3.4 CJK Ideographs

This area of the Unicode standard encodes the ideographic Han characters.

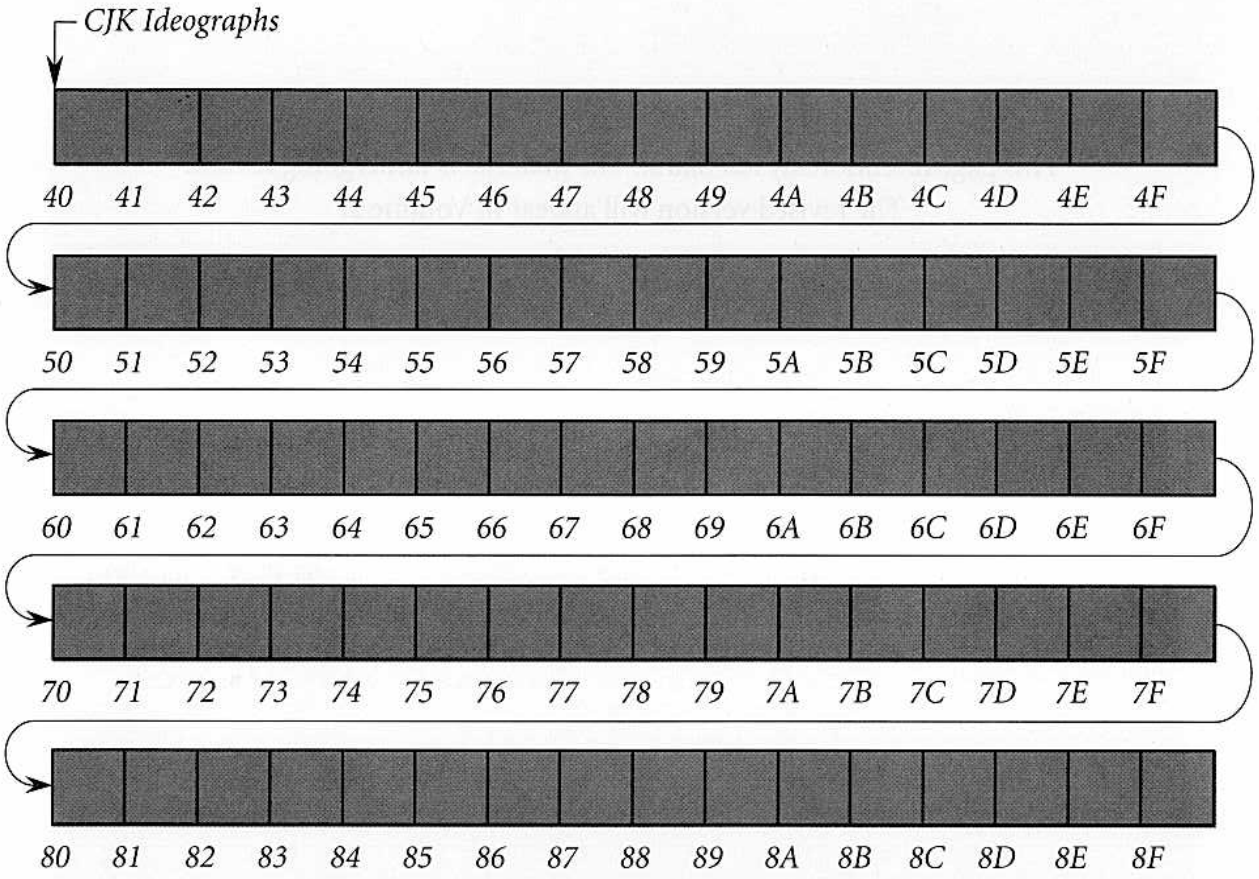

Figure 3-12. CJK Ideographs

Chinese/Japanese/Korean Ideographs U+4000 → U+8BFF

The Unicode standard encodes in this block a set of international Han ideographic characters used in the Chinese, Japanese, and Korean languages.

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]

Until its recent replacement by the English alphabet, written Chinese was the accepted written standard of East Asia. The impact on the writing of the modern Chinese, Japanese, and Korean languages is similar to the impact of Latin on the vocabulary and syntax 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) now used in the orthographies of Japan and Korea.

Han Character	*Chinese*	*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

*The superscripted numbers in this table represent Chinese tone marks.

The evolution of character shapes and semantic drift over the centuries have sometimes 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"

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

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 similarities in appearance and meaning are dramatic and more than justify the Unicode concept of a generic Han script that transcends language.

The "nationality" of the Han characters only became an issue when each country began to create coded character sets (for example, China's GB 2312-80, Japan's JIS X0208-1978, and Korea's KS C 5601-86) based on purely local needs. This problem appears to have arisen more from the priority placed on local requirements, different levels of computerization in the respective countries, and lack of coordination with other countries, rather than out of conscious design.

