SINO-PLATONIC PAPERS

Number 360 December, 2024

Kanji and the Computer:

A Brief History of Japanese Character Set Standards

by
James Breen

Victor H. Mair, Editor
Sino-Platonic Papers
Department of East Asian Languages and Civilizations
University of Pennsylvania
Philadelphia, PA 19104-6305 USA
vmair@sas.upenn.edu
www.sino-platonic.org

SINO-PLATONIC PAPERS

FOUNDED 1986

ISSN
2157-9679 (print) 2157-9687 (online)

SINO-PLATONIC PAPERS is an occasional series dedicated to making available to specialists and the interested public the results of research that, because of its unconventional or controversial nature, might otherwise go unpublished. The editor-in-chief actively encourages younger, not yet well established scholars and independent authors to submit manuscripts for consideration.

Contributions in any of the major scholarly languages of the world, including romanized modern standard Mandarin and Japanese, are acceptable. In special circumstances, papers written in one of the Sinitic topolects (*fangyan*) may be considered for publication.

Although the chief focus of *Sino-Platonic Papers* is on the intercultural relations of China with other peoples, challenging and creative studies on a wide variety of philological subjects will be entertained. This series is *not* the place for safe, sober, and stodgy presentations. *Sino-Platonic Papers* prefers lively work that, while taking reasonable risks to advance the field, capitalizes on brilliant new insights into the development of civilization.

Submissions are regularly sent out for peer review, and extensive editorial suggestions for revision may be offered.

Sino-Platonic Papers emphasizes substance over form. We do, however, strongly recommend that prospective authors consult our style guidelines at www.sino-platonic.org/stylesheet.doc.

Manuscripts should be submitted as electronic files in Microsoft Word format. You may wish to use our sample document template, available here: www.sino-platonic.org/spp.dot.

All issues of *Sino-Platonic Papers* are free in PDF form. Issues 1–170, however, will continue to be available in paper copies until our stock runs out.

Please note: When the editor goes on an expedition or research trip, all operations may cease for up to three months at a time.

Kanji and the Computer:
A Brief History of Japanese Character Set Standards

James Breen

Monash University

Melbourne, Australia

ABSTRACT

This paper describes the development of the character coding systems and standards that enable Japanese text to be recorded and used in computer systems. The Japanese coding systems, which were first developed in the late 1970s, pioneered the approaches to handling the large numbers of *kanji* characters and established a pathway that was adopted in other standards for Asian languages. The paper covers the development of the major Japanese standards and their evolution into the Unicode character standard, which is now the basis for all language coding.[1]

INTRODUCTION

These days a person familiar with Japanese can type *kotoba* into a computer, tablet, or mobile phone, have it turn into ことば, and select the preferred *kanji*, e.g. 言葉 (word). Those *kanji* can then be used in an email, web page, etc., with total confidence that people will be able to read them. We take it for granted: of course computers can "do" *kanji*, just as they can do the Latin alphabet or Cyrillic or all those

1 An earlier version of this paper was published on the "Joy o' Kanji" website (https://www.joyokanji.com/).

other funny scripts people use. It's important to be aware that this wasn't always the case, and that the road to where we are now with text in digital form was rather long and complicated, especially with the huge collections of *hanzi* and *kanji*.

For the first couple of decades of computing, digital storage was limited and expensive, and text was usually coded using 6-bit numbers, which allow for 64 combinations. Each number was associated with a specific character, and 64 codes were enough to include the Latin alphabet (uppercase only), numerics, and a selection of punctuation and other special characters. These coding systems were commonly called binary-coded decimal (BCD) systems. Things began to improve in the 1960s when the computing industry started to move to 8-bit units, for which IBM coined the term "bytes," a word that has stuck with us. These allowed for up to 256 combinations, so we finally could use lowercase alphabetic characters as well, and non-English languages could potentially add characters such as é, ö, and ç.

CHARACTER SET STANDARDS: WHO DEVELOPS THEM?

As the exchange of textual information between systems is important, e.g., between companies using different types of computers, standard coding systems are needed. Before getting too far into the details of these coding systems, we need to understand how these sorts of standards are developed and approved. These standards form part of the sets of "industrial standards" (i.e., standards for a given industry) that are an important part of modern life. They cover many topics, and in some areas, such as equipment safety, food preparation and handling, etc., may be enforced by the legal systems. Often, they are concerned with the interworking and interoperability of computer systems and services, and it is in this area that character set standards lie.

Most nations have national organizations with the specific role of developing and maintaining industrial standards. These organizations are typically established as joint government/industry activities, and they prepare, approve, and publish industrial standards in a range of areas. Examples of these organizations are the American National Standards Institute (ANSI) and in Germany the Deutsches Institut für Normung (DIN). The national organizations represent their countries on the UN-related International Organization for Standardization (ISO), which carries out a similar process for

worldwide standards. Many national standards are, in fact, localized versions of international standards. The development of international standards covering information technology is carried out jointly by ISO and the International Electrotechnical Commission (IEC), and hence are typically designated as ISO/IEC standards.

