This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (http://www.unicode.org/errata/). For information on more recent versions of the standard, see http://www.unicode.org/standard/versions/enumeratedversions.html.

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, http://www.mehallo.com

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: http://www.awprofessional.com

Library of Congress Cataloging-in-Publication Data
The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-321-18578-1 (alk. paper)
1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004
005.7'2—dc21

2003052158

ISBN 0-321-18578-1
Text printed on recycled paper
1 2 3 4 5 6 7 8 9 10—CRW—0706050403
First printing, August 2003

Chapter 11

East Asian Scripts

This chapter presents the following scripts:

- Han
- Bopomofo
- Hiragana
- Katakana
- Hangul
- 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 11.1, Han*.

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, but as Han ideographs were adopted for unrelated languages, extensive modifications were required.

Thus 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 11.3, 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 11.4, Hangul*. The shapes of Hangul syllables or the letter-like *jamos* of which they are comprised 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 twentieth century and are still 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 11.5, Yi*.

Bopomofo, discussed in *Section 11.2, 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 running 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 A, 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.

11.1 Han

CJK Unified Ideographs

The three blocks of CJK unified ideographs contain a set of unified Han ideographic characters used in the written Chinese, Japanese, and Korean languages.[1] 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).

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.

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 subsections. The first two subsections, CJK Standards and Blocks, describe the character set standards used as sources, and the way in which the Unicode Standard divides Han ideographs into blocks. These subsections are 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 order of their arrangement. For a detailed account of the background and history of development of the Unified Han character encoding, see also *Appendix A, Han Unification History*.

CJK Standards

The Unicode Standard draws its Han character repertoire of 70,207 characters from a number of character set standards. These standards are grouped into six sources, as indicated in *Table 11-1*. The primary work of unifying and ordering the characters from these sources

1. 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.

was done by the Ideographic Rapporteur Group (IRG), a subgroup of ISO/IEC JTC1/SC2/WG2.

Table 11-1. Initial Sources for Unified Han

G source:	G0	GB2312-80
	G1	GB12345-90 with 58 Hong Kong and 92 Korean "Idu" characters
	G3	GB7589-87 unsimplified forms
	G5	GB7590-87 unsimplified forms
	G7	General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi
	GS	Singapore Characters
	G8	GB8565-88
	GE	GB16500-95
T source:	T1	CNS 11643-1992 1st plane
	T2	CNS 11643-1992 2nd plane
	T3	CNS 11643-1992 3rd plane with some additional characters
	T4	CNS 11643-1992 4th plane
	T5	CNS 11643-1992 5th plane
	T6	CNS 11643-1992 6th plane
	T7	CNS 11643-1992 7th plane
	TF	CNS 11643-1992 15th plane
J source:	J0	JIS X 0208-1990
	J1	JIS X 0212-1990
	JA	Unified Japanese IT Vendors Contemporary Ideographs, 1993
K source:	K0	KS C 5601-1987 (unique ideographs)
	K1	KS C 5657-1991
	K2	PKS C 5700-1 1994
	K3	PKS C 5700-2 1994
KP source:	KP0	KPS 9566-97
	KP1	KPS 10721-2000
V source:	V0	TCVN 5773:1993
	V1	TCVN 6056:1995
U source:		KS C 5601-1987 (duplicate ideographs)
		ANSI Z39.64-1989 (EACC)
		Big-5 (Taiwan)
		CCCII, level 1
		GB 12052-89 (Korean)
		JEF (Fujitsu)
		PRC Telegraph Code
		Taiwan Telegraph Code (CCDC)
		Xerox Chinese
		Han Character Shapes Permitted for Personal Names (Japan)
		IBM Selected Japanese and Korean Ideographs

The G, T, J, K, KP, and V sources represent the characters submitted to the IRG by its member bodies. The G source consists of submissions from mainland China, the Hong Kong SAR, and Singapore. The other five are the submissions from Taiwan, Japan, South and North Korea, and Vietnam, respectively. The U source represents character set standards that were not submitted to the IRG by any member body but that were used by the Unicode Consortium.

For each of the IRG sources, the table contains an abbreviated source name in the second column and a descriptive source name in the third column. The abbreviated names are used in various data files published by the Unicode Consortium and ISO/IEC to identify the specific IRG sources.

