Various schools and universities at the national, provincial and prefectural levels have been offering Lisu courses for many years. Globally, the script is also widely used in a variety of Lisu literature.

Structure. There are 40 letters in the Lisu alphabet. These consist of 30 consonants and 10 vowels. Each letter was originally derived from the capital letters of the Latin alphabet. Twenty-five of them look like sans-serif Latin capital letters (all but "Q") in upright positions; the other 15 are derived from sans-serif Latin capital letters rotated 180 degrees.

Although the letters of the Lisu script clearly derived originally from the Latin alphabet, the Lisu script is distinguished from the Latin script. The Latin script is bicameral, with case mappings between uppercase and lowercase letters. The Lisu script is unicameral; it has no casing, and the letters do not change form. Furthermore, typography for the Lisu script is rather sharply distinguished from typography for the Latin script. There is not the same range of font faces as for the Latin script, and Lisu typography is typically monospaced and heavily influenced by the conventions of Chinese typography.

Consonant letters have an inherent [ɑ] vowel unless followed by an explicit vowel letter. Three letters sometimes represent a vowel and sometimes a consonant: U+A4EA ʟɪsᴜ ʟᴇᴛᴛᴇʀ ᴡᴀ, U+A4EC ʟɪsᴜ ʟᴇᴛᴛᴇʀ ʏᴀ, and U+A4ED ʟɪsᴜ ʟᴇᴛᴛᴇʀ ɢʜᴀ.

Tone Letters. The Lisu script has six tone letters which are placed after the syllable to mark tones. These tone letters are listed in *Table 18-12*, with the tones identified in terms of their pitch contours.

Each of the six tone letters represents one simple tone. Although the tone letters clearly derive from Western punctuation marks (full stop, comma, semicolon, and colon), they do not function as punctuation at all. Rather, they are word-forming modifier letters. Furthermore, each tone letter is typeset on an em-square, including those whose visual appearance consists of two marks.

Table 18-12. Lisu Tone Letters

Code	Glyph	Name	Tone
A4F8	.	mya ti	55
A4F9	,	na po	35
A4FA	..	mya cya	44
A4FB	.,	mya bo	33
A4FC	;	mya na	42
A4FD	:	mya jeu	31

The first four tone letters can be used in combination with the last two to represent certain combination tones. Of the various possibilities, only ";" is still in use; the rest are now rarely seen in China.

Other Modifier Letters. Nasalized vowels are denoted by a nasalization mark following the vowel. This word-forming character is not encoded separately in the Lisu script, but is represented by U+02BC MODIFIER LETTER APOSTROPHE, which has the requisite shape and properties (General_Category=Lm) and is used in similar contexts.

A glide based on the vowel *A*, pronounced as [ɑ] without an initial glottal stop (and normally bearing a 31 low falling pitch), is written after a verbal form to mark various aspects. This word-forming modifier letters is represented by U+02CD MODIFIER LETTER LOW MACRON. In a Lisu font, this modifier letter should be rendered on the baseline, to harmonize with the position of the tone letters.

Digits and Separators. There are no unique Lisu digits. The Lisu use European digits for counting. The thousands separator and the decimal point are represented with U+002C COMMA and U+002E FULL STOP, respectively. To separate chapter and verse numbers, U+003A COLON and U+003B SEMICOLON are used. These can be readily distinguished from the similar-appearing tone letters by their numerical context.

Punctuation. U+A4FE "-." LISU PUNCTUATION COMMA and U+A4FF "=" LISU PUNCTUATION FULL STOP are punctuation marks used respectively to denote a lesser and a greater degree of finality. These characters are similar in appearance to sequences of Latin punctuation marks, but are not unified with them.

Over time various other punctuation marks from European or Chinese traditions have been adopted into Lisu orthography. *Table 18-13* lists all known adopted punctuation, along with the respective contexts of use.

U+2010 HYPHEN may be preferred to U+002D HYPHEN-MINUS for the dash used to separate syllables in names, as its semantics are less ambiguous than U+002D.