In Japan, industrial standards are developed, approved, and issued by the JSA (Japanese Standards Association, 日本規格協会, *Nihon Kikaku Kyōkai*) in conjunction with the JISC (Japanese Industrial Standards Committee, 日本産業標準調査会, *Nihon Sangyō Hyōjun Kōsakai*). Starting in 1985, computer industry standards were the specific domain of the Information Technology Research and Standardization Center (INSTAC), which was supported and resourced by both government and industry. In 2010, INSTAC activities were absorbed into the JISC.

Most national standard sets are referred to by a coded title that includes identification of the country. In the case of Japan, standards start with the code JIS, for Japanese Industrial Standard. (American standards use the code ANSI, British standards use the code BS, etc.)

As it became apparent that standardizing the representation of text characters in computer files was an important issue, several national and international standards organizations began work on developing more-or-less compatible coding systems. The most famous early standard was ASCII (American Standard Code for Information Interchange), which the American Standards Association (now ANSI) initially approved in 1963. Around the same time ISO approved the very similar coded character set standard known as ISO 646.

THE EARLY JAPANESE CODING STANDARDS

Before proceeding, it is appropriate to summarize briefly the basic aspects of Japanese text. Japanese is written using a combination of one or more of four scripts:

a. *kanji* (Chinese characters). e.g. 辞書, 言葉, etc. About 2,000 *kanji* are studied at school and several thousand more are in reasonably common use. They are used primarily to write nouns and the roots of verbs, adjectives, adverbs, etc.

b. the *hiragana* syllabary (あいうえおかきくけこ...) This consists of 46 characters plus several with diacritics (だ, ぽ, etc.). It is used primarily for particles, conjunctions, short common words, and the inflections of verbs and adjectives.
c. the *katakana* syllabary (アイウエオカキクケコ...) which also consists of 46 characters plus diacritics. In modern Japanese orthography it is used primarily to transcribe loanwords and non-Japanese names. It is also commonly used for the taxonomical names of plants, animals, etc., for slang and colloquial terms, and for emphasis.
d. the Roman alphabet. This is mainly used for acronyms and initialisms, both of foreign terms and names (e.g. CD, DJ) and Japanese terms (e.g., NHK, JR). On occasion Japanese terms are written in this alphabet, a practice known as ローマ字 (*rōmaji*: literally "Roman letters").

After some struggles with 6-bit codes, JSA and associated industry organizations initially concentrated on producing a Japanese equivalent of the 8-bit ISO 646 standard.

As well as the basic alphabetic characters and numerics, this initial standard included the *katakana* syllabary, which was the main Japanese script then used in computing and telecommunications. As the coding space was limited, and a degree of code compatibility with older systems was considered essential, the diacritic marks 濁点 *dakuten* _゛_ as in ブ and 半濁点 *handakuten* _゜_ as in プ were encoded as separate characters. Thus words like パブ (pa-bu: pub) were coded using four characters and typically displayed or printed as ﾊﾟﾌﾞ, in what is now known as 半角カナ *hankaku kana* "half-width *(kata)kana*" The first version of this standard was published as JIS C 6220-1969 in 1969. (JIS C 6220 was later renamed JIS X 0201, and that name will be used here.)

THE ARRIVAL OF *KANJI*

During the 1970s work began in a rather confusing set of government and industry committees on selecting a set of *kanji* that could be included in a national standard. (In parallel with this, a number of companies, such as Sharp and Toshiba, began developing what were to become the first Japanese word-processing systems.) It was realized that, given the number of possible characters, the code would need to use two bytes per character, an approach that came to be called "double-byte coding." For technical

reasons to do with the prevailing methods of transmitting data, a model was chosen based on ISO/IEC 2022 (a standard for encoding multiple character sets in a document), which limited each byte to the values assigned to the 94 printable ASCII characters (33 to 126). This effectively put a ceiling of 8,836 (94×94) on the number of characters that could be encoded.

For the *kanji* selection, there were, of course, the 1,850 当用漢字 *tōyō kanji* (general-use *kanji*; established by the Ministry of Education in 1946 as part of a script reform process) and the 人名用漢字 *jinmeiyō kanji* (additional *kanji* approved for use in personal names, initially 92 but expanded to 120 in 1976), however there were many other *kanji* in reasonably common use. In 1971 the Information Processing Society of Japan drew up a list of 6,086 suitable *kanji*, and in 1975 the Administrative Management Agency of the government identified 2,817 used in the bureaucracy. Also considered were *kanji* used in the registration of the names of persons (3,044) and of administrative districts (3,251). Of course, many of these lists overlapped. Then there were all the *kana*, alphabetic characters, special characters, etc., that needed to be included.