In some cases, the entire ideographic repertoire of the original character set standards was *not* included in the corresponding source. Three reasons explain this decision:

1. Where the repertoires of two of the character set standards within a single source have considerable overlap, the characters in the overlap might be included only once in the source. This approach is used, for example, with GB 2312-80 and GB 12345-90, which have many ideographs in common. Characters in GB 12345-90 that are duplicates of characters in GB 2312-80 are not included in the G source.
2. Where a character set standard is based on unification rules that differ substantially from those used by the IRG, many variant characters found in the character set standard will not be included in the source. This situation is the case with CNS 11643-1992, EACC, and CCCII. It is the only case where full round-trip compatibility with the Han ideograph repertoire of the relevant character set standards is not guaranteed.
3. KS C 5601-1987 contains numerous duplicate ideographs included because they have multiple pronunciations in Korean. These multiply encoded ideographs are not included in the K source but are included in the U source to provide full round-trip compatibility with KS C 5601-1987 (now known as KS X 1001:1998).

Blocks

Ideographs are found in five blocks of the Unicode Standard: the CJK Unified Ideographs block (U+4E00–U+9FFF, common ideographs), the CJK Unified Ideographs Extension A block (U+3400–U+4DFF, rare ideographs), the CJK Unified Ideographs Extension B block (U+20000–U+2A6DF, rare ideographs), the CJK Compatibility Ideographs block (U+F900–U+FAFF, duplicates, unifiable variants, and corporate characters), and CJK Compatibility Ideographs Supplement block (U+2F800–U+2FA1F, CNS unifiable ideographs).

Characters in the CJK Unified Ideographs, CJK Unified Ideographs Extension A, and CJK Unified Ideographs Extension B blocks are defined by the IRG and are derived entirely from the G, T, J, K, and V sources.

The CJK Unified Ideographs block represents characters submitted to the IRG prior to 1992 and consists of commonly used characters. Characters in the CJK Unified Ideographs Extension A block and the CJK Unified Ideographs Extension B block are rarer and not unifiable with characters in the CJK Unified Ideographs block. They were submitted to the IRG during 1992–1998 and 1998–2002, respectively.

The only difference in the unification work done by the IRG on these three blocks is that the source separation rule was applied only to the CJK Unified Ideographs block and not to the CJK Unified Ideogaphs Extension A and Extension B blocks. This rule states that ideographs that are distinct in a source must not be unified. (For further discussion, see the subsection “Principles” later in this section.)

Characters unique to the U source 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 are either duplicates or unifiable variants of characters in the one of the other blocks and are included in the Unicode Standard for reasons of round-trip compatibility.

General Characteristics of Han Ideographs

The authoritative Japanese dictionary *Koujien* 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).[1]

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 11-2*).

Table 11-2. 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 char-

1. Lee Collins' translation from the Japanese, *Koujien*, Izuru, Shinmura, ed. (Tokyo: Iwanami Shoten, 1983).

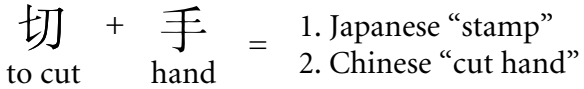
acters is fundamentally independent of language, as shown by dictionary definitions, vocabulary lists, and encoding standards.

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 the purpose of 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 11-6*.)

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.

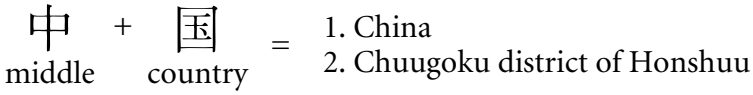
Similarly, the Han characters are often combined to "spell" words whose meaning may not be evident from the constituent characters. For example, the two characters "to cut" and "hand" mean "postage stamp" in Japanese, but the compound may appear to be nonsense to a speaker of Chinese or Korean (see *Figure 11-1*).

Figure 11-1. Han Spelling



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 11-2*).

Figure 11-2. Context for Characters



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" (*jiantizi*), used in most parts of the People's Republic of

China and Singapore, and "traditional Chinese" (*fantizi*), 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 is 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 說 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 simplifications exist (U+2EC8 讠 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 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 Guangzhou province), Wu, Min, Hakka, Gan, and Xiang.

Prior to the twentieth 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. Of Chinese languages, Cantonese is occasionally found in printed materials; the others almost never. 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 11-6*.)