Efforts to create an international Han character encoding are at least as venerable as the existing national standards. The Chinese Character Code for Information Interchange (CCCII) developed in Taiwan has been in use since 1980. It contains characters for use in China, Taiwan, and Japan. In somewhat modified form, it has been adopted for use in the United states as ANSI Z39.64-1989, also known as the East Asian Character Code (EACC) for bibliographic use. In 1981, Takahashi Tokutaro of Japan's National Diet Library proposed standardization of a character set for common use among East Asian countries.[2]

Of particular relevance to the Unicode standard is China's GB 13000, a universal Han character set that contains all of the characters from the CNS, GB, JIS, and KS C standards. The designers of GB 13000 and of the Unicode Han character set consulted closely for several years on the development of both standards. Because of overlap in their goals and design criteria, in April, 1991, both groups decided to merge their efforts so that the repertoire and ordering of GB 13000 was aligned with that of the Unicode Han character set.

These efforts are all based on the intuitive notion of those literate in the Han script that the identity of the Han characters is independent of language. The existence of an enormous body of cognate characters can be backed by objective evidence including dictionary definitions and vocabulary lists. Statistics assembled by China and the Unicode consortium further show that the overlap among characters encoded in each of the local standards is significant enough for this effort to be worth undertaking.

Distinguishing Han Character Usage between Languages. There is some concern that unifying the Han characters can lead to confusion because they are sometimes used differently in the three languages. Computationally, Han character unification presents no more problems than having a single character set for the Roman alphabet that is used to write languages as different as English and Vietnamese. 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 con-

2. Cited in John Clews, *Language Automation Worldwide: The Development of Character Set Standards*, (Harrowgate, England: Sesame Computer Projects, 1988).

fused 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 which 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 may be nonsense to a speaker of Chinese or Korean. Chinese and Korean use the word 郵票 (Chinese: *youpiao*, Korean: *wuphyo*). However, as a result of Japan's colonial occupation of Korea, many older Koreans probably recognize "cut hand" (*celswu* in Korean) as the Japanese word for "stamp."

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

Even within one language, for a computer to distinguish the meanings of words represented by coded characters requires context. The word *tyhuugoku* in Japanese, for example, may refer to China or to a district in central, west Honshuu:

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

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

Standards. The Unicode standard draws its Han character repertoire from the following Han character standards:

Standard	*Number of Characters*
ANSI Z39.64-1989 (EACC)	13,481
Big Five	13,053
CCCII, level 1	4,808
CNS 11643-1986	13,051
CNS 11643-1986 User Characters	3,418
GB 2312-1980 (GB_0)	6,763
GB 12345-90 (GB_1)*	2,176
GB 7589-87 (GB_3)	7,327
GB 7590-87 (GB_5)	7,039
General Use Characters for Modern Chinese (GB_7)†	41
GB 8565-89 (GB_8)‡	287

Standard	*Number of Characters*
GB 12052-89 (Korean)	94
Han Character Shapes Permitted for Personal Names§	103
IBM Selected Japanese	360
IBM Selected Korean	6
JEF (Fujitsu)	3,149
JIS X 0208-1990	6,355
JIS X 0212-1990	5,801
KS C 5601-1987	4,888
PRC Telegraph Code	~8,000
Taiwan Telegraph Code	9,040
Xerox Chinese	9,776
Total characters covered	~119,016
Total unique characters	21,001

* Characters not already included in GB_0.
† Characters not already included in GB_0, GB_1, GB_3, GB_5, or GB_8.
‡ Characters not already included in GB_0, GB_1, GB_3, or GB_5.
§ The Japanese title of "Han characters Shapes Permitted for Personal Names (Japan)" is "Jinmei-you kanzi kyoyou zitai-hyou," from "Zyouyou kanzi-hyou gendai kanazukai" (Ministry of Finance Printing Department; Tokyo, 1991).

Selection of Han Characters. The Unicode standard preserves the identity of characters across the combined source standards. This allows the Unicode standard to maintain any distinctions in character shape and usage defined as significant in each standard and guarantees a unique mapping to and from the source standards. Thus, where variant forms are given separate codes within one standard, they are also kept separate within the Unicode standard.

For example, the Unicode standard preserves all six variants of the character "sword" found in JIS X 0208-1990:

劍 劍 劒 劔 劒 釼

sword

Note, though, that 劍 and 劔 from KS C 5601-1987 are "unified" with the corresponding characters from JIS X 0208-1990.

Also, the Unicode standard separately codes the approximately 2,000 modern Chinese simplified characters which have corresponding traditional variants in extensions to the GB standards.