In addition to the question of which *kanji* were to be included, there was the matter of how they were to be ordered. At that time the input methods we use today were years in the future, and the typewriters and typesetting facilities at the time generally used visual identification to select *kanji*.

The first Japanese industrial standard to include *kanji* was released by JSA on 1 January 1978 as JIS C 6226-1978 (情報交換用漢字符号系 *Jōhō Kōkan'yō Kanji Fugōkei*), with the rather clumsy English title of "Code of Japanese Graphic Character Set for Information Interchange."

The standard has the characters organized into 94 rows, each containing up to 94 characters. The first byte of the code indicates the row, and the second indicates the column, i.e., the position in the row. This row-column approach is referred to in Japanese as the 区点 *kuten* system.

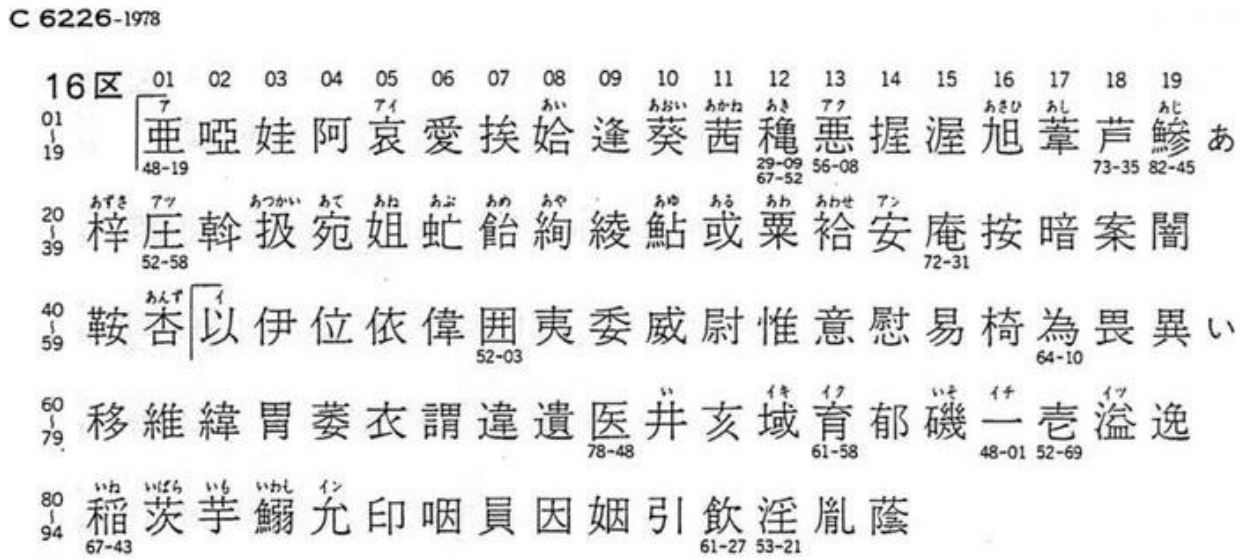
The standard contained the following:

- 453 non-*kanji* characters, including *hiragana*, *katakana*, the Latin, Greek and Cyrillic alphabets, and punctuation and other special characters. Each block was identified by the value of the first byte: row 3 (coded as hexadecimal 23) contained the alphanumeric characters, row 4 the *hiragana*, row 5 the *katakana*, etc.
- 2,965 *kanji* designated "Level 1 *Kanji*" in rows 16 to 47. These included the 1,850 *tōyō kanji* plus other relatively common *kanji*. The *kanji* are arranged in order of their readings; either the *on*-reading

or, in its absence, the *kun*-reading, presumably to enable them to be looked up in an ordered list. For example, row 16 begins with 亜, 唖, 娃, 阿, ..., all of which have the reading ア (*a*).

- 3,384 *kanji*, designated "Level 2 *Kanji*" in rows 48 to 83. In contrast to the Level 1 *kanji*, these *kanji* are in the traditional radical/stroke-count order commonly used in *kanji* dictionaries, e.g., in row 48, there is the sequence 亠, 亢, 亰, 亳, ... of *kanji* with the 亠 radical in ascending stroke-count order.

The following extract from the printed JIS C 6226 standard shows the contents of the first row of *kanji* (row 16). The readings used to order the *kanji* are above the first in each sequence; *katakana* for the *on*-readings and *hiragana* for the *kun*-readings. The number-pairs that are included below some of the kanji, e.g., 56-08 in the case of 悪, are the codes for the matching 異体字 *itaiji* (variant characters) (惡 in the case of 悪.)



