Japanese Character Input: Its State and Problems

Ichiko MORITA: Ohio State University, Columbus.

Computer processing of information is highly advanced in Japan, and it continues to be researched and improved by the cooperative efforts of the government, private corporations, and individual scientists, who are among the best in the world. This paper introduces various approaches to the computer input of information currently developed in Japan, and discusses the possibility of their applications to the processing of East Asian–vernacular language materials in large research libraries in this country.

Processing of catalog information through an on-line shared-cataloging system has become a part of American libraries' common practice, and its financial and temporal savings have been proven. However, there are some materials not yet considered appropriate for computer processing. The Library of Congress' plans for romanizing catalog information for all non–roman language materials and putting them on MARC tapes for quick distribution of information have been objected to by a large number of specialists in the field. The opponents' reason has been that computerization of vernacular languages by means of transliteration is not satisfactory. Such materials are best handled in their own writing systems (the languages in this category include Chinese, Japanese, Korean, Hebrew, Arabic, and various languages in India). Those specialists in the field who see systems working for roman-alphabet materials generally agree that automated systems are very efficient and useful for their research. It would be best if non–roman language materials could be processed through computers using their own writing systems.

As far as technology goes, it is possible to process such materials in their original form. Systems that have the capability of handling those languages directly have been developed; among the most advanced are the Japanese systems. Japan has overcome numerous difficulties in developing systems that are capable of handling Japanese characters. Although automation of libraries is not as widespread as in the United States (due perhaps to a delay in the development of computers), some Japanese libraries have already a decade of experience with advanced

Manuscript received August 1980; accepted December 1980.

systems. Many others have recently started to adopt them. Wide utilization of these systems seems to be just a matter of time.

It will be beneficial to review Japanese methods and consider possible adaptation of them to our systems. In the following sections, various Japanese approaches to inputting the Japanese language are explained with an eye to future automation of non–roman language materials in this country.

THE JAPANESE LANGUAGE AND THE COMPUTER

It should be noted, first of all, that the Japanese language is an entirely different language from Chinese, although they are often confused because they both use the same Chinese ideographs in writing. Each Chinese ideograph, or character, symbolizes a certain object or denotes a certain meaning. The Japanese use them in the Japanese language with its own pronunciation in the context of its own grammar, whereas the Chinese use them in the Chinese language with its own pronunciation in the context of its own grammar. This means that a Chinese ideograph could mean the same thing in both languages, but be pronounced or read differently and used in different grammatical environments. The Chinese ideographs used in Japanese are referred to as *Kanji,* which are, to complicate the matter, used along with Japanese syllabaries called *Kana*. *Kana*, in two styles called *Hiragana* and *Katakana*, total about 170 characters. Depending on whether a *Kanji* is used with another *Kanji* or *Kana*, the reading of it varies. At different times one set of *Kanji* may be read in two or three different ways.

The total number of *Kanji* is about 50,000. In comprehensive dictionaries, about 40,000 or more *Kanji* are included. Medium-sized ones, such as Ueda's *Daijiten*, include about 15,000; concise ones about 8,000 to 10,000.[1] According to several tests on frequency of *Kanji* occurrence made in various Japanese institutions, approximately 3,000 *Kanji* appear in high frequency, 3,000 are of moderate frequency, and several thousand more are of infrequent occurrence. As for geographical names, 2,279 *Kanji* will cover most of Japan and 1,500 *Kanji* will suffice to cover personal names, except for very unusual names.[2] Approximately 6,300 characters are needed for major newspapers such as *The Asahi* and *The Nikkei*.

The trends in the use of *Kanji* are to simplify the characters themselves, and not to use difficult *Kanji* with many strokes. In 1946, the Japanese government established 1,850 *Kanji* as those for daily use,[3] and today newspapers and official documents use only those *Kanji,* except for some personal and geographical names. The implication of this trend for computerization of *Kanji* is that, depending on the documents to be covered, the need in number and kind of *Kanji* varies. That is, institutions that deal with scientific or current information do not need as many *Kanji* as other types of institutions that handle documents cover-ing longer periods and larger areas of knowledge. For example, Japan Information Center for Science and Technology, which mainly handles the latest scientific information, claims that with approximately 6,000 *Kanji* it can function satisfactorily. An example from the other extreme is the National Institute of Japanese Literature, whose collection covers older historical periods, during which a great number of *Kanji* were used and many *Kanji* went through changes, mostly simplification in style. The latter institute is constantly adding new *Kanji* to its system.

It is obvious then that the first problem in the computerization of Japanese materials is the number and kind of *Kanji* to be included in the system. This is a problem of hardware.

The other problem concerns software. When Japanese is written, its words are not divided as in English, for combination of *Kanji* and *Kana* helps visually to make sentences understandable without word division. Also, compound nouns are made by adding other words to a noun, so that, if a set of *Kanji* represents one noun, one can expand its meaning by adding another *Kanji* to it. Though word division has been a problem in transliteration and not new in computerization, both arbitrarily divided words and undivided words in particular become serious problems in the computer files and in the retrieval of information.

A question may be raised as to why we need *Kanji* processing in spite of these problems; why isn't computer handling of alphanumerics and *Kana*, which is in use today, sufficient? The answer to this is mainly that *Kanji* possess a definite visual effect. Also, if only romanized languages or *Kana* alone are used, many homonyms may make the meaning ambiguous. While it is quite possible to write Japanese only in *Kana* or in the romanized forms, as proven by the systems in use, it is better, for efficiency and precision, to express the language in the way it is actually written.