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.

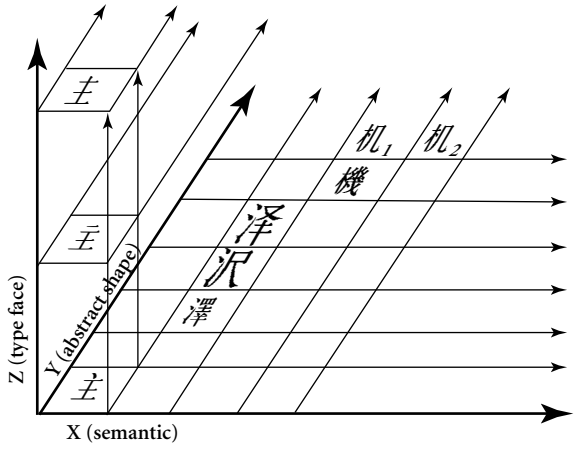
There may be a wide variation in the glyphs used in different countries and for different applications. The most commonly used typefaces in one country may not be used in others.

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

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 11-3*).

Figure 11-3. 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.

Only characters that have the same abstract shape (that is, occupy a single point on the *X* and *Y* axes) are potential candidates for unification. *Z* axis typeface and stylistic differences are generally ignored.

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 or URO). The IRG dropped this rule in 1992 and will not use it in future work.

Figure 11-4 illustrates six variants of the CJK ideograph meaning "sword."

Figure 11-4. Source Separation

"sword"

Each of the six variants in *Figure 11-4* is separately encoded in one of the primary source standards—in this case, J0 (JIS X 0208-1990), as shown in *Table 11-3*.

Table 11-3. Source Encoding for Sword Variants

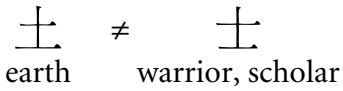
Unicode	JIS
U+5263	J0-3775
U+528D	J0-5178
U+5271	J0-517B
U+5294	J0-5179
U+5292	J0-517A
U+91FC	J0-6E5F

Because the six sword characters are historically related, they are not subject to disunification by the Noncognate Rule (R2 below), and thus would ordinarily have been considered for possible abstract shape-based unification by R3 below. 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 consideration of shape-based unification. Note that 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 derivation (noncognate characters), then they are not unified.

For example, the ideographs in *Figure 11-5*, although visually quite similar, are nevertheless not unified because they are historically unrelated and have distinct meanings.

Figure 11-5. Not Cognates, Not Unified



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.

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 identified 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.

Ideograph Component Structure. The component structure of each ideograph is examined. A component is a geometrical combination of primitive elements. Various ideographs 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 11-6* and *Figure 11-7*).

Figure 11-6. Component Structure

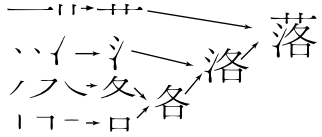
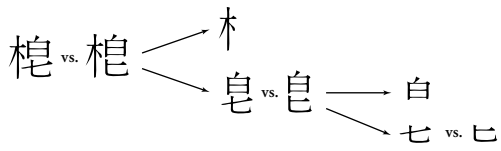


Figure 11-7. The Most Superior Node of a Component



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

- Number of components
- Relative position of components in each complete ideograph
- Structure of a corresponding component

- Treatment in a source character set
- Radical contained in a component

Uniqueness. 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.

Unification. If all of these features are identical between the ideographs, the ideographs are considered to have the same abstract shape and are therefore unified.

The examples in *Table 11-4* represent some typical differences in abstract character shape. The ideographs are therefore *not* unified.

Table 11-4. Ideographs Not Unified

Characters	Reason
崖 ≠ 厓	Different number of components
峰 ≠ 峯	Same number of components placed in different relative position
拡 ≠ 擴	Same number and same relative position of components, corresponding components structure differently
区 ≠ 區	Characters treated differently in a source character set
祕 ≠ 秘	Characters with different radical in a component
爲 ≠ 為	Same abstract shape, different actual shape

Differences in the actual shapes of ideographs that *have* been unified are illustrated in *Table 11-5*.

Table 11-5. 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 11-6* with their priorities.