Extract from JIS C 6226-1978

The establishment of the JIS C 6226 standard was a major milestone in the coding of characters used in East Asian languages, and it set a direction that was to be followed in other coding standards in the following years. Many of the structures and features in the standard were followed by other countries in their equivalent standards. For example, the GB 2312-80 standard in the PRC also used two

levels for the coding of *hanzi*, and even had *hiragana* and *katakana* in the same places. The JIS standard did, however, include a few blunders, which will be discussed below. This standard, too, was later renamed, becoming JIS X 0208.

JAPANESE TEXT ENCODING

Unlike the single-byte JIS X 0201 characters, which were broadly equivalent to ASCII/ISO 646, the codes used in the JIS X 0208 standard were not suitable for using freely in computer text, especially if they were being mixed with the single-byte alphabetic and numeric characters from those standards. To enable their use in such mixing of codes, three mutually-incompatible encoding (also known as encapsulation) approaches were devised to enable the double-byte characters to be distinguished from the single-byte characters, and mixed with them in computer text:

- JIS Encoding, which is an approach based on the JIS X 0202 standard, derived from the ISO/IEC 2022 standard (mentioned above). JIS Encoding uses "escape sequences" of special characters to switch between strings of text using the single-byte JIS X 0201 characters and two-byte JIS X 0208 characters. This technique was particularly used in email, bulletin boards, and related systems, as they were usually constrained to using only 7-bit characters.
- EUC (Extended Unix Code). As with JIS Encoding, EUC is also based on options in ISO/IEC 2022, and for Japanese text it typically took the form of using pairs of 8-bit characters. For the JIS X 0208 characters, it meant adding 128 to the numerical value of each byte, which results in what is known as the "most significant bit" (MSB) of each byte in the character being set to "1." This distinguishes them from the regular alphabetic, numeric, etc., characters where this bit is set to "0." Half-width *katakana* (described above) was also encoded using two-byte sequences.
- Shift-JIS. This method was developed by several computer companies, in particular Microsoft, and it combined characters from both the JIS X 0201 and JIS X 0208 sets in a two-byte structure. To achieve this the numeric codes of the *kanji* and some other characters were "shifted" in value, which gave the method its name. A goal in the method was to assist with back-compatibility with older office equipment by having half-width *katakana* encoded as a single byte, as in JIS X 0201.

These encoding methods were used in parallel for several decades, and text-handling software often had to detect which method was being used, and to convert text from one method to another. Often the encodings would be damaged, particularly with the JIS Encoding approach, which led to the all-too-common problem of corrupted text, popularly known in Japanese as 文字化け *mojibake.*

MORE *KANJI*, CHANGED *KANJI*

What we now know as JIS X 0208 had barely begun being implemented when it became out-of-date. In 1981 the Ministry of Education replaced the 1,850 *tōyō kanji* with the 1,945 常用漢字 *jōyō kanji* (daily-use *kanji*), adding 95 *kanji*. Moreover, the number of 人名用漢字 was increased to 166, some of which were not in the standard. If that wasn't enough, the ministry-preferred forms of some of the *kanji* were changed. To accommodate these changes, four more *kanji* were added to the Level 2 *kanji*, 22 *kanji* were swapped between the levels, and the forms of a number of *kanji* were changed *in situ*. For example:

- 壺, which was in Level 1, was swapped with 壷, which had been in Level 2;
- 堯 was moved out of Level 1 to the end of Level 2 and replaced by the brand-new 尭 *kanji*;
- 啞 at the start of Level 1 had its form changed to 唖.

In 1990 a further revision of JIS X 0208 was released. For the *kanji*, the main changes were:

- the addition of two more *kanji*. This was to support the expansion of 人名用漢字, to which another 118 characters had been added. This brought the total number of *kanji* in the standard to 6,355;
- 225 *kanji* had their forms changed, mostly in subtle ways. A few reversed the changes made in 1983.

The other major development in 1990 was the release of a "supplementary" character set standard (JIS X 0212) which, along with additional alphabetic and special characters, added 5,801 *kanji*. As expected, these *kanji* were ones that were not in everyday use in Japan; however, it was interesting that some of the forms that had been dropped from the original version of JIS X 0208, such as 唖, made a return in the new standard. (Another returnee was 頬, which had been replaced in JIS X 0208–1983 by 頬.)

In fact, the additional *kanji* introduced in JIS X 0212 were not really available to most Japanese computer users. While the JIS and EUC encoding techniques were extended to include them, the Shift-JIS technique could not easily be modified and, thus, since it was the main encoding system in use, users of Windows PCs and most workstations had no access to the additional *kanji*.