As for the problem of word division, study is in progress on methods of dividing words systematically and automatically, incorporating the latest research in the field of applied linguistics. This is more concerned with the development of software, and this paper will not delve into it.

INPUTTING

Various Japanese approaches to inputting *Kanji* and *Kana* are organized below into six major groupings according to different inputting devices. They are: (1) full keyboard, (2) component pattern input, (3) *Kana* keyboard, (4) stenotype, (5) optical character recognition, and (6) voice recognition. These six methods are further divided into subvariations as shown in table 1.[4]

Full keyboard

The main feature of this approach is use of a full character keyboard as the inputting device. The operator uses the full character keyboard

Table 1. Input Systems

Major Approaches	*Variations*	*Subvariations*	*Training Needed*	*Characters/ Minute*	*Characters Accommodated*
Full keyboard	Kanji teletypewriter		Medium–Extensive	40–100	2,300–4,000
	Japanese typewriter	Character location	Medium	30–50	2,205
		Coded-plate scanning			2,863
		Coded typeface			2,200–3,000
		Modified coded typeface			
	Tablet style	Electromagnetic	Medium–Small	30–70	3,000–4,096
		Electrostatic			2,800–4,000
		Photoelectric			2,800–4,000
Component pattern input					
Kana keyboard	Two-key stroke	Location correspondence	Extensive	60–120	4,096
			Association memory		
	Display selection		Small	20–30	
	Kana-Kanji conversion	Word conversion			
		Sentence conversion			
Stenotype					
Optical character recognition					1,000–2,500
Voice recognition					

rather than codes or other symbols. The keyboard varies depending on models, usually consisting of frequently used *Kanji* and both sets of *Kana*, supplemented by Arabic numerals, Roman, Cyrillic, and Greek alphabets in upper and lower cases, often with italics, signs, and diacritical marks. To each character, a two-byte binary code (expressed by a four-digit numeral) is assigned, so that when the inputter types a character the code for the character is punched on paper or cassette tape.

Kanji Teletypewriter

The oldest method for *Kanji* inputting, still widely in use, is the *Kanji* teletypewriter system or multishift system. One variation of this approach, developed by the National Diet Library at an early stage of its computerization, has 192 character keys, each having fourteen characters in three columns and five lines, as shown in figure 1. In addition, there are fourteen selection keys arranged in three columns and five rows on the lower left of the keyboard to correspond to the pattern of characters on each character key. When an operator strikes the character key *B* with the right hand and the selection key *A* with the left hand at the same time, the code for the character *C* is punched on the tape.

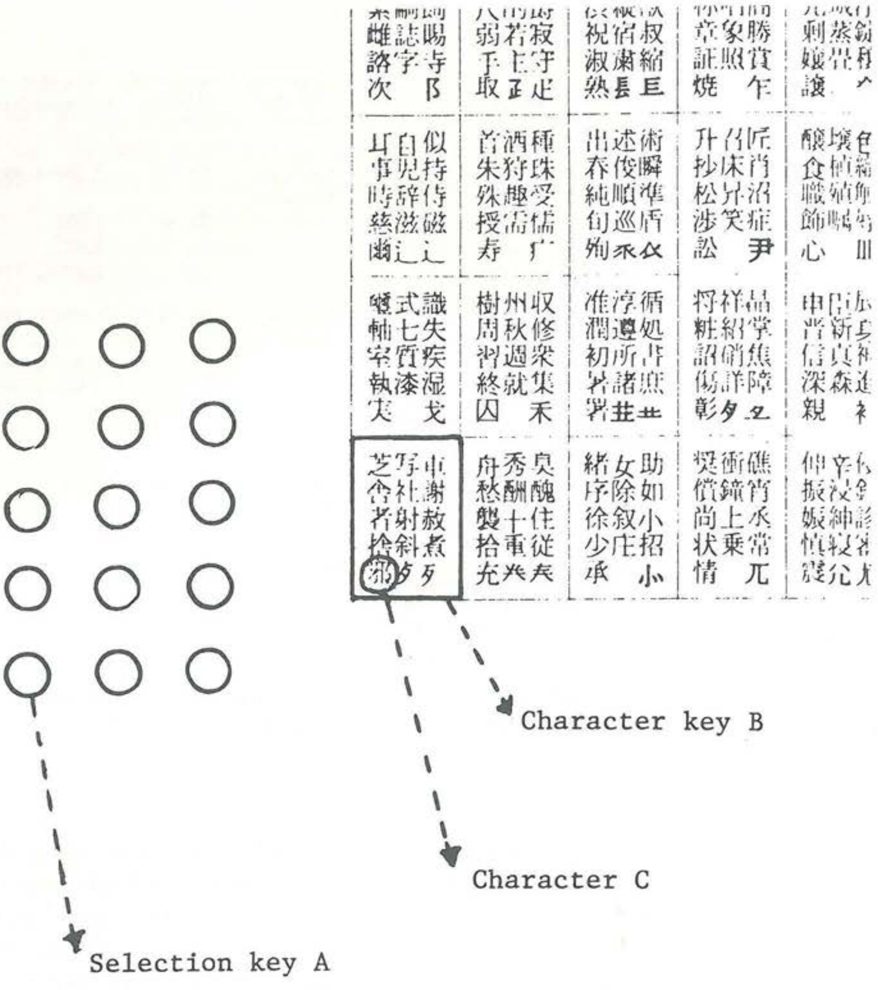


Fig. 1. Kanji Teletypewriter Keyboard of the National Diet Library.

