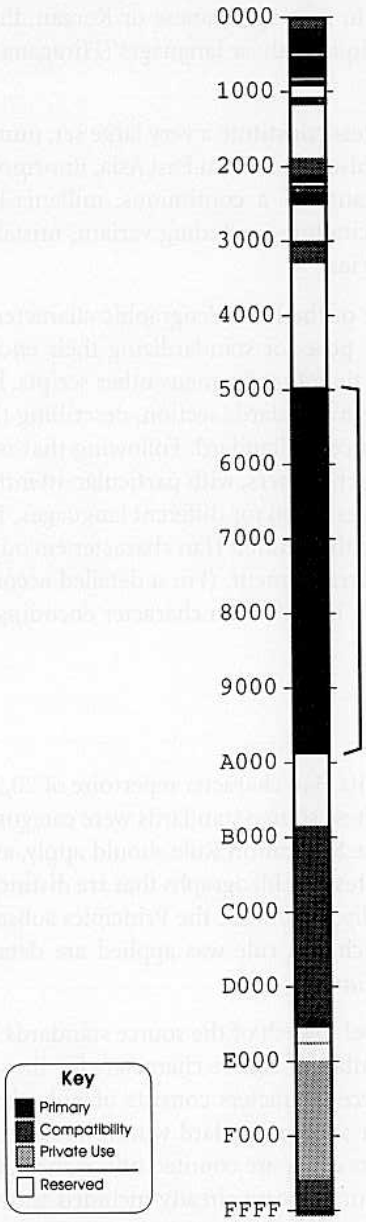
6.4 CJK Ideographs Area

The CJK Ideographs Area of the Unicode Standard encodes the ideographic Han characters (see Figure 6-22).

Figure 6-22. CJK Ideographs



CJK Unified Ideographs: U+4E00—U+9FFF

This block contains a set of Unified Han ideographic characters used in the written Chinese, Japanese, and Korean languages.[1] The term *Han*, derived from *Han Dynasty*, refers generally to Chinese traditional culture. The Han ideographic characters comprise 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 Han ideographic characters constitute a very large set, numbering in the tens of thousands. They have a very long history of use in East Asia. Enormous compendia of Han ideographic characters exist because of a continuous, millenia-long scholarly tradition of collecting all Han character citations, including variant, mistaken, and nonce forms, into annotated character dictionaries.

Because of the very large size of the Han ideographic character repertoire, and because of the particular problems they pose for standardizing their encoding, this character block description is more extended than that for many other scripts. It is divided into the following subsections. First, there is a Standards section, describing the source standards for the repertoire encoded in the Unicode Standard. Following that is an extended discussion of the characteristics of the Han characters, with particular attention to the problem of unification of encoding for characters used for different languages. Then 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 E, Han Unification History.*)

CJK Standards

The Unicode Standard draws its Han character repertoire of 20,902 characters from a number of character standards. These source standards were categorized into two classes: those standards to which the Source Separation Rule should apply, and those to which this rule would not apply. This rule states that ideographs that are distinct in a source standard must not be unified. (For further discussion, see the Principles subsection later in this section.) The source standards to which this rule was applied are detailed in *Table 6-21 Primary Source Standards for Unified Han.*

This table gives a category label to each of the source standards for which the Source Separation Rules applies. The number of source characters for those standards is calculated as follows: The number of source characters consists of only those characters that are not already included from a prior source standard within its same category: G, T, J, or K. For example, only 2,352 characters of G_1 are counted out of the 6,866 characters in GB12345-90 since many characters from G_1 were already included through G_0. However, in some categories, such as the J category, every source character of each standard in that category is counted as a new inclusion since there is no overlap between the standards that comprise the category.

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.

Table 6-21. Primary Source Standards for Unified Han

Category	Standard	Number of Source Characters
G_0	GB2312-80	6,763
G_1	GB12345-90	2,352
G_3	GB7589-87	4,835
G_5	GB7590-87	2,842
G_7	General Use Characters for Modern Chinese	42
G_8	GB8565-88	290
T_1	CNS 11643-1986/1st plane	5,401
T_2	CNS 11643-1986/2nd plane	7,650
T_e	CNS 11643-1986/14th plane	4,198
J_0	JIS X 0208-1990	6,356
J_1	JIS X 0212-1990	5,801
K_0	KS C 5601-1987	4,620
K_1	KS C 5657-1991	2,856

Other contributing standards included by unification without the application of the Source Separation Rule are listed in Table 6-22 without a formal category label.

Table 6-22. Secondary Source Standards for Unified Han

Standard	Number of Source Characters
ANSI Z39.64-1989 (EACC)	13,053
Big-5 (Taiwan)	13,481
CCCII, level 1	4,808
GB 12052-89 (Korean)	94
JEF (Fujitsu)	3,149
PRC Telegraph Code	~8,000
Taiwan Telegraph Code (CCDC)	9,040
Xerox Chinese	9,776

In addition to the standards in the preceding two tables, the following standards were also included by unification without the application of the Source Separation Rule: Han Character Shapes Permitted for Personal Names (Japan) and IBM Selected Japanese and Korean ideographs.

General Characteristics of Han Ideographs

The authoritative Japanese dictionary *Kouzien*, 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 *kanzi* (Han, that is, Chinese, characters) including the "national characters"(*kokuzi*) 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]

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

For many centuries, written Chinese was the accepted written standard throughout East Asia. The impact of the Chinese language and its written form on the modern East Asian languages is similar to the impact of Latin on the vocabulary and written forms of languages in the West. This 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 6-23).