THE RISE OF UNICODE

By the 1980s, it became apparent to many people involved with handling text in computer systems that having overlapping and conflicting codes for different national character set standards was a significant problem in need of resolution. It was also recognized that a major aspect of the problem was the coding of the large number of *hanzi* and *kanji*. In 1986 people in several computer companies began exploring the possible creation of a common coding system for all languages and scripts, and around the same time ISO began preparing a unified code standard. Even before that, a code-set that combined around thirteen thousand *hanzi* and *kanji* had been compiled in Taiwan, and eventually a snapshot of it was included in the ANSI Z39.64 standard in 1989 where it was termed EACC – East Asian Character Code.

The coding system being developed by the computer companies was given the name Unicode (from Universal Coded Character Set), and the companies formed the Unicode Consortium as a vehicle for its development. The initial focus was on using 16-bit (two-byte) codes in order to handle the large number of characters, especially the many *hanzi* and *kanji*. The early Unicode and ISO proposals were not compatible, the ISO draft being broader and more complex, however in 1991 agreement was reached to effectively merge the essential aspects of the proposals into a common standard, largely based on the Unicode approach.

The initial set of unified *hanzi*/*kanji*, which the Unicode Standard calls "ideographs," consisted of 20,902 characters. It was developed by taking the major character standards from China, Japan, Korea, and Taiwan (in the case of Japan, these were JIS X 0208 and JIS X 0212, and subjecting the aggregated 120,000 characters in them to what is termed the "Han Unification" process. The process, which is quite complex, can be summarized as follows:

- characters were classified according to their semantic values (i.e., basic meaning) and their abstract shapes.

- characters that had the same meaning and abstract shape were "unified," i.e., were treated as the same character. For example, the 三 character from the Chinese, Japanese, and Korean standards were treated as one character and given one code.
- characters sharing the same semantic value and abstract shape, but differing in (usually minor) typeface details, were also unified. Examples include character forms that varied only in their use of alternative versions of the *shinnyō* "movement" radical (辶, 辶). Alternatively, the 学 and 學 characters were not unified because, although the semantic values are the same, the abstract shapes differ.

Under what was termed the "Source Separation Rule," characters were not unified if they were separately coded in a source standard. For example, 剣, 剱, 釼, 劍, 劒 and 劔 — all of which mean "sword," share the Japanese reading *tsurugi*, and look very similar — were not candidates for unification as they are coded as separate characters in JIS X 0208.

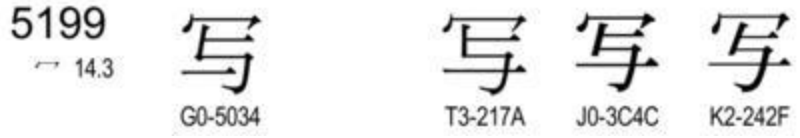
The first edition of the Unicode standard was published in 1991, and in 1992 the second volume came out, containing what became known as the "CJK" (Chinese, Japanese, Korean) codings. The matching ISO standard, ISO/IEC 10646, was first published in 1993, and a Japanese equivalent, JIS X 0221, was approved and published by JSA in 1995.

The extract below is from the code tables in Unicode 3.0, published in 2000. As can be seen, the tables indicate only the characters themselves and their Unicode code values.

Extract from the Unicode CJK tables

The Unicode Consortium has also compiled an extensive database, known as the "UniHan database" of information about each CJK character. The UniHan database contains for each CJK character such things as the readings in various languages, the broad meaning, and references to national standards, character dictionaries, etc. (See https://www.unicode.org/reports/tr38/)

Despite the fact that the early drive for a unified CJK coding system came in part from Japanese organizations such as the National Diet Library, and that Japanese people were involved in the unification process, there was an initial negative reaction to Unicode in Japan. This was, in part, due to the published standard using typically Chinese forms for many of the CJK characters. By contrast, the ISO and JIS printed standards contained representative national forms of the characters, as can be seen in the extract below from JIS X 0221, which depicts the typical Chinese (simplified and traditional), Japanese and Korean forms for 写 (Unicode character U+5199). The codes under each character refer to the source national standard. "G0-5034" is Unicode shorthand for "code 5034 in the (PRC) GB 2312 standard" and J0-3C4C similarly references code 3C4C in JIS X 0208.



Extract from JIS X 0221

The Unicode Consortium eventually adopted a similar approach of including representative national styles of the characters in its published standard. This began with Unicode Version 5.2 (2009).

