KANJI Input for Data Processors

KATSUHISA SAITO*, ETSUO NAGAI*, KOZO YAMADA*
AND AKIKAZU OSAKO*

Abstract

Recently in the field of information retrieval, there has been a strong trend toward computer processing using the Japanese language. One of the main problems in Japanese information processing is the method of feeding Kanji (Chinese character) used in Japanese into the machine. Since the Japanese language employs a large number of ideographic Kanji, a simple and practical machine like the English typewriter can not be used.

This paper describes a practical method of encoding Kanji using a Kanji display system. The encoding, a practical method based on pronunciation.

1. *Introduction*

The systems used in writing Western languages are very different from those used in Oriental languages. In this age of computers, methods of conversation between man and machine also differ greatly in the West and in the East. The alphabets used in Europe and the America consist of only a few characters which can easily be fed into a machine using a keyboard resembling that of a typewriter. In this way, all Western languages can be fed into a machine and the words of that language can be used in the machine. In Japan, Western words have been used in the interface between man and machine ever since computers were first developed in this country. The Japanese language which includes a large number of Kanji characters brought from China was not used. However, with the recent advances in computer technology, computers have been introduced for large scale processing of the Japanese language and information retrieval represented in the Japanese language. The main problem in these information processing systems using the Japanese language with all of its complex Kanji is how to load several thousand different characters into a machine quickly and easily. This article describes a practical Kanji input method which uses a character display system.

2. *Kanji encoding*

There are three main methods for Kanji encoding: the method according to the shape of the Kanji, the method according to the pronunciation of the

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 5 (1969), pp. 294–298.
* Oki Electric Industry Co., Ltd.

Kanji and the method of matching the Kanji with English letters and numbers. Almost all investigations up to now have dealt with the first method which also can be divided into two types: methods concerned with the actual shape of the characters and those in which the elements making up the character are analyzed. Many studies have been made but there have been no definite conclusions, which indicates the complexity of the Kanji. In this article, the authors will describe a practical method of Kanji encoding which utilizes classification of Kanji according to pronunciation and a Kanji display system.

The problems when selecting Kanji according to pronunciation is that there are so many different Kanji in Japanese which have the same pronunciation (hereafter referred to as homonym Kanji). However, with this method, operation is simple and the number of steps needed are fewer since Japanese pronunciation is familiar to Japanese rather than stroke order and component parts of the Kanji. Also, "Kana" (Japanese syllabry) or roman letter keyboard can be employed. There are two ways to read Kanji: the "on" reading based on old Chinese sounds and the "Kun" reading based on native Japanese words applied to the Kanji. It is almost impossible to determine only one character from the pronunciation because in the case "on" reading, many different characters have the same pronunciation. Also in "Kun" reading, the characters which were decided merely from the pronunciation are less than 30%. There is therefore a problem in the handling of these different characters with the same pronunciation. These homonym Kanji which could not be readily determined were shown on a Kanji display and a method of selecting these Kanji form this display has been investigated. When the homonym Kanji in the 1,850 general use and 28 supplementary Kanji were investigated, results as shown in Table 1 were obtained. However, characters with two or more pronunciations are double count (taken from the "Kadokawa Kokugo Jiten", a Kanji dictionary). The column headed both "on" and "kun" reading is added number which "on" and "kun" reading are same pronunciation. In the example of "ア" (a) there are "on" reading (亜) and "kun" readings (合, 会, 飽, 明, 上, 揚……).

In the case of "on" reading only, there is a total of 286 types of pronunciation and the number is less than "kun" reading. However, maximum number of homonym Kanji is 61, and the selection of homonym Kanji is more difficult than "kun" reading. As 40% of all Kanji have no "kun" reading, only "kun" reading is not practical. Also it is inconvenient for Japanese to use only "on" reading in this method. When considering both "on" and "kun" readings, there are over 900 pronunciation, the number of homonym Kanji increases and there are many characters to be selected from. As can be seen from Table 2, there are only about 30 Kanji which require 4 or more kana to represent their pronunciation and these are limited to certain "kun" readings. These can thus

Table 1. Number of Homonym Kanji

No. of Homonym Kanji	No. of "on" Reading	No. of "kun" Reading	Both "on" "kun" Reading	No. of Homonym Kanji	No. of "on" Reading	No. of "kun" Reading	Both "on" "kun" Reading
1	56	514	546	21	4	1	3
2	47	81	118	22	1		1
3	31	24	54	23	1		1
4	20	15	27	24	4		4
5	22	9	19	25	0		1
6	12	5	15	26	0		0
7	16	2	22	27	1		2
8	12	2	15	28	2		2
9	12	2	13	29	1		1
10	2	5	4	30	2		4
11	9	2	9	31	1		1
12	6	1	6	35	1		0
13	4	1	8	63	1		1
14	4	0	4	43	1		2
15	1	2	3	46	1		0
16	0	0	3	51	0		1
17	0	0	1	52	0		1
18	1	1	0	61	2		2
19	4	0	2				
20	4	0	5	Total	286	667	901

Table 2. Kana Representation of Kanji Pronunciation.