Table 11-6. Han Ideograph Arrangement

Priority	Dictionary	City	Publisher	Version
1	*KangXi Zidian*	Beijing	Zhonghua Bookstore, 1989	7th edition
2	*Dai Kan-Wa Jiten*	Tokyo	Taishuukan Shoten, 1986	Revised edition
3	*Hanyu Da Zidian*	Chengdu	Sichuan Cishu Publishing, 1986	1st edition
4	*Dae Jaweon*	Seoul	Samseong Publishing Co. Ltd, 1988	1st 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 Unihan.txt in the Unicode Character Database. It should be used if characters from the three unified ideograph blocks (CJK Unified Ideographs, CJK Unified Ideographs Extension A, and CJK Unified Ideographs Extension B) and the compatibility ideographs are to be properly interleaved.

The form of the charts for the CJK unified ideograph blocks is described in the introduction to *Chapter 16, Code Charts*. A full radical-stroke index is also provided in *Chapter 17, Han Radical-Stroke Index*, to help users locate characters in the main charts.

Mapping to Standards

The mappings defined by the IRG between the ideographs in the Unicode Standard and the IRG sources are specified in Unihan.txt in the Unicode Character 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 mapping 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. Because of this, 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 Ext. B: U+20000–U+2A6D6

The ideographs in the CJK Unified Ideographs Extension B block represent an additional set of 42,711 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 the ideographs in the CJK Unified Ideographs Extension B block.

The ideographs in this block are derived from the six IRG sources: G source, H source, T source, J source, K source, and V source. There is no U source for ideographs in the CJK Unified Ideographs Extension B block. The H source represents a new IRG source beyond the ones used for earlier blocks of Han ideographs and is used for characters derived from standards published by the Hong Kong SAR.

The standards and other references associated with these six IRG sources are listed in *Table 11-7*. For each of the six IRG sources, the second column of the table contains an abbreviated name of the source; the third column gives a descriptive name. The abbreviated names are used in various data files published by the Unicode Consortium and ISO/IEC to identify the specific IRG sources. For a more detailed explanation of the format of *Table 11-7*, refer to *Table 11-1*.

Table 11-7. Sources Added for Extension B

G source:	G_KX	KangXi dictionary ideographs (including the addendum) not already encoded in the BMP
	G_HZ	Hanyu Da Zidian ideographs not already encoded in the BMP
	G_CY	Ci Yuan
	G_CH	Ci Hai
	G_HC	Hanyu Da Cidian
	G_BK	Chinese Encyclopedia
	G_FZ	Founder Press System
	G_4K	Siku Quanshu
H source:	H	Hong Kong Supplementary Character Set
T source:	T4	CNS 11643-1992, 4th plane
	T5	CNS 11643-1992, 5th plane
	T6	CNS 11643-1992, 6th plane
	T7	CNS 11643-1992, 7th plane
	TF	CNS 11643-1992, 15th plane
J source:	J3	JIS X 0213:2000, level 3
	J4	JIS X 0213:2000, level 4
K source:	K4	PKS 5700-3:1998
V source:	V0	TCVN 5773:1993
	V2	VHN 01:1998
	V3	VHN 02:1998

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 be slightly different than published versions of these standards.

As with other CJK unified ideographs, the names for these characters are algorithmically assigned. Thus, CJK UNIFIED IDEOGRAPH-20000 is the name for the ideograph at U+20000.

These ideographs may be used in Ideographic Description Sequences, which are described in the following subsection on "Ideographic Description."

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 encoded 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*. They are in all ways identical in shape to their nominal counterparts, and so 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.

In addition, another 34 ideographs from various regional and industry standards were encoded in this block, primarily to achieve round-trip conversion compatibility. Twelve of these 34 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.

The names for the compatibility ideographs are also algorithmically derived. Thus, the name for the compatibility ideograph U+F900 is CJK COMPATIBILITY IDEOGRAPH-F900.

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.

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. They are not encoded in any 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.

See also enclosed CJK letters and months (U+3200..U+32FF) and CJK compatibility (U+3300..U+33FF).

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. *Chapter 17, Han Radical-Stroke Index*, provides information on how to use radical-stroke lookup to locate ideographs encoded in the Unicode Standard.

