STARS: WORD PROCESSING FOR THE JAPANESE LANGUAGE

Editor's note: *This month we continue a series of reprints from the IEEE Global History Network's STARS articles.*[1] *STARS is an online compendium of invited, peer-reviewed articles on the history of major developments in electrical and computer science and technology. Some light editing has been done to make the article more suitable for a journal publication.*

The Japanese word processor was not merely an adaptation of an existing system, but a highly innovative system realized by entirely new software and hardware that revolutionized communications and information processing in Japan.

In the English language, any text of any length can be written using only 26 different letters of the alphabet, because any English word can be spelled by some combination of those letters. In practice, users want to have each of the 26 letters in two forms, capitals and lower case, and also want to have the Arabic numerals of 0 to 9, as well as punctuation marks and various special symbols. Thus, a typical keyboard for the English language has 44 keys, which can produce 88 characters by means of a shift function that permits each key to produce two characters. This many characters are sufficient for most applications.

I. CHALLENGES OF THE JAPANESE LANGUAGE

By contrast, the written Japanese language uses three character sets combined: a large number of kanji, or Chinese characters, plus two sets of native symbols collectively called kana. Today, both sets of kana and thousands of different kanji are used in daily life.

The Japanese began adopting the Chinese characters to record their native language in the 5th century CE. They used the Chinese ideographic characters as symbols to represent objects and concepts, though pronouncing them with Japanese sounds for words. They added new kanji and new meanings. The Japanese also used characters purely to represent native sounds, including those having grammatical significance in a manner similar to word endings in English such as -ed, -ing, and -ly. This latter group of characters was developed into the two sets of kana, called hiragana and katakana, and comprised 48 basic symbols each. This development has been attributed to the Buddhist priest, Kukai, in the 9th century. Katakana are traditionally used to spell foreign words. The present sets of kana were codified by the 1900 Elementary School Order and their orthography standardized in 1946 by the Cabinet announcement on Gendai Kanazukai, or Modern Kana Usage.

Developing an input system for written Japanese proved difficult for obvious reasons. Along with kana, namely the two phonetically identical sets of 48 symbols, a great many kanji have to be dealt with. Besides, each kanji comprises many strokes. This makes it far more complex than spelling an English word using the 26 Roman letters. If the strokes in each kanji are not clearly demarcated, the thousands of characters cannot be distinguished from each other. This complexity contributes to the beauty so often seen in handwritten Japanese text. But this beauty comes at a price. Starting in the early 1900s, home and office typewriters were in common use for English, French, German, and other languages, in which words are spelled by an alphabet. But there was no similar machine that could handle the Japanese language.

Even though one fifth of all people in the world use ideographic languages as their mother tongue, only alphanumeric characters were available on computers until the 1980s. Some people before then insisted that Japanese should stop using kanji and kana and instead use the letters of the Roman alphabet to spell Japanese words. Against this background, newspaper companies wanted an input and output device that could

[1]Please refer to the STARS website for additional information and to view the full article: http://www.ieeeghn.org/wiki/index.php/STARS:Word_Processing_for_the_Japanese_Language.

Digital Object Identifier: 10.1109/JPROC.2013.2295876

0018-9219 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Timeline

1954	Tele-typesetting system is developed for Chinese characters
1967	Toshihiko Kurihara develops the fundamental kana-to-kanji conversion technology
1971	Toshiba starts R&D of the Japanese-language word processor
1972	Akira and Tadasi Kawakami present the two-stroke method for Chinese characters input
1977	Prototypes of Japanese-language word processors presented by Toshiba and Sharp
1978	Toshiba exhibits the Japanese-language word processor JW-10
1979	Major Japanese electronics companies market Japanese-language word processors
1980	Fujitsu markets word processor with non-JIS thumb-shift keyboard
1982	Desk-top Japanese-language word processor JW-1 released by Toshiba
1985	Laptop Japanese-language word processor JW-R10 released by Toshiba
1985	Kana-to-kanji conversion software is available on personal computers
1989	27 percent of Japanese homes possess one Japanese-language word processor
1999	Japanese word processor market is minimized by personal computers
2010	Over 100 million mobile phones provide kana-to-kanji conversion technology

handle over 6000 Chinese ideograms. The professional typists of Japanese typewriters had to locate a target character from that total, and even they needed much time to input newspaper manuscripts.

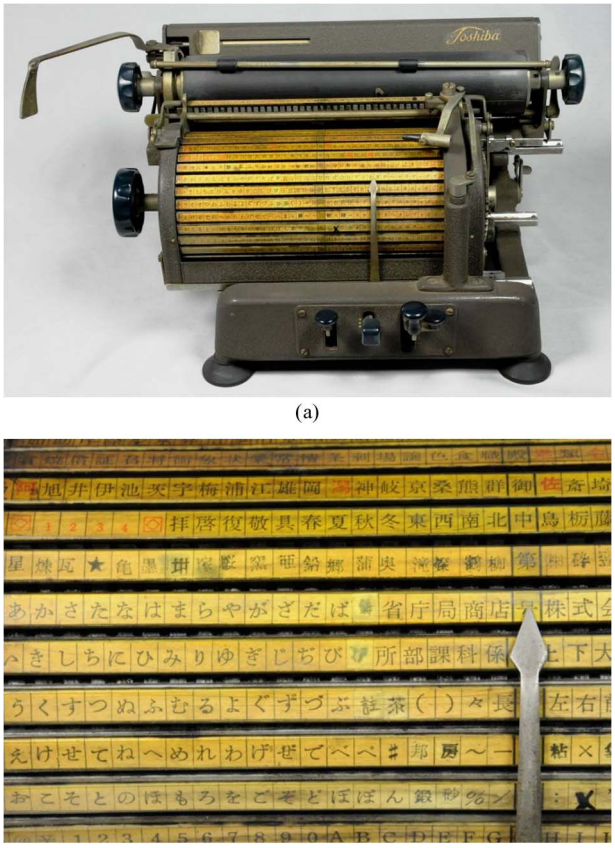
II. EARLY INPUT METHODS FOR CHINESE CHARACTERS

The expansion of Japan's industries and government in the early 20th century increased the demand for mechanized writing. Academic researchers and corporate engineers responded by innovating five different character-input techniques before more sophisticated electronic word-processing systems were introduced to the market in the 1970s. In their efforts to devise input methods for a writing system with three distinct character sets, the central problem to be addressed was how to select and incorporate the appropriate kanji from a vast library of such characters.

The first successful input method was Kyota Sugimoto's typewriter, patented between 1915 and 1929 (Japan Patent 27 