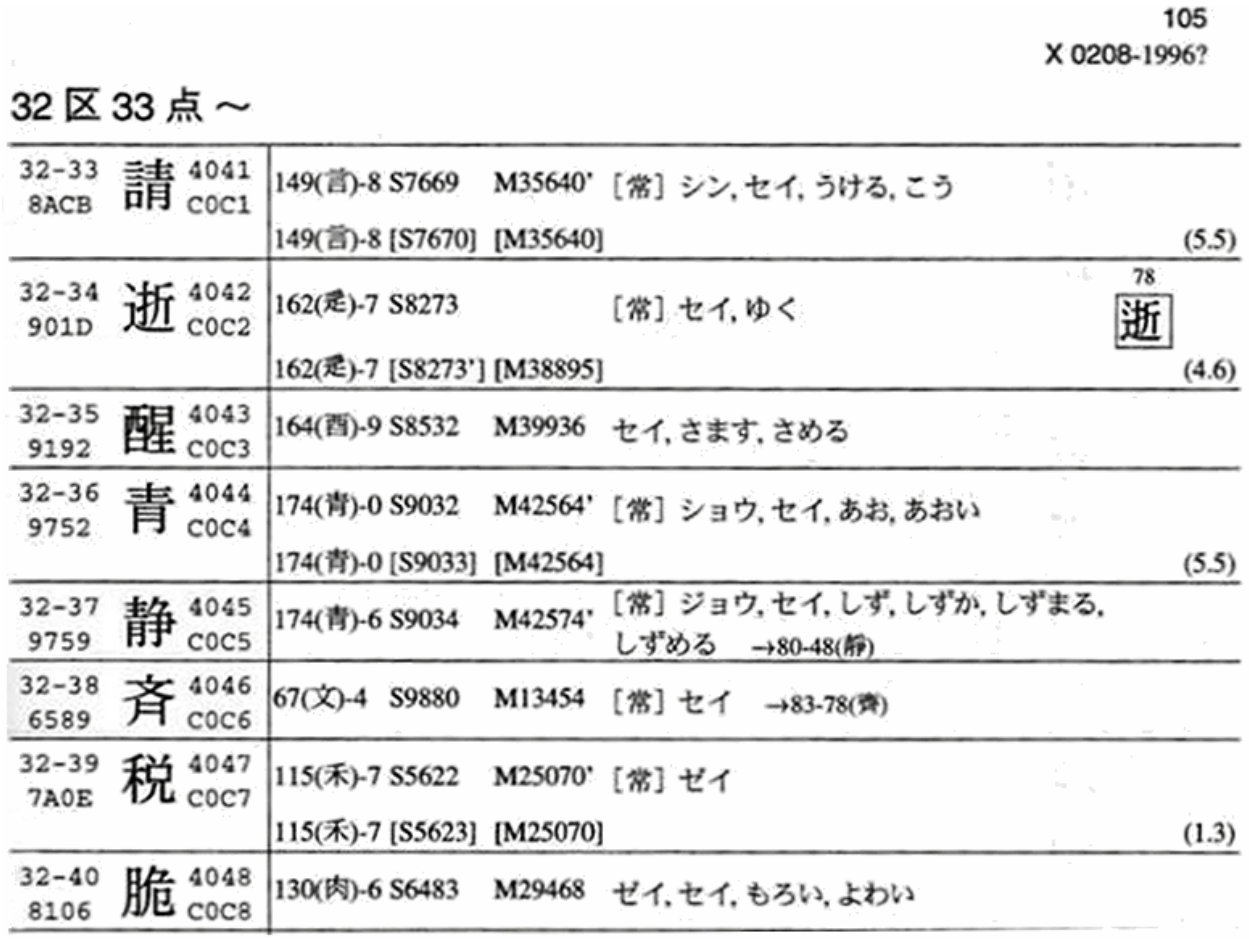
For a number of technical reasons, most of the numeric codes in the Unicode and ISO/IEC 10646 standards are not suitable for direct inclusion in text, use as file names, etc. For example, 恰 has been given the code of "6070" in Unicode. If these numbers were used directly in text, they would be treated as the ASCII characters "<" (60) and "F" (70). The exceptions are the basic alphabetic, numeric, punctuation, etc., characters that are compatible with ASCII. As with the encoding of characters in the JIS standards (discussed above), other Unicode character codes need to be converted into a compatible format. The encoding method that is most commonly used is UTF-8 (Unicode Transformation Format 8-bit), in which characters are encoded as sequences of two or more bytes. Most *kanji* are encoded as three- or four-byte sequences.

REVISED AND EXPANDED JIS STANDARDS

The JSA committee dealing with character coding, following the release of the 1990 version of JIS X 0208 and the new JIS X 0212, turned its attention to a thorough review of JIS X 0208. The goal was not to produce an expanded standard, but rather to resolve some outstanding issues with earlier versions, make the standard more usable in conjunction with others such as JIS X 0201, and provide more details on the principles behind the compilation, unification, character forms, etc. With all this additional information, the result, JIS X 0208:1997, was more than three hundred pages longer than the previous edition. (As a comparison, the last edition of Unicode to be printed as a book (5.0 in 2000) was over 1,400 pages, not including the CJK and *hangul* (Korean phonetic script) tables that were provided in a CD-ROM.)