Included on this keyboard are:

Kanji	2,006
Kana	90
Western alphabets	144
Numerals	20
Symbols and marks	210
Kanji pattern [5]	40
Kanji components	139
Space	1
Total	2,650[6]

By using shift keys on the upper left of the keyboard, *Kana* in both styles and alphabets in upper and lower cases can be input. For satisfactory operation, the keyers must be professionally trained, and it is said that one to three months are necessary for them to be fully trained and able to input an average of fifty to sixty *Kanji* per minute. This is not as fast as most other methods discussed.

Japanese Typewriter

The second of the full keyboard approaches is the Japanese typewriter method, which uses a modification of the standard Japanese typewriter with a tray filled with *Kanji* printing types. The operator finds a character in the tray and punches it by moving a metal handle as the type bar is punched down to print the character. This is rather primitive and different in its operation from the English typewriter, which uses the ten-finger touch method. There are four variations:

Character Location Method. Kanji are arranged on a keyboard by their codes, so that when a key is punched, the *Kanji* is typed on regular paper as if it had been done by a regular Japanese typewriter. At the same time, the code is automatically read from the location of the key and is punched on tape.

Code-plate Scanning Method. Each type bar has a plate attached on its side, and the code for the character is marked on its plate. When a key is typed, the *Kanji* is printed on paper and the code from the plate is optically scanned at the same time.

Coded Typeface Method. Each typeface is made with a character on the upper half and a code for it on the lower half. When a key is typed, both the character and code are printed. The code on the bottom half is optically scanned from the printed paper.

Modified Coded Typeface Method. Instead of typing both characters and codes on the paper, this method prints only the characters on the front of the paper and, at the same time, prints a bar code on the back of the paper. The machine capable of doing this is complicated. The size of the character on a typeface can be bigger than in the variation above, and the bar code can be larger to make the scanning of the code easier and more precise.

As the discussion of the four variations indicates, the Japanese typewriter offers the advantage of being able to monitor input at the time of keying.

Since the Japanese typewriter has been in use for a long time in offices where a quantity of official documents are dealt with, and since ordinary Japanese typists can use this system without any additional training, the use of equipment similar in operation was considered advantageous. However, it should be noted that Japanese typewriters have never become as prevalent as English typewriters, and the demand for computers comes from more areas than just those where Japanese typewriters are used. For this reason, the use of Japanese typewriters is not as advantageous as its proponents claim. An obvious disadvantage is its slow speed of operation—thirty to fifty characters per minute on the average. Another disadvantage is that the number of characters on the keyboard is limited to about 3,000.

Tablet Style

This method, also known as pen-touch method, was recently developed. Each character has a key, and characters are arranged in a certain order. The location of the characters on a matrix sheet determines the two-byte binary code, which consists of a two-digit numerical abscissa and two-digit numerical ordinate. The operator touches the key with a pen-shaped detector and the code for the character is punched on the paper tape. The operation is one-handed, requiring only a light touch of the key by a detector. Keys are on one flat keyboard and are color-coded by sections to make it easier for the operator to locate them. Light touch operation reduces operator fatigue. This method does not require special training. However, the number of *Kanji* on a keyboard of reasonable size is limited to approximately 3,500. By shifting, twice as many characters can be handled, though all characters are not indicated on the keyboard. Speed of input is not very high—thirty to seventy characters per minute. This system, already used in many libraries, is becoming increasingly popular because of its easy operation. There are three different technologies used: electromagnetic, electrostatic, and photoelectric. There are no differences in actual input operation for those electronically different methods.

Component Pattern Input

Although not a full keyboard method, component pattern input is closely related to these methods.

The idea behind this approach is that most *Kanji* are composed of one or more basic component units, two or more of which can be put together into one *Kanji* according to one predetermined pattern out of forty general patterns. The inputting device has keys for those forty patterns along with keys for individual components on a special keyboard. To compose a *Kanji*, a key for an appropriate pattern is selected and typed, and components are chosen to fill each individually numbered block of the selected pattern, following the established order as shown below.[7] Each pattern has a code, and so does each component. When a key is typed, the code is punched on a paper tape as shown in figure 2. There are cases where a *Kanji* with two components can be a component of another *Kanji*, as shown in the first and second examples in figure 2. A *Kanji* is constructed by punching at least three codes: one for a pattern and at least two for components. Then, a *Kanji* dictionary consisting of several thousand master-code combinations (see figure 3) is stored in a magnetic drum, and the several codes to compose a *Kanji* punched on paper or cassette tapes are converted through this diction-