The use of the U+003F "?" QUESTION MARK replaced the older Lisu tradition of using a tone letter combination to represent the question prosody, followed by a Lisu full stop: "..:="

Linebreaking. A line break is not allowed within an orthographic syllable in Lisu. A line break is also prohibited before a punctuation mark, even if it is preceded by a space. There

Table 18-13. Punctuation Adopted in Lisu Orthography

Code	Glyph	Name	Context
002D	-	hyphen-minus	syllable separation in names
003F	?	question mark	questions
0021	!	exclamation mark	exclamations
0022	"	quotation mark	quotations
0028/0029	( )	parentheses	parenthetical notes
300A/300B	《 》	double angle brackets	book titles
2026	...	ellipsis	omission of words (always doubled in Chinese usage)

is no line-breaking hyphenation of words, except in proper nouns, where a break is allowed after the hyphen used as a syllable separator

Word Separation. The Lisu script separates syllables using a space or, for proper names, a hyphen. In the case of polysyllabic words, it can be ambiguous as to which syllables join together to form a word. Thus for most text processing at the character level, a syllable (starting after a space or punctuation and ending before another space or punctuation) is treated as a word except for proper names—where the occurrence of a hyphen holds the word together.

18.9 Miao

Miao: U+16F00–U+16F9F

The Miao script, also called Lao Miaowen ("Old Miao Script") in Chinese, was created in 1904 by Samuel Pollard and others, to write the Northeast Yunnan Miao language of southern China. The script has also been referred to as the Pollard script, but that usage is no longer preferred. The Miao script was created by an adaptation of Latin letter variants, English shorthand characters, Miao pictographs, and Cree syllable forms. (See *Section 20.2, Canadian Aboriginal Syllabics.*) Today, the script is used to write various Miao dialects, as well as languages of the Yi and Lisu nationalities in southern China.

The script was reformed in the 1950s by Yang Rongxin and others, and was later adopted as the "Normalized" writing system of Kunming City and Chuxiong Prefecture. The main difference between the pre-reformed and the reformed orthographies is in how they mark tones. Both orthographies can be correctly represented using the Miao characters encoded in the Unicode Standard.

Encoding Principles. The script is written left to right. The basic syllabic structure contains an initial consonant or consonant cluster and a final. The final consists of either a vowel or vowel cluster, an optional final nasal, plus a tone mark. The initial consonant may be preceded by U+16F50 MIAO LETTER NASALIZATION, and can be followed by combining marks for voicing (U+16F52 MIAO SIGN REFORMED VOICING) or aspiration (U+16F51 MIAO SIGN ASPIRATION and U+16F53 MIAO SIGN REFORMED ASPIRATION).

Tone Marks. In the Chuxiong reformed orthography, vowels and final nasals appear on the baseline. If no explicit tone mark is present, this indicates the default tone 3. An additional tone mark, encoded in the range U+16F93..U+16F99, may follow the vowel to indicate other tones. A set of archaic tone marks used in the reformed orthography is encoded in the range U+16F9A..U+16F9F.

In the pre-reformed orthography, such as that used for the language Ahmao (Northern Hmong), the tone marks are represented in a different manner, using one of five shifter characters. These are represented in sequence following the vowel or vowel sequence and indicate where the vowel letter is to be rendered in relation to the consonant. If more than one vowel letter appears before the shifter, all of the vowel glyphs are moved together to the appropriate position.

Rendering of "wart". Several Miao consonants appear in the code charts with a "wart" attached to the glyph, usually on the left-hand side. In the Chuxiong orthography, a dot appears instead of the wart on these consonants. Because the user communities consider the appearance of the wart or dot to be a different way to write the same characters and not a difference of the character's identity, the differences in appearance are a matter of font style.

Ordering. The order of Miao characters in the code charts derives from a reference ordering widely employed in China, based in part on the order of Bopomofo phonetic characters. The expected collation order for Miao strings varies by language and user

communities, and requires tailoring. See Unicode Technical Standard #10, "Unicode Collation Algorithm."

Digits. Miao uses European digits.

Punctuation. The Miao script employs a variety of punctuation marks, both from the East Asian typographical tradition and from the Western typographical tradition. There are no script-specific punctuation marks.