The extract from the JIS X 0208 code table below illustrates the level of detail in the standard.

For example, the 逝 *kanji* has on the left the JIS coding in 区点 format (32-34) and the Unicode code (901D). In the central part it has the radical and stroke-count details, the reference numbers in the *Shinjigen* and *Daikanwajiten kanji* dictionaries (S8273 and M38895), an indication that it is a *jōyō kanji* [常] and the *on*- and k*un*-readings in *katakana* and *hiragana* respectively. It also shows the form of the *kanji* as it appeared in the 1978 version of the standard.



105
X 0208-1996?

32 区 33 点 ～

32-33 8ACB	請	4041 C0C1	149(言)-8 S7669	M35640'	[常] シン, セイ, うける, こう
			149(言)-8 [S7670]	[M35640]	(5.5)
32-34 901D	逝	4042 C0C2	162(辵)-7 S8273		[常] セイ, ゆく 78 逝
			162(辵)-7 [S8273']	[M38895]	(4.6)
32-35 9192	醒	4043 C0C3	164(酉)-9 S8532	M39936	セイ, さます, さめる
32-36 9752	青	4044 C0C4	174(青)-0 S9032	M42564'	[常] ショウ, セイ, あお, あおい
			174(青)-0 [S9033]	[M42564]	(5.5)
32-37 9759	静	4045 C0C5	174(青)-6 S9034	M42574'	[常] ジョウ, セイ, しず, しずか, しずまる, しずめる →80-48(静)
32-38 6589	斉	4046 C0C6	67(文)-4 S9880	M13454	[常] セイ →83-78(齊)
32-39 7A0E	税	4047 C0C7	115(禾)-7 S5622	M25070'	[常] ゼイ
			115(禾)-7 [S5623]	[M25070]	(1.3)
32-40 8106	脆	4048 C0C8	130(肉)-6 S6483	M29468	ゼイ, セイ, もろい, よわい

Extract from JIS X 0208 code table (1996 draft)

One task that was carried out by the JSA committee was reviewing and validating the sources of all the *kanji* included in the initial 1978 version. Questions had arisen about several of the *kanji* that had been included and that did not appear in any published Japanese or Chinese character dictionaries. These *kanji* had come to be known as "ghost characters" (幽霊文字 *yūreimoji*). The review encountered

problems with missing or incomplete source documentation from the initial compilation, but it was eventually able to confirm the sources of many of the *kanji*, although some may have been variant transcriptions of other *kanji*. Some of the anomalies that were discovered included:

- the 峹 *kanji*, where the central horizontal stroke appeared to be the result of a join or fold in a document containing the 岌 *kanji*.
- the 彁 *kanji*, where it was thought that a penciled annotation beside a 哥 *kanji* was interpreted as the 弓 radical.

In addition, the 1997 revision made it clear that it was not intended to define the precise forms of the characters. Some years before, JIS standards had been established for character forms, e.g., JIS X 9051 and JIS X 9052, but these had often not been followed in all details by the designers of modern fonts, mainly because they defined very coarse, by today's standards, bitmap images, 16×16 and 24×24, respectively.

Although the revision of JIS X 0208 did not add any *kanji* to that standard, there appeared to be a need to make more characters (both *kanji* and other characters) available to Japanese computer users. Many of these desired characters had been defined in JIS X 0212, but, as mentioned above, the inability of the dominant Shift-JIS encoding method to include those characters meant they were effectively not available to most users. In addition, the review of the sources for the previous standards had identified a number of *kanji* that had not been defined in either of them. For example, the *kanji* 朧, which can be found in the family name 集朧 (Shūki), was missing.

To address the perceived need to add more *kanji* than would fit into the structure of JIS X 0208, the committee chose to establish a new standard, JIS X 0213, by taking the existing JIS X 0208 standard and expanding it. The extra characters in JIS X 0213 include:

- a range of special, alphabetic, etc., characters. Many of these were also in JIS X 0212 but some, such as the International Phonetic Alphabet (IPA) characters and *katakana* extensions for writing the Ainu language, were new;
- 3,685 *kanji* (later expanded to 3,695). Of these, 2,743 were already in JIS X 0212 and hence in Unicode, and 649 of the other 952 had already been added to Unicode from non-Japanese

sources. The remaining 303 were new to Unicode and were added to Version 3.1 in 2001 as part of Extension B. The total number of *kanji* in the standard was 10,040 (later increased to 10,050).

Some of the additional characters were encoded using unoccupied places in the JIS X 0208 table, but 2,436 *kanji* were encoded by adding a second 94×94 table of coding places. (In coding standards these tables of coding places are often referred to as "planes." The earlier JIS coding standards, such as JIS X 0201 and JIS X 0212, each consisted of only one plane.) As the expanded set of characters could not be handled in the existing Shift-JIS encoding method, the standard also proposed a modified version of Shift-JIS capable of supporting the increased number of characters.