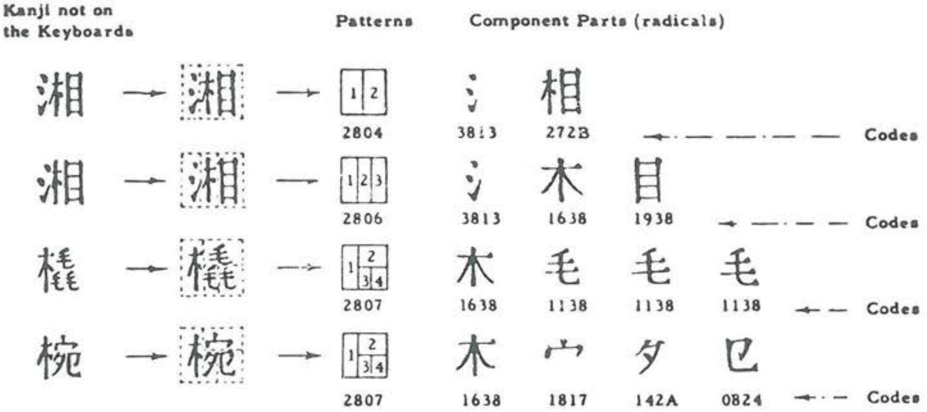
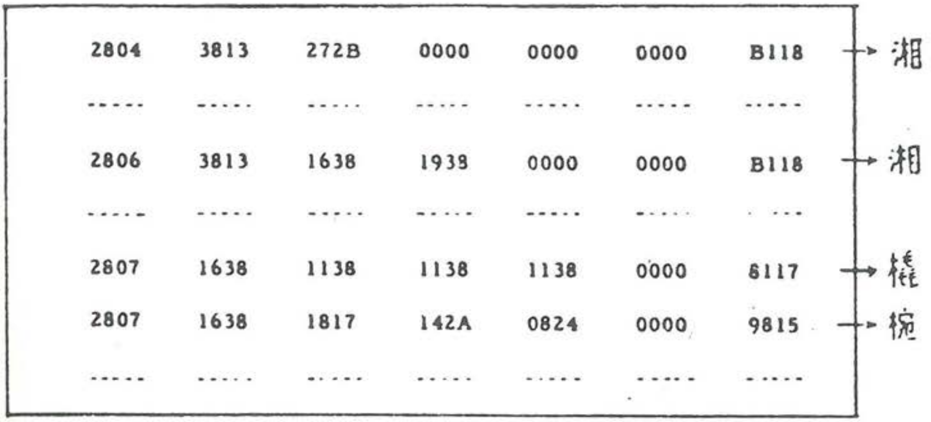


Fig. 2. Component Pattern Input.



2804	3813	272B	0000	0000	0000	B118	→ 湘
-----	-----	-----	-----	-----	-----	-----	
2806	3813	1638	1938	0000	0000	B118	→ 湘
-----	-----	-----	-----	-----	-----	-----	
2807	1638	1138	1138	1138	0000	8117	→ 橇
2807	1638	1817	142A	0824	0000	9815	→ 椀
-----	-----	-----	-----	-----	-----	-----	

Fig. 3. Kanji Dictionary.

ary to a two-byte binary code assigned to that particular *Kanji*. These are then handled as other *Kanji* with an individual code.

Though this can be a stand-alone approach to inputting *Kanji*, the principle has been adopted by the National Diet Library to supplement the inputting of *Kanji* on the full keyboard *Kanji* teletypewriter. The National Diet Library uses this system when inputting *Kanji* that are not included in its keyboard. Instead of having a special separate keyboard, the *Kanji* teletypewriter of the National Diet Library integrates patterns and components as equivalents to other characters. Its keyboard includes forty patterns and approximately 140 components.

This was the most elementary approach to computerize *Kanji*. Conceived in the early developmental stage of *Kanji* processing, it used one of the characteristics of *Kanji*, the composition from several components. In actual situations, this technique requires at least three key strokes for one *Kanji* and consumes time to locate the needed component on the keyboard. Furthermore, it requires the complicated extra step of putting input codes through a *Kanji* dictionary to combine component codes into a code per *Kanji*. No library is currently using this system by itself.

Kana Keyboard System

The keyboard of a Japanese syllabary typewriter has adapted the conventional English typewriter keyboard and has standard roman alphabet keys that contain *Katakana* in shift (figure 4). Since the number of *Katakana* exceeds that of roman letters, the *Katakana* keys are extended to keys for numerals and punctuation marks. This means that this typewriter can be used either for *Kana* or roman letters by changing its mode.

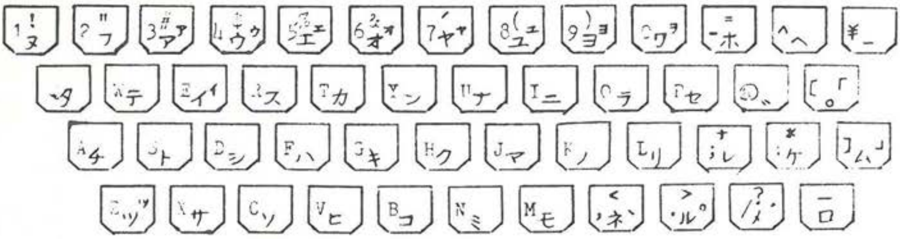


Fig. 4. Kana Typewriter Keyboard.