Table 6-23. Common Han Characters

Han Character	*Chinese*[a]	*Japanese*	*Korean*	*English translation*
天	tian[1]	ten, ame	chen	heaven, sky
地	di[4]	ti, tuti	ci	earth, ground
人	ren[2]	zin, hito	in	man, person
山	shan[1]	san, yama	san	mountain
水	shui[3]	sui, mizu	swu	water
上	shang[4]	zyou, ue	sang	above
下	xia[4]	ka, sita	ha	below

a. The superscripted numbers in this table represent Chinese (Mandarin) tone marks.

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 *tang* (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, while "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. But the identity of the Han characters is fundamentally independent of language, as shown by dictionary definitions, vocabulary lists, and encoding standards.

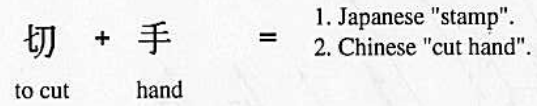
Terminology. Several standard romanizations of the term used to refer to East Asian ideographic characters are commonly used. These include *hanzi* (Chinese), *kanzi* (Japanese), *kanji* (colloquial Japanese), *hanja* (Korean), and Chữ hán (Vietnamese). The standard English translations for these terms 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 dictionary on which the Unified Repertoire and Ordering, Version 2.0 was based.

Distinguishing Han Character Usage Between Languages. There is some concern that unifying the Han characters may lead to confusion because they are sometimes used differ-

ently 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.

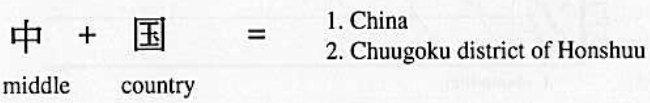
Similarly, the Han characters are often combined to "spell" words whose meaning may not be evident from the constituent characters. For example, the two characters "to cut" and "hand" mean "postal stamp" in Japanese, but the compound may appear to be nonsense to a speaker of Chinese or Korean (see Figure 6-23).

Figure 6-23. 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 6-24).

Figure 6-24. 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.

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 useable ordering in any of these schemes; ancillary data is usually required. (See *Table 6-26 Han Ideograph Arrangement.*)

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 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.

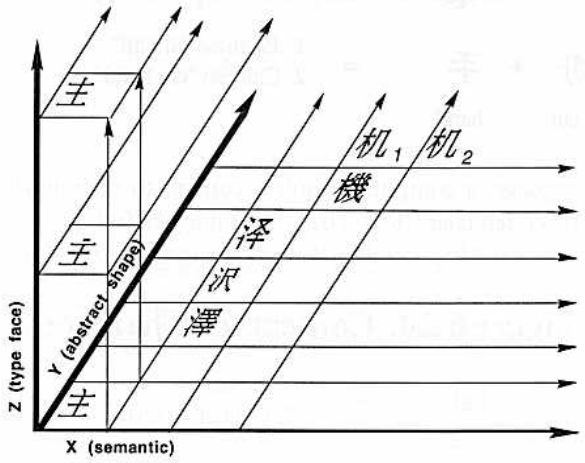
There may be a wide variation in the glyphs used in different countries and for different applications. The most commonly used typefaces in one country may not be used in others.

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 application.

Principles

Three-Dimensional Conceptual Model. In order 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 6-25).

Figure 6-25. 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 semantic 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.

For example, the following ideographs would normally be subject to unification by rule R3; however, their unification is prevented since they are distinct in the primary source standard J_0 (JIS X 0208-1990) (see Figure 6-36).

Figure 6-26. Preserving Variants

劍 劍劒 劒 劒 釼

"sword"

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

R2 ***Non-Cognate Rule. In general, if two ideographs are unrelated in historical derivation (non-cognate characters), then they are not unified.***

For example, the following ideographs (in Figure 6-27) although visually quite similar, are nevertheless not unified since they are historically unrelated, and have distinct meanings.

Figure 6-27. 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 their unification is not disallowed by either the source separation rule or the non-cognate 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.

In order to determine differences in abstract shape and actual shape, the structure and features of each component of an ideograph is 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 Figures 6-28 and 6-29).

Figure 6-28. Component Structure

Figure 6-29. 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 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 6-24 represent some typical differences in abstract character shape. The ideographs are therefore *not* unified.

Table 6-24. 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, Difference in Actual Shape

Differences in actual shape of ideographs that *have* been unified are illustrated in Table 6-25.

Table 6-25. 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
亜 ≈ 亞	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 Arrangment. The arrangement of the Unicode Han characters is based on the position 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 here with their priorities:

Table 6-26. Han Ideograph Arrangement

Priority	Dictionary	City	Publisher	Version
1	*KangXi Zidian*	Beijing	Zhonghua Bookstore, 1989	7th edition
2	*Dai Kanwa Ziten*	Tokyo	Taisyuukan Syoten, 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 Kanwa Ziten*, it is given a position extrapolated from the *KangXi* position of the preceding character in *Dai Kanwa Ziten*. When it is not found in either *KanXi* or *Dai Kanwa*, 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 non-simplified character having 言 as a radical. The arrangement for these simplified characters is that of the *Hanyu Da Zidian*.

The few characters which are not found in any of the four dictionaries are placed following characters with the same *KangXi* radical and stroke count.

Encoding Structure. The Unified CJK Ideographs block occupies range:

U+4E00 → U+9FA5 Unified CJK Ideograph Repertoire

➥ The form of the charts for the Unified CJK Ideographs block differs from that for other blocks; it is described in the introduction to *Chapter 7, Code Charts*. A full radical/stroke index is also provided in *Chapter 8, Han Radical-Stroke Index*, to help users locate characters in the main charts.