The process of merging Han characters from the different source character sets is as follows:

1. Group each character from one of the source character sets with cognate characters from the other sets.

2. Encode cognates that are separate in any one of the source national standards as separate characters in the unified set. This permits a simple round-trip mapping between the unified set and each source set.

3. Encode cognates whose appearance is sufficiently dissimilar and unique to a single source set as separate characters in the unified set.

This algorithm is the method used in creating the Unicode standard and GB 13000. The validity of this approach was verified in 1991 by an independent team of East Asian experts at the University of Toronto. (See the *Unicode CJK Unification Verification Project Final Report.* Kazuko Nakajima, Project Leader, Associate Professor, Department of East Asian Studies, University of Toronto.)

The most interesting question in unifying the Han characters is how to handle variations in character shape across the standards. In unification, the Unicode standard attempts to preserve the same tolerance for variation allowed within any single standard. The Unicode standard relies primarily on the guidelines published by JIS, the Chinese proposal for a common Han character encoding, ISO/JTC1/WG2/N480, and recommendations from the University of Toronto. Where these guidelines suggest that two forms constitute a minor difference, the Unicode standard assigns a single code. Otherwise, separate codes are assigned.

JIS X 0208-1990, §3.4 漢字の異体字の取扱 "The handling of variant Han characters" gives the following six examples of minor differences:

1. Differences in the direction, length, or curve of a stroke:

   羽 ≈ 羽 , 說 ≈ 説

2. Whether strokes touch or intersect each other:

   包 ≈ 包 , 雪 ≈ 雪

3. Fusion of strokes:

   研 ≈ 研 , 每 ≈ 每

4. The addition or subtraction of a stroke:

   者 ≈ 者 , 近 ≈ 近

5. Differences in stroke type:

   靑 ≈ 青 , 喝 ≈ 喝

6. Stroke simplification:

   卽 ≈ 即 , 社 ≈ 社

These rules are applied after filtering out characters with distinct semantics such as

土 ≠ 士

earth warrior, scholar

where a superficial application of rule 1, for example, would result in a false unification.

The small number of national characters invented outside of China such as Korean 𡯂 (*mal*, phonetic used in place names), Japanese 辻 (*tsuji*, cross roads) and 峠 (*touge*, mountain pass) are of course coded separately. Note, however, that 峠 (Korean reading *sang*) is also found in KS C 5601-1987, and therefore unified with the JIS character in the Unicode standard.

Han Character Ordering. The ordering of the Unicode Han characters follows GB 13000. The GB 13000 ordering is based on the position of characters as they are listed in four major Han character dictionaries. In order of priority, these are: the *Kang Xi Zidian* (general East Asia), the *Dai Kanwa Ziten* (Japan), *Hanyu Da Zidian* (China), and the *Dae Jaweon* (Korea). The *Kang Xi 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.

Characters are first assigned a *Kang Xi* ordering if they have one. This becomes the basis for the ordering of the main character list. Characters not found in *Kang Xi* are then ordered according to their position in the *Dai Kanwa Ziten.* These characters are then merged into the main list by placing each of them after the closest preceding *Dai Kanwa Ziten* character that also has a position in the main list. Where there are conflicts, the *Dai Kanwa Ziten* ordering is followed in placing characters after the common character from the main list. The process is then repeated for characters that appear only in the *Hanyu Da Zidian* and the *Dae Jaweon.*

GB characters with simplified *Kang Xi* radicals are placed in a group following the traditional *Kang Xi* radical from which the simplified radical is derived. For example, characters with the simplified radical 讠 corresponding to *Kang Xi* radical 言 follow the last non-simplified character having 言 as a radical. The sub-ordering 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 *Kang Xi* radical and stroke count.

Selection of Glyphs for the Charts. Since the Unicode standard is not a glyph standard, the selection of font for any particular character should not be considered normative. Rather, the intent is to suggest an acceptable range of appearance based on JIS X 0208-1990, §3.4 and on the Chinese principles for recognition of common Han characters, ISO/JTC1/WG2/N480.

The selection of a particular glyph is based on the availability of that glyph in a font. Where several glyphs are available, the preferred order reflects the legibility of the glyphs in the available font. This ordering from clearest to least clear is: Morisawa Ryumin Light, Song, Macintosh Myongjo,

Macintosh Simplified Chinese Screen font, and CCCII bitmap. Many glyphs have been drawn just for the Unicode standard.

Geta in Code Cells. Space within the Unicode Han block has been left for a small number of characters defined in GB 13000 and included in the Unicode standard, but for which glyphs were not available for printing. The proper glyphs will be added in a subsequent version of the Unicode standard.