Two-key Stroke Method

This variation of the *Kana* keyboard system is referred to as the two-key stroke system, and uses *Kana* as codes not as letters. Roman letters can be used as codes, too. There are two different subvariations. They are:

Location Correspondence. Keys are divided into two sections: one for right hand, and the other for left hand. If two keys are to be stroked, there will be four possible combinations of key strokes: (1) left hand twice, (2) left and right, (3) right and left, and (4) right twice. The keyboard is accompanied by a *Kanji* table in which characters are arranged in several blocks and in a certain order within each block. Each block, which contains twenty-six *Kanji* in a four-by-six arrangement, is made according to each combination of strokes: first block is left and left; second block is left, right, etc. Within each block, the ordinate consists of keys for the first stroke and the abscissa for the second. A *Kanji* which is at the intersection of the above indicates which keys are to be typed. When *Kanji A* is to be typed (see figure 5), since it is in block *A* indicating the stroke combination as left and left, the operator types *A* and *W* by left hand. If *Kanji B* is to be typed, the operator types key *A* by left hand and key *P* by right. Each key has a byte code and a combination of two key strokes makes a composite, a two-byte binary code, for a *Kanji*. The bit may be changed by shifting, and different *Kanji* can

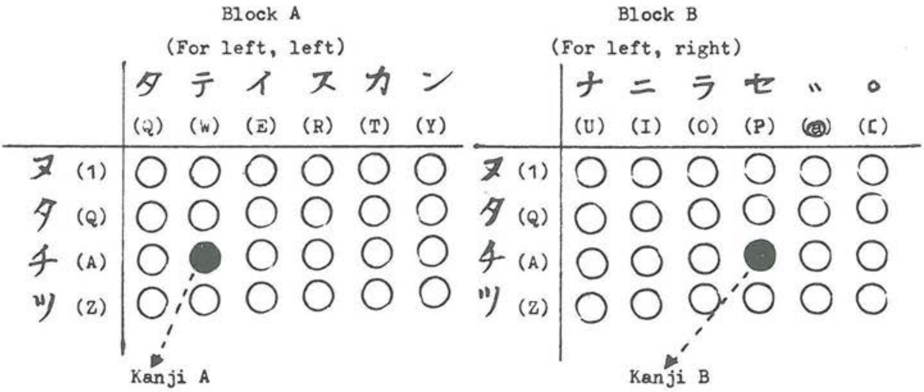


Fig. 5. Kanji Table for Location Correspondence Method.

be typed if another table is prepared for *Kanji* with different bits.

Association Memory Method. In this method, each *Kanji* is given two *Kana* which usually represent a reading of that *Kanji.* The operator associates a *Kanji* to be input with two *Kana* assigned to that *Kanji*, and types them with two strokes using the *Kana* keys.

Both of the key-stroke methods are economical as well as convenient because of the wide availability of *Kana* typewriters. Mainly for that reason, both of these systems have been well accepted and are expected to grow further. Since this touch method does not require the operator to look for the character on the keyboard to input, it is the fastest to operate and is considered suitable for input in quantity. It is possible to input 60 to 120 characters per minute. The only drawback is that the operator must get acquainted with the arrangement of *Kanji* in the first variation, and must memorize all the associated *Kana* spelling for many *Kanji* in case of the second variation. In either case, the operator must be professionally trained.

The Japan Information Center for Science and Technology, which indexes many scientific publications, employs a vendor who uses the location correspondence variation of this system for inputting information.

Display Selection

This also uses a *Kana* typewriter with a screen in front. When a word is typed in *Kana*, a group of *Kanji* with that sound are displayed on the screen. The operator chooses the right *Kanji* with a light pen—a slow but accurate operation. The operator does not have to be specially trained for this.

Kana-Kanji Conversion

In contrast to the conventional approach of full keyboard inputting, an entirely new method for inputting *Kanji* is gaining popularity as the availability of sophisticated software increases. This uses a *Kana* typewriter keyboard to input Japanese in syllabary or romanized form, converting them to *Kanji* by software. There are two ways of conversion: one that converts word by word, and the other sentence by sentence.

Stenotype

The stenotype is a typewriterlike device. The operator must be able to take shorthand. When the stenotype is used, it punches words in paper tapes. Therefore, inputting is high speed. However, the operator must receive proper training.

Optical Character Recognition

This system, developing quickly and expected to gain wider use, can scan a maximum of 2,500 printed *Kanji*.[8] One variation connects a writing tablet to a computer so that as the operator writes *Kanji* on the tablet, the computer scans them in stroke order. This function of scanning by the stroke order is considered to be an advantage for processing some types of Japanese documents. The drawbacks are that the system is still very expensive, and the number of recognizable characters is fewer than 2,000.

Voice Recognition

This is an oral-visual system, in which the human voice is read by a computer. Obviously the most difficult to develop, this system is still in an experimental stage. However, a prototype has been demonstrated at various exhibitions, and the system apparently possesses great potential.

Summary

Pattern configuration and output devices for Japanese characters are basically the same as those for English. However, the pattern generation of characters is mechanically more complicated than that of the roman alphabet, because *Kanji* has a more complicated structure than the roman alphabet and the number of components is greater. Each *Kanji* is represented by a two-byte binary code rather than one byte as in roman alphabet. Because of this, the efficiency of retrieval is low. Presently, hard copy and typesetting for printing of hard copy are the major output forms, and very little on-line retrieval of information with *Kanji* is in current operation.

PROBLEMS PARTICULAR TO KANJI PROCESSING

Among numerous problems in processing *Kanji* through computers, major ones are: (1) which *Kanji* are to be included; (2) how many characters are to be handled; (3) what code should be assigned and how it should be arranged on the keyboard or table; and (4) how the *Kanji* not included on the keyboard should be treated.

In the early stage of *Kanji* computer development, different institutions handled the problems in ways best suited to their individual needs, according to the nature of the literature covered, the amount of literature processed, and the kinds of output needed. They experimented with the then best available capabilities. As a result, the finished systems are all independent and mutually incompatible. Standardization is obviously necessary for exchange of information among the systems.

