KANA TO KANJI-AND-KANA CONVERSION SYSTEM
(Information Processing of Japanese Language)

Yutaka Matsushita*, Haruaki Yamazaki* and Fumikazu Sato*

1. Introduction

There is an increasing demand for computer processing of Kanji (Chinese character) information by the press, publishers, personnel information management and other related circles. As computer systems gained increasing acceptance, work slips, books and direct mail printed in Kana (a form of Japanese letters) or Romanized Japanese came to be used so much that they posed problems. In view of the fact that it would be difficult to remove Kanji from Japanese language information and that removal of it would present difficulties, the importance of Kanji as information conveyance media was recognized again, and various Japanese-language data processing systems and Kanji input-output devices were developed.

Under such circumstances, the writers began studying systems primarily desgined for translation of correspondence a few years ago. This report introduces the Kana to Kanji-and-Kana Conversion System (hereinafter called the system) which was developed by the writers at the request of the Information Processing Promotion Association (IPA) in 1972. A system capable of Kanji translation of correspondence was reported by Fujii in 1971, but our system is quite different from his in that it is not dependent on a specific kind of business, and that the universal use of software is regarded as especially important. Section 2 deals with problems concerning how Japanese-language information can be processed by the computer system; Section 3 discusses the process of conversion, particularly the conversion of homonyms; and finally Section 4 evaluates the processing of homonyms.

This paper first appeared in Japanese in Joho－Shori (Journal of the Information Processing Society of Japan), Vol. 15, No. 1 (1974), pp. 2～9.

* Oki Electric Industry Co., Ltd.

2. Problems concerning the processing of Japanese-language information

2.1 Homonyms

Japanese words in Kana do not always correspond to words in Kanji. In other words, a single word in Kana generally represents a plurality of words that sound the same but mean different (therefore Kanji expressions are different). Distinguishing each of these words from one another is the greatest difficulty that lies in converting Kana words into Kanji words by machine.

2.2 Isolating Kana words for conversion to Kanji

Written Japanese words are not separated one from another by spaces. Thus, in converting them, it is necessary to isolate Kana words that can be converted into Kanji words. However, ambiguity is unavoidable in the process.

2.3 Stored glossary

As is clear from a comparison of the sports and political columns of newspapers, for example, the usage of words is quite different in various fields. Japanese words are so numerous that it is next to impossible to store all of them in a conversion dictionary.

Accordingly, many conversion systems had to be dependent on a specific kind of field. Therefore, this system has two dictionaries - a common dictionary which stores frequently used common words that are not technical, and a private dictionary that stores proper nouns, technical terms and abbreviations and other special words used exclusively in special fields - for effective use of them. In selecting common words, we referred to glossary studies by the National language Research Institute. The technical dictionary has an improved utility group for dictionary maintenance to deal with the frequent deletions and additions anticipated.

3. Conversion system

Generally, there are two ways of converting Kanji, one converting character (Kanji character) by character and the other by words. The former uses small-capacity dictionaries and is simple in processing so that it is used for Kanji input terminal equipment, etc.

The latter, on the other hand, requires large dictionaries and is complicated, but is expected to be effective in sharply reducing the wrong selection of Kanji words having the same sound, but different meanings, which is the greatest problem for Kanji conversion systems.

3.1 The conversion process of this sytem

This system is based on the word-by-word conversion process to lessen the burden on the system user and to increase conversion efficiency. Homonyms are rated according to the frequency of their use, and their ambiguity is reduced by analyzing declensional Kana and by affix processing.

An outline of the homonym processing by the system is described below. The system selects grammatical categories to which homonyms belong, and if homonyms belong to the same category, the one which has the higher frequency of use is selected for output. The categories roughly conform to the classification of Japanese grammar. Particularly, nouns are subdivided as they are so numerous.

3.2 Forms of succeeding letters

Generally, Kanji words as appearing in sentences have suffixes that may take any of the following three forms.

1) Suffixed by Kanji

   Example: Sato-shi (Mr. Sato)

2) Suffixed by Kana

   Example: A-karui (bright, underlined part is Kanji.)

3) Suffixed by Alphabet, numeral, or other special letters.

   Example: 300m (300 is Kanji part.)

Item (3) includes cases where sentences end with Kanji words.

This system can quite easily identify the forms of suffixes. Input data into the system is always provided with control codes to specify Kanji, Katakana (a form of Kana), or Alphabet. Specifically, that part which is bracketed by the control code " <" is to be converted into Kanji; that part which is bracketed by control code " ((", into Katakana; that part which is bracketed by the control code " ' " into Alphabet; and that part which is bracketed by no control code, into

Hiragana (a form of Kana), numeral or special symbol.

Fig. 1 shows the theoretical tree illustrating the process of gramatical category selection.

3.3 Process of analyzing declensional Kana

In case of Katakana or Hiragana suffixes, the system traces the grammatical tree and finds the grammatical category.