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 from what it is 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 through U+2FD5), which contains the base forms for the 214 radicals, and the CJK Radicals Supplement block (U+2E80 through 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 of 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 meaning. U+2F00 KANGXI RADICAL ONE is not equivalent to U+4E00 CJK UNIFIED IDEOGRAPH-4E00. 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.

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

Although the Unicode Standard includes more than 70,000 ideographs, many thousands of extremely rare ideographs were nevertheless 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, ideograph descriptions are more akin to the English phrase, "an 'e' with an acute accent on it," than to the character sequence <U+006E, 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.

Ideographic Description Sequences. Ideographic Description Sequences are defined by the following grammar. The list of characters associated with the *Unified_CJK_Ideograph* and *CJK_Radical* properties can be found in the Unicode Character Database. See *Section 0.3, Notational Conventions*, for the notational conventions used here.

IDS := *Unified_CJK_Ideograph* | *CJK_Radical* | *IDS_BinaryOperator IDS IDS*
| *IDS_TrinaryOperator IDS IDS IDS*

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

In addition to the above grammar, Ideographic Description Sequences have two additional length constraints:

- No sequence can be longer than 16 Unicode code points in length.
- No sequence can contain more than six *Unified_CJK_Ideographs* or *CJK_Radicals* in a row without an intervening Ideographic Description character.

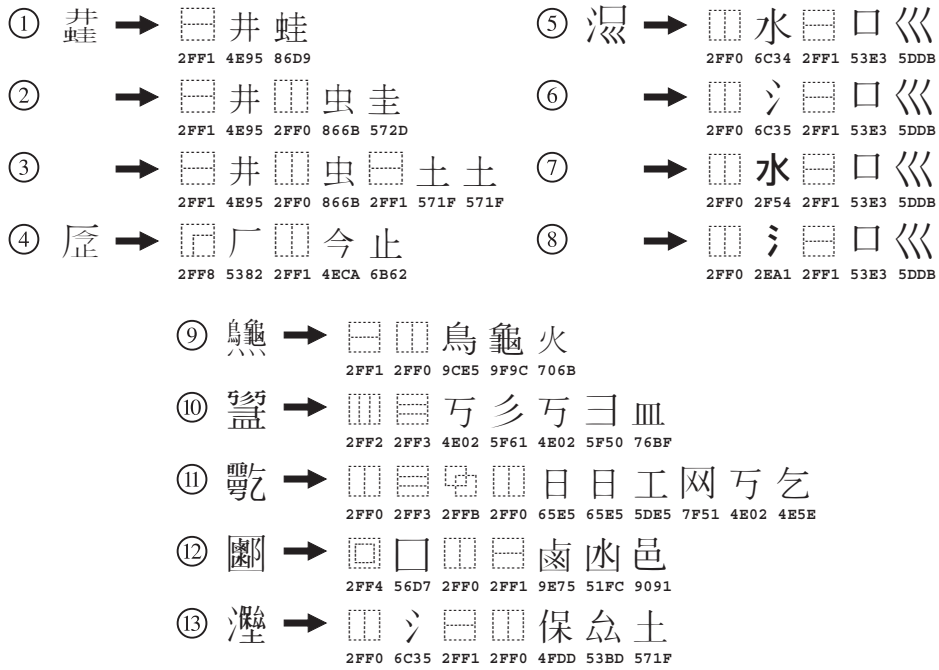
A sequence of characters that includes Ideographic Description characters but does not conform to the above grammar and length constraints is not an Ideographic Description Sequence.

The operators indicate the relative graphic positions of the operands running from left to right and from top to bottom.

Note that non-unique compatibility ideographs (U+F900–U+FA6B and U+2F800–U+2FA1D, but not U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28, or U+FA29) are not counted as unified ideographs for the purposes of this grammar, although they do have the ideographic property (see *Section 4.9, Letters, Alphabetic, and Ideographic*). Non-unique compatibility ideographs are excluded from Ideographic Description Sequences to incrementally reduce the ambiguity of such sequences. Non-unique compatibility ideographs have canonical equivalences, and are excluded on that basis. Some *CJK_Radical* characters have compatibility equivalences to unified ideographs, but compatibility equivalence is not considered a basis for exclusion from Ideographic Description Sequences, because the shape differences involved may be relevant to description of the forms of unencoded ideographs.

Figure 11-8 illustrates the use of this grammar to provide descriptions of unencoded ideographs.