In order to set standards for selection of characters and assignment of codes, JIS (Japan Industrial Standard) C6226-1978 has been compiled by the Japan Association for Development of Information Processing. This is a table of characters designed for information exchange (a portion of which is shown in figure 6). It has a one-byte code as its abscissa and another as its ordinate. Characters are arranged so that the intersection of abscissa and ordinate determines a *Kanji* whose code consists of four numerals, two from the abscissa and two from the ordinate. Included in the table are *Kana* in both styles, Roman, Greek, and Cyrillic alphabets in upper and lower cases, diacritical marks, numerals, and punctuation marks, as follows:

1. Special characters	108
2. Numerals (Arabic)	10
3. Roman alphabets	52
4. *Hiragana*	83
5. *Katakana*	86
6. Greek alphabets	48
7. Cyrillic alphabets	66
8. *Kanji*	6,349
Total	6,802[9]

In the first section of the table, numerals, alphabets, *Kana*, and special characters are grouped. In the second section, the total of 2,965 frequently used *Kanji* are arranged as the first priority group, and an additional 3,384 *Kanji* are selected as the second group[10] in the bottom half of the table. *Kanji* are printed in the preferred style for printing typeface. This table will resolve problems 1 to 3 mentioned above. Institutions that had arranged their own codes for *Kanji*, including the National Institute of Japanese Literature, are now automatically translating their own codes into JIS codes.

In cases where needed *Kanji* are not included on the keyboard, handling varies. With the Japanese typewriter, because each *Kanji* is inscribed on a typeface, only the *Kanji* on that typeface is printed when the type bar is stroked. Therefore, only *Kanji* that have typefaces can be input in this system, while some other handling is possible in other methods.

While the number of characters that can be accommodated on keyboards is limited to 2,000 to 3,500, depending on the type of equip-

							第2バイト	b7	0	0	0	0	0	0	0	0	0	0	0	0	0
								b6	1	1	1	1	1	1	1	1	1	1	1	1	1
								b5	0	0	0	0	0	0	0	0	0	0	0	0	0
								b4	0	0	0	0	0	0	0	1	1	1	1	1	1
								b3	0	0	0	1	1	1	1	0	0	0	0	1	1
								b2	0	1	1	0	0	1	1	0	0	1	1	0	0
								b1	1	0	1	0	1	0	1	0	1	0	1	0	1
第1バイト								点	1	2	3	4	5	6	7	8	9	10	11	12	13
b7	b6	b5	b4	b3	b2	b1	区														
0	1	0	0	0	0	1	1		SP	、	。	，	．	・	：	；	？	！	゛	゜	´
0	1	0	0	0	1	0	2		◆	□	■	△	▲	▽	▼	※	〒	→	←	↑	↓
0	1	0	0	0	1	1	3														
0	1	0	0	1	0	0	4		ぁ	あ	ぃ	い	ぅ	う	ぇ	え	ぉ	お	か	が	き
0	1	0	0	1	0	1	5		ァ	ア	ィ	イ	ゥ	ウ	ェ	エ	ォ	オ	カ	ガ	キ
0	1	0	0	1	1	0	6		Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν
0	1	0	0	1	1	1	7		А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	Л
0	1	0	1	0	0	0	8														
0	1	0	1	0	0	1	9														
0	1	0	1	0	1	0	10														
0	1	0	1	0	1	1	11														
0	1	0	1	1	0	0	12														
0	1	0	1	1	0	1	13														
0	1	0	1	1	1	0	14														
0	1	0	1	1	1	1	15														
0	1	1	0	0	0	0	16		亜	唖	娃	阿	哀	愛	挨	姶	逢	葵	茜	穐	悪
0	1	1	0	0	0	1	17		院	陰	隠	韻	吋	右	宇	烏	羽	迂	雨	卯	鵜
0	1	1	0	0	1	0	18		押	旺	横	欧	殴	王	翁	襖	鴬	鴎	黄	岡	沖
0	1	1	0	0	1	1	19		魁	晦	械	海	灰	界	皆	絵	芥	蟹	開	階	貝
0	1	1	0	1	0	0	20		粥	刈	苅	瓦	乾	侃	冠	寒	刋	勘	勧	巻	喚
0	1	1	0	1	0	1	21		機	帰	毅	気	汽	畿	祈	季	稀	紀	徽	規	記

Fig. 6. Code of the Japanese Graphic Character Set for Information Interchange.

ment, character generators have the capability of outputting more than the number of characters on the keyboard. Figure 7 shows their relationship. Characters that are in the generator but not on the keyboard must be frequently processed, because the number of characters needed for most documents could reach 6,000 to 6,500. Using a shift key to enter another mode is a fairly common technique for inputting uncommon *Kanji*. The keyboard may not have a character but, if the character generator has it, the code for that character can be input by shifting. For example, if a character on the keyboard has a code 0117, a bit is changed so the code 8117 can be typed by shifting and typing that key. If the code 8117 is assigned to another *Kanji* not on the keyboard but indexed in the dictionary, it can be input. This applies for the *Kanji* teletypewriter, tablet style, and the two-key stroke variations of the *Kana* typewriter.