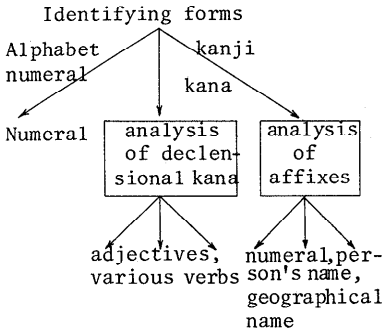

Fig.1 The character category decision process

Greater connection		
↑	Numeral	Suffix to numeral
	Numeral	Numeral
	Person's name	Suffix to person's name
	Person's name	Person's name
	Geographical name	Suffix to geographical name
	Geographical name	Geographical name
	Common noun	Suffix to common noun

Fig.2 Contents of a 'Setuji' table

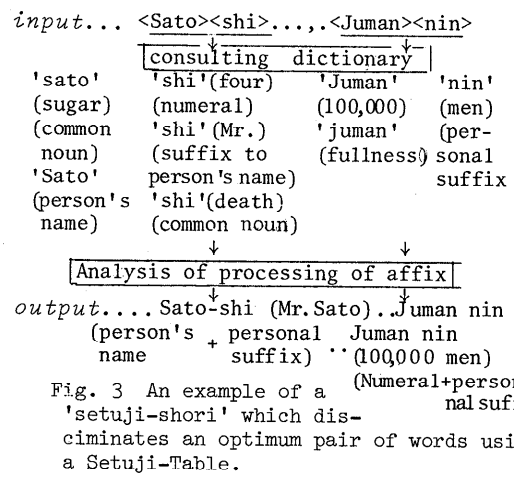

Fig. 3 An example of a 'setuji-shori' which discriminates an optimum pair of words using a Setuji-Table.

3.4 Processing suffixes

This system processes nouns (numericals, personal names, geographical names, etc.) and affixes in their connecting relationship. Fig.2 shows a table of affixes representing connection relationships which are processed in preference to others. For example, when an input is fed into the system, it selects the most suitable pair in reference to the table of affixes. If there are 2 or more pairs in the table of affixes, however, the one in the higher place, that is, the one which is considered to have the greater connection, is selected in preference to the others.

4. Evaluation

The effect of converting common sentences by this system is considered attributive to the general effectiveness of processing various homonyms as described in the foregoing pages. To evaluate the effectiveness of processing homonyms, a news article was picked out at random from a

newspaper which was considered to use the most standard Japanese language, and was processed by the system. Of the 360 sample Kanji, 87 % were correctly converted.

Then the relationship between the number of homonyms and correct conversion rate and the relationship between the number of homonyms and the product of the correct conversion and reference rates were evaluated as shown in Table 1, Fig. 4 and Fig. 5.

Only common dictionary was used, and non-convertible words (whose corresponding Kanji words are not stored in the dictionaries) were omitted. That the correct conversion rate of 100% was not obtained for a sample having a single homonym (that is, having only 1 corresponding Kanji word) is due to the fact that there are in fact homonyms which are special words that are rarely, if ever, used and are not included in common dictionaries and as a result the wrong conversion was made.

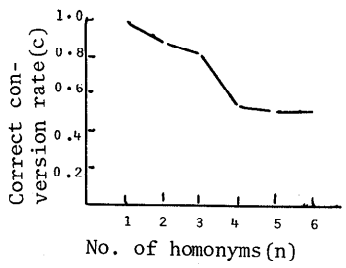

Fig. 4 No. of homonyms vs rate of correct conversion

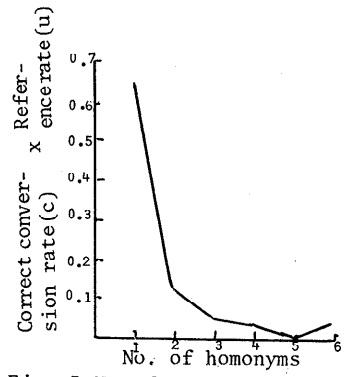

Fig. 5 No. of homonyms vs the product of correct conversion and reference rates

Table 1 Conversion program efficiency

No. of homonyms (n)	Reference rate (u)	Correct conversion rate (c)	[u x c]
1	64.8%	99%	0.642
2	14.5%	87%	0.126
3	6.5%	80%	0.052
4	7.1%	52%	0.037
5	0.7%	50%	0.004
6 and over	6.4%	50%	0.032

No. of samples: N
No. of references of homonyms: dn
No. of correct conversions: (u)

$$u = \frac{dn}{N}$$

$$c = \frac{tn}{dn}$$

5. Conclusion

The system was developed as a software system for general use on the

basis of the primary research conducted at Kyushu University up to 1966 and subsequent research conducted by our research staff based on it. This program is unique in that it does not depend on specific kinds of business or fields. Although the system now uses only common dictionaries, a higher level of conversion efficiency can be expected by employing technical dictionaries for specific fields. One common dictionary stores about 15,000 commonly used words. The definition of common words is not yet certain, and there may be common words which are not yet included in common dictionary. Thsese problems pose the greatest difficulties in developing this system.

6. Acknowledgement

In concluding this report, we wish to express heartfelt thanks to the late Professor Kurihara of the Engineering Department of Kyushu University for his helpful instructions and guidance, to Mr. Kurosaki of Oki Electric for his assistance and encouragement, to the people of the Information Processing Promotion Association, and others for their cooperation in our efforts to develope this Kana-to-Kanji conversion system.