No. of Kana	No. of Kanji
1	59
2	620
3	195
4	28
5	3

be disregarded and 3 kana or less can be used for Kanji representation.

An example of a Kanji selection system based on pronunciation is shown in Fig. 1. Read only memory 2 stores a code of all Kanji included in the display in order of pronunciation and therefore the homonym Kanji code is in a continuous address. Read only memory 1 on the other hand stores the kana code according to pronunciation and the lead address of the read only memory 2 which contains the homonym Kanji code for these pronunciations.

When the Kanji to be selected is punched on the kana keyboard according to pronunciation, this code is set in the kana code register and coincidence with register 1 is obtained by reading out read only memory 1 in sequence. The contents of register 2 when concidence is achieved are set in the address

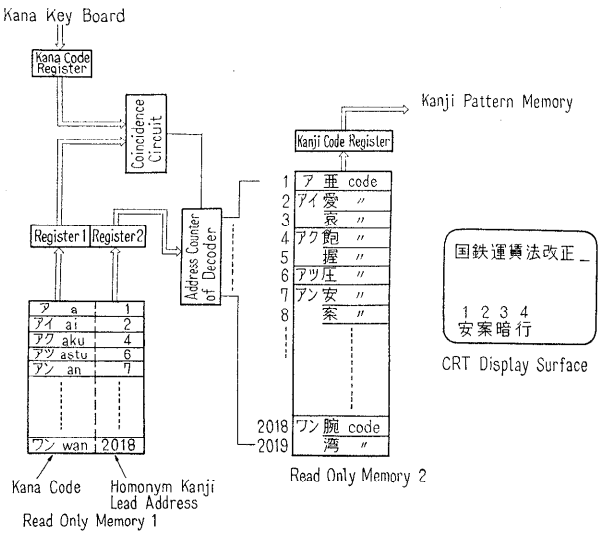

Fig. 1. Kanji Decoder by Pronunciation.

counter of read only memory 2 and all the Kanji with the same pronunciation are displayed on the CRT.

When there are no homonym Kanji, determination is possible immediately but otherwise selection must be made. For example, in the CRT display shown in Fig. 1, the Kanji "案" is desired as input. Wnen アン (an) is punched on the kana keyboard, 4 different Kanji with the pronunciation "an" are displayed on the lower part of the CRT. From these, the desired Kanji "案" is selected.

In the hardware of this Kanji coding system, the method of selecting displayed homonym Kanji also presents a problem in addition to the decorder using the above described read only memory. When the input is from a kana keyboard, it is not natural to select the displayed Kanji using a light pen; a method using input of corresponding number is better. If the selection of homonym Kanji is emphasized, the light pen method is desirable. In such a case, operation is easier if a part of the display is used instead of the kana keyboard and input operation is done by only the CRT. Then the CRT is divided into 2 parts, one for Kanji input and one for information display. It is convenient that the input CRT is placed horisontally.

Next there is the problem of the number of Kanji selected from the CRT. The percentages of reading numbers with up to 20 homonym Kanji where selection is relatively easy are 65% in the case of "on" readings only and 72% in the case of "on" and "kun" reading combined. Readings with more than 20 homonyms are handled in a special way. They are either displayed by using several parallel lines of characters arranged in accordance with frequency of use or by dividing them into several display frames. Another method is to reduce

the homonym Kanji to be displayed for selection. For this we use main component types of Kanji, in addition to the pronunciation. For example, by using about 20 main component types of common sense, 61 homonym Kanji of "コウ" (ko) and "ショウ" (sho) as shown in Table 1 can be reduced to about 25. If it is possible to use voice input with single "kana" sounds in the future, Kanji input using this system will become practical in conjunction with high speed CRT display.

3. *Conclusion*

A practical methods of encoding Kanji is discussed. Because of the complexity of Kanji, the coding system are not simple either to use pronunciation or shape like phonetic characters. From the point of view which it is difficult to simplify such a system only with machines, a practical method emplying a display system was investigated.

A method was described in which a set as small as posible, including Kanji to be selected, was made by using the pronunciations and components of Kanji and from this set operator made selection.

This article is based on a compilation of development and research work performed as a part of a large scale project sponsored by the Agency for Industrial Science and Technology of the Ministry of International Trade and Industry.