In the *Kanji* teletypewriter system used by the National Diet Library, the keyboard accommodates 2,650 characters, while its character generator has the capability for 5,717. Operators in the National Diet Library input *Kanji* that are not on the keyboard by using component pattern input method. Or, if the operator finds the *Kanji* code in the specially compiled dictionary in which codes for *Kanji* are indexed, a shift key is used to change the bit, thus creating the code for *Kanji* not on the keyboard. Most other tablet systems use code dictionaries. In the two-key stroke variations of *Kana* typewriters, tables of *Kanji* for second and third or more shifts can be built, especially when the location association method is used.

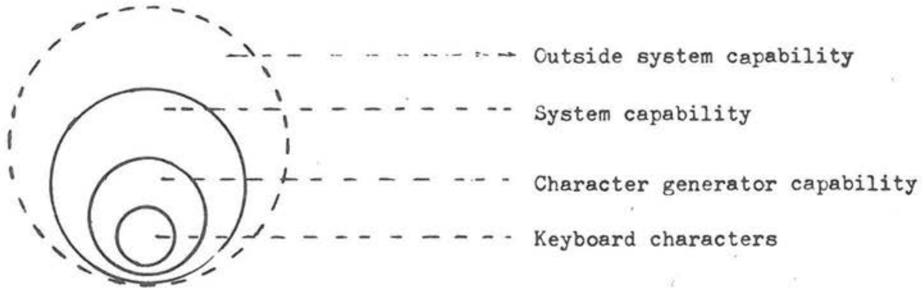


Fig. 7. Kanji Creating Capability.

The handling of *Kanji* that are not in character generators is more difficult. Only the digital character generator, the kind that uses either dot or stroke, can add characters fairly easily. In the flying spot system, characters can be added, but it must be done professionally with an additional character cylinder and is very costly. The National Diet Library, which now uses flying spot, limits addition of *Kanji* to a minimum. Because its output is solely in printed book form, the National Diet Library inputs a fill character for *Kanji* not in the system. When the phototypeset masters are made, the fill characters are replaced by typeset characters. The use of a fill character suffices only when the output is phototypeset, because there is a step to replace fill characters by typeface. However, as long as the data base includes many fill characters on the magnetic tapes, the on-line retrieval of information or later utilization of tapes becomes unsatisfactory.

The National Institute of Japanese Literature uses a dot matrix and prints by wiredot impact. If a *Kanji* is not in the character generator, the institute's staff composes the *Kanji* in an enlarged dot matrix and creates the capability for printing in the generator. If the *Kanji* made in such a way is used only once, the *Kanji* pattern is not stored in the character generator, so that the generator does not reach its full capacity quickly. The enlarged dot composite for *Kanji* created in the institute is filed and indexed for future use.

Most other institutions simply do not use those less commonly used *Kanji*, and substitute *Kana* for them.

In addition to the problems common to any character output, such as size and number of dots, the problem of the space for *Kanji* in relation to other characters and the choice of vertical or horizontal printing of Japanese sentences with *Kanji* must be considered.

Kanji have many strokes and, as mentioned before, are expressed by two-byte codes. Each *Kanji* needs a double space when displayed on screens or printed. When a *Kanji* is used with numerals or *Kana*, the *Kanji* part looks fine but the numerical part has too much space between each numeral. Therefore, input of *Kanji* is done in a *Kanji* mode and input of *Kana*, roman alphabets, and numerals are in a *Kana*-numerical mode. In this way a multidigit figure looks like one whole figure rather than a line of one-digit figures.

Some formal documents must be printed in the traditional vertical arrangement. To cope with this situation, some line printers have the capability to precompose a vertical page before printing it.

There are multicolor CRTs on the market that can be used for the retrieval of library-related information, e.g., main entry in red, series statement in yellow.

One last problem that must be considered is that most of these systems require trained operators, or else the operation is very slow. The information is edited and compiled by the editors and prepared for input in the form of worksheets. So are the revisions. At various stages of revising the text, the information must be printed, given to the editors, and revised. Further developments in simplifying input and revising texts for efficient flow are to be expected.

APPLICATION OF KANJI SYSTEMS

Processing of vernacular-language materials in their own writing systems is considered vital for research libraries in this country. In adopt-ing the *Kanji* systems in such libraries, there are three major factors that must be considered: the objectives and needs of the institution, the cost, and the personnel.

First, the institution must know what it must accomplish by means of such a system. The needs may not be the same for all institutions. Is the system for retrieving catalog information, or for inputting catalog and other information? Is it for internal processing or patron use? Is it for a large bibliographic utility to distribute information to its subscribers, or for an individual institution to process its own information? Could the system be shared by the department of Asian studies in any way? The character set needs of the institution are a major factor in choosing the system.

Since input and output devices are different, i.e., one cannot input *Kanji* on a CRT and retrieve *Kanji* from the same CRT, the institution must consider how much it will need to input, or whether it can rely on available data bases. Some institutions may not need any input equipment if they utilize available data bases. If Japan MARC and other tapes are made accessible by a large bibliographic utility in this country, the institutions will be able to obtain bibliographic information in *Kanji* on the screen. If they want only catalog cards or a COM catalog, they will not need any equipment except the terminals supported by the utility. If they want to input, they must consider what form or forms of output they need, how to create the characters not included in the system, in addition to which system to choose.

Second, cost is an important factor. Is the expense justified in terms of the other needs of the library? What can be accomplished per dollar spent? The *Kanji* systems are still expensive, though the cost will eventually be reduced. How much can be spent and how much continuing support can be expected are factors that modify system expectations. The budget must include not only the one-time hardware cost, but also the software, maintenance, and personnel.