The first version of JIS X 0213 was released in 2000, and a revision was made in 2004 that added ten more *kanji* and made minor modifications to the printed forms of 168 others. (For example, the radical on the left of 辻, which had hitherto been the three-stroke 辶, was changed to the four-stroke 辶.)

The actual implementation of JIS X 0213 in computer systems as an alternative to JIS X 0208, which had been the character standard for decades, turned out to be largely a non-event. There appear to be two main reasons for this:

- The real demand for the extra characters was not that high, with the majority of Japanese users apparently satisfied with what was available via JIS X 0208. This probably led to some reluctance on the part of computer companies to invest in the expanded fonts and software upgrades required to support the new standard.
- By this time there was increased attention within the computer industry on Unicode. By the 2000s, major software packages were becoming internationalized, i.e., the goal was to have a single software version with configurable text modules for each supported language, rather than the earlier approach of language-specific versions. In this environment it made sense to use a single text coding method as far as possible, and Unicode was really the only choice.

In fact, the main lasting impact of the JIS X 0213 standard will probably be the additional 303 *kanji* it contributed to Unicode.

THE TRIUMPH OF UNICODE

As mentioned above, Unicode had its inception in moves by a number of computer companies to develop a common coding system. By 2000, with the release of Unicode 3.0, most of the major companies had committed to using Unicode for all their forward development. As companies such as Microsoft, Apple, Sun, etc., released new versions of operating systems, word processing packages, etc., they were increasingly built using Unicode as the basis for encoding text. Moreover, as new platforms were developed and released, such as Android and iOS for mobile devices, they invariably used Unicode from the beginning.

As a measure of the general acceptance of Unicode, a survey of Japanese web pages in 2020 indicated that more than 95% used Unicode/UTF-8 as their coding.

As with other national standards organizations, JSA has virtually ceased work on revising national character standards; all the activity associated with character standards now largely takes place within the framework of updates to the Unicode and ISO/IEC 10646 standards. The last updated edition of a JIS character standard was the 2004 revision of JIS X 0213; however, both JIS X 0208 and JIS X 0213 were reissued in 2012, mainly to add the expanded list of 2,136 *jōyō kanji* established in 2010.

The JIS character standards were important pioneering components of the computing fabric. They were the first to establish coding systems for the large major character sets used in East Asian countries, and they also were the first to deal with the issues of multi-byte character codes in computer text. Their destiny now is to be seen as part of the history of the uniform international text coding system, which is expected to endure for a very long time.

FURTHER READING

For further information on the history, structure, etc., of Japanese and other character standards, the following sources are recommended:

CJKV Information Processing (2nd Ed), Ken Lunde, O'Reilly Media, 2008

Dr Ken Lunde's monumental book (860 pages!) is the first place to look for information on this topic. Dr Lunde has spent virtually his whole working life dealing with CJKV characters within major computer companies, and he leads the Unicode Consortium's project team in the area.

JIS 漢字字典, Shibano Kōji (ed.), JSA, 2002 (*in Japanese*)

This is a character dictionary based on the 10,050 *kanji* plus other characters in the JIS X 0213 standard, but it also includes essential explanatory information from the standard itself and also from JIS X 0208-1997. Here you can read the details of the exploration of the "ghost characters." Professor Shibano was in IBM Japan for many years before he took up a professorship at the Tokyo University of Foreign Studies. He chaired the JSA committee which carried out the 1997 revision of JIS X 0208 and compiled JIS X 0213. His colleague Professor Masayuki Toyoshima was also on the JSA committee. (There was an earlier edition of the JIS 漢字字典 in 1997 that covered only the JIS X 0208 characters.)

The Unicode Consortium Standards and Website

The Unicode website at https://unicode.org/main.html is a goldmine of information about all aspects of the project and standards. Of particular interest are the sections on the CJK activities (https://www.unicode.org/consortium/cjkunihan.html) and the history of Unicode (https://www.unicode.org/history/).

Wikipedia Pages

The main JIS character standards have explanatory Wikipedia pages. The ones for JIS X 0212 (https://en.wikipedia.org/wiki/JIS_X_0212) and JIS X 0213 (https://en.wikipedia.org/wiki/JIS_X_0213) are relatively basic. The one for JIS X 0208 (https://en.wikipedia.org/wiki/JIS_X_0208), although structurally quite a mess, contains a lot of interesting and useful information about the standard and its development.

ACKNOWLEDGMENTS

The expert feedback and suggestions of Ken Lunde and Eve Kushner in the preparation of this article were most welcome.

All issues of *Sino-Platonic Papers* are accessible to readers at no charge via our website.

To see the complete catalog of *Sino-Platonic Papers*, visit

www.sino-platonic.org