Figure 11-8. Using the Ideographic Description Characters



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, but there is no requirement that these rules be followed. Beyond that, the shortest possible Ideographic Description Sequence is preferred.

The length constraints allow random access into a string of ideographs to have well-defined limits. Only a small number of characters need to be scanned backward to determine whether those characters are part of an Ideographic Description Sequence.

The fact that Ideographic Description Sequences can contain other Ideographic Description Sequences means that implementations may need to be aware of the *recursion depth* of a sequence and its *back-scan length*. The recursion depth of an Ideographic Description Sequence is the maximum number of pending operations encountered in the process of parsing an Ideographic Description Sequence. In *Figure 11-8*, the maximum recursion

depth is shown in the eleventh example, where four operations are still pending at the end of the Ideographic Description Sequence.

The back-scan length is the maximum number of ideographs unbroken by Ideographic Description characters in the sequence. None of the examples in *Figure 11-8* has more than six ideographs in a row; for many, the back-scan length is two.

The Unicode Standard places no formal limits on the recursion depth of Ideographic Description Sequences. It does, however, limit the back-scan length for valid Ideographic Description Sequences to be six or less.

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 11-8*) or because duplications appear within the Unicode Standard (examples 5 and 6 in *Figure 11-8*).

The Unicode Standard does not define equivalence for two Ideographic Description Sequences that are not identical. *Figure 11-8* contains numerous examples illustrating how different Ideographic Description Sequences might be used to describe the same ideograph.

In particular, Ideographic Description Sequences are not to be used to provide alternative graphic representations of encoded ideographs. 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. They are not to be treated as control characters. 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, but 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.

11.2 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. The 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 or the final parts of a Chinese syllable. The initials are just consonants, as for an alphabet. However, 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. And 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 standard GB 2312 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 11-8*.

Table 11-8. 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. The character U+3127 BOPOMOFO LETTER I may be rendered as either a horizontal stroke or a vertical stroke. Often the glyph is chosen to stand perpendicular to the text baseline (for example, a horizontal stroke in vertically set text), but other usage is also common. In the Unicode Standard, the form shown in the charts is a horizontal stroke; the vertical stroke form is considered to be a rendering variant. The variant glyph is not assigned a separate character code.

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 refer-

ence 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 Mandarin dialects. They have the same shapes as Bopomofo "b", "d", "k", and "h", respectively, but are rendered smaller than the initial consonants; they are also generally shown close to the syllable medial vowel character. These final letters are encoded separately so that 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.

Extended Bopomofo Tone Marks. In addition to the Mandarin tone marks enumerated in *Table 11-8*, 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 11-9*. 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 11-9. Minnan and Hakka Tone Marks

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

Rendering of Bopomofo. Bopomofo is rendered left to right in horizontal text, but also commonly appears in vertical text. It may be used by itself in either orientation, but most commonly appears in interlinear annotation of Chinese (Han character) text. It is not uncommon for children's books to be 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.

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 Latin digits 0–9 to express changes in actual tonetic values resulting from juxtaposition of basic tones.

11.3 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 corresponding Katakana syllables.

Standards. The Hiragana block is based on the JIS X 0208-1990 standard, extended by the nonstandard syllable U+3094 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 semi-voiced 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. As most implementations and JIS standard treat these marks as spacing characters, the Unicode Standard also 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 used only when Hiragana is displayed vertically.

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, 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 used only when Katakana is displayed vertically.

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.

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.

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 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 Phonetics and Symbols Area and the Unified CJK 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 characters from the Symbols Area are replicated here (U+FFE8..U+FFEE) with explicit halfwidth semantics.

As with other compatibility characters, the preferred Unicode encoding is to use the nominal counterparts of these characters and use rich text font or style bindings to select the appropriate glyph size and width.

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

11.4 Hangul

Hangul Jamo: U+1100–U+11FF

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 the set of conjoining Hangul jamo. This set of conjoining Hangul jamo can be used to encode all modern and ancient syllable blocks. For a description of conjoining jamo behavior and precomposed Hangul syllables, see *Section 3.12, Conjoining Jamo Behavior,* and the description of the Hangul Syllables block (U+AC00..U+D7A3), which follows in this section.

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). In the following discussion, these classes are abbreviated as *L* (leading consonant), *V* (vowel), and *T* (trailing consonant).