Third, the availability of personnel will affect the choice of system. What degree of language expertise does the system require in each stage of operation, such as inputting, maintenance, and programming? Does it need terminal operators trained in those languages? What other personnel does the system need as far as language-related qualification is concerned?

Apart from the three major factors discussed above, there are some technical aspects that must be adjusted to library situations in this country. Since Japanese, Chinese, and Korean use the same Chinese ideographs to different degrees and in different ways, libraries considering automated processing of these language materials are probably expected to handle all three languages by the same system, to say nothing about the other non-roman scripts. Problems will arise in selecting characters for inclusion in the system. As pointed out earlier with regard to

Japanese character processing, there are simply too many characters for the present capacity of any computer. If Korean and Chinese languages are to be handled by the same computer, this problem multiplies. The Korean alphabet, called *Hangul*, would have to be included. Chinese has more characters than Japanese. Worse yet is the fact that some *Kanji* are simplified in different ways in Japan and China, so that they are neither recognizable nor interchangeable between them. It will be an enormous task to accommodate both in the same system.

Another problem is the arrangement and indexing of *Kanji*. If a full keyboard, a Japanese typewriter keyboard, or two-key stroke system, especially its location association method by *Kana* typewriter, is considered for Japanese, Chinese, and Korean, the arrangement of the characters must be indexed and accessed for the three languages, in addition to the multiple readings found in Japanese. For example, *Kanji* on the Japanese keyboard are usually arranged by the initial sound of the Japanese reading of the *Kanji*. This arrangement will be useless for Chinese and Korean, because Japanese readings are not the same as Chinese or Korean readings. The arrangement of *Kanji* on the keyboards must be on some new principle common to these languages.

Even if the *Kana-Kanji* conversion is used, and roman alphabet–*Kanji* conversion software is adopted, software to handle those three languages must be developed. Such software would have to be highly sophisticated. The presence of many homonyms in Chinese will cause a great problem to the extent that the system relies on transliterated or romanized forms of the language. Recognition of the many identical spellings in different language contexts will be extremely difficult.

The above discussion is based on what is currently available in Japan. The combination of existing inputting, generating, and outputting equipment developed by Japanese technology opens up various possibilities for us to build effective systems in this country.

ACKNOWLEDGMENT

This article is based on a study conducted in Japan as a Japan Foundation professional fellow, and as a visiting research fellow of the Center for Research on Information and Library Science, University of Tokyo.

REFERENCES

1. National Institute of Japanese Literature, *Implementation of a Computer System and a Kanji Handling System at NIJL* (Tokyo: NIJL, 1978), p.16.
2. Toshio Ishiwata, "Kanji Shori Kenkyū ni Motomerareru Mono" ["Requirements for Study on Kanji Processing"] *Computopia* no.9 (1977), p.35.
3. *Gendai Yōgo no Kiso Chishiki, 1980 [Basic Knowledge on Current Terms, 1980]* (Tokyo: Jiyūkokuminsha, 1980), p.999.
4. Figures are taken from the following two sources and compiled by the author: Hasegawa, Jitsurō. "Kanji Shori Sōchi" ["Kanji Processing Devices"] *Jōhō Shori [Information Processing]* 19, no.4:353 (April 1978).

Sugai, Kazurō. "Kanji Nyū-shutsuryoku Sōchi mo Kaihatsu Dōkō" ["A Trend in Development of Kanji Input-Output Devices"] *Business Communication* 16, no.7:41 (1979).

5. Used for the pattern input mentioned in the following component pattern input system.
6. National Diet Library, *Library Automation in the National Diet Library* (Tokyo: The Library, 1979), p.4.
7. Ibid., p.7.
8. Asia Business Consultants is using an optical character recognition system that can scan handwritten *Kana* and numerals in a small scale to input and process catalog information for a library collection.
9. "Jōhō Kōkan no Tame no Kanji Fugō no Hyōjunka" ["Standarization of Kanji Code for Information Interchange"] *Kagaku Gijitsu Bunken Sābisu [Scientific and Technical Documents Service]* no.50 (1978), p.29.
10. Ibid., p.28.

Ichiko Morita is assistant professor in library administration and head, Automated Processing Division, the Ohio State University Libraries.

EDITOR'S NOTES

Most *JOLA* readers are aware of significant delays in publication in the last volume. Susan K. Martin, a former editor of *JOLA*, and Richard D. Johnson, a former editor of *College & Research Libraries*, gave freely of their time and energy to bring the journal back on schedule. Mary Madden, Judith Schmidt, and the members of the Editorial Board under the leadership of Charles Husbands all worked closely with Sue and Richard in this effort. This was a second time around for Sue, who undertook a similar task when she assumed the *JOLA* editorship in 1972. The *JOLA* readership and this editor owe debts of gratitude to Sue, Richard, and all the others who helped.

We do not foresee major changes in the format of the journal as established principally under the editorships of Kilgour and Martin. We look for increased strength in our Book Reviews section under the editorship of David Weisbrod. The addition of Tom Harnish as assistant editor for Video Technologies indicates our recognition of the growing importance of video-based information systems.

We encourage reader suggestions. We welcome brief communications of successes or failures that might be of interest to other readers. Letters to the editor about any of our feature articles or communications are solicited.