For use in composition, two invisible filler characters act as placeholders for *choseong* or *jungseong*: U+115F HANGUL CHOSEONG FILLER and U+1160 HANGUL JUNGSEONG FILLER.

Collation. The unit of collation in Korean text is normally the Hangul syllable block. Because of the arrangement of the conjoining jamo, their sequences may be collated with a binary comparison. For example, in comparing (a) *LVTLV* with (b) *LVLV*, the first syllable block of (a) *LVT* should be compared with the first syllable block of (b) *LV*. Supposing the first two characters are identical, the *T* would compare as greater than the second *L* in (b) because all trailing consonants have binary values greater than all leading consonants. This result produces the correct ordering between the strings. The positions of the fillers in the code charts were also chosen with this condition in mind.

- As with any coded characters, collation cannot depend simply on a binary comparison. Odd sequences such as superfluous fillers will produce an incorrect sort, as will cases where a non-jamo character follows a sequence (such as comparing *LVT* with *LVX*, where *X* is a Unicode character above U+11FF, such as U+3000 IDEOGRAPHIC SPACE).

If mixtures of precomposed syllable blocks and jamo are collated, the easiest approach is to decompose the precomposed syllable blocks into conjoining jamo before comparing.

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 can be done by inserting any non-Korean character.

- U+200B ZERO-WIDTH SPACE is recommended where the characters are to allow a line break.
- U+2060 WORD JOINER can be used where the characters are not to break across lines.

Figure 11-9 illustrates how two Hangul compatibility jamo can be separated in display, even after transforming them with NFKD or NFKC.

Figure 11-9. 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. The latter encoding is supported by the Hangul Jamo block (U+1100..U+11FF). The syllabic encoding method is described here.

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, for 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, KS C 5657-1991).

Equivalence. Each of the Hangul syllables encoded in this block may be encoded by an equivalent sequence of conjoining jamo; however, the converse is not true because thousands of archaic Hangul syllables may be encoded only as a sequence of conjoining jamo. Implementations that use a conjoining jamo encoding are able to represent these archaic Hangul syllables.

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. Conversely, 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 11-10*.

Table 11-10. Line-Based Placement of Jungseong

Vertical		Horizontal		Horizontal and Vertical	
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				

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.

See also enclosed CJK Letters and Months (U+3200..U+32FF), CJK Compatibility (U+3300..U+33FF), and Halfwidth and Fullwidth Forms (U+FF00..U+FFEF).

11.5 Yi

Yi: U+A000–U+A4CF

The Yi syllabary is used to write the Yi language, a member of the Sino-Tibetan language family. The script is also known as Cuan or Wei.

The Yi, also known as Lolo and Nuo-su, are one of the largest non-Han minorities in the People's Republic of China (PRC). Most live in southwestern China, but others live in Myanmar, Laos, and Vietnam. Yi is one of the official languages of the People's Republic of China.

The earliest surviving samples of classical Yi, an ideographic script, date from about 500 years ago. Unlike other Sinoform scripts, the ideographs themselves appear not to be derived from Han ideographs. There are some 8,000 to 10,000 characters in the classical Yi script, although the exact ideographs used varied from region to region.

To improve literacy in Yi, the Yi syllabary was introduced in the 1970s. This syllabary is encoded in the Unicode Standard; the classical ideographic Yi script is not encoded at this time.

Each Yi syllable consists of a consonantal initial, a final, and a tone. The core Yi syllabary consists of 820 signs for syllables with the first three tones (high, low, and middle low), plus a mark added to the form for the middle low tone to indicate a fourth tone (middle high).

Standards. In 1991, a national standard for Yi was adopted by China as GB 13134-91. This encoding includes all 1,165 Yi syllables and is the basis for the encoding used by the Unicode Standard, which also includes all 1,165 Yi syllables.

Naming Conventions and Order. The Yi syllables are named on the basis of their romanized sound values. The tone is indicated by appending a letter to the romanization: "t" for the high tone, "p" for the low tone, "x" for the middle high tone, and no letter for the middle low tone.

Rendering. Yi follows the writing rules for Han ideographs. Characters are generally written left to right or occasionally 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, a set of radicals has been invented. The Yi repertoire is divided into several subsets, each of which shares a common stroke (radical). The name used for the radical is that of the corresponding Yi character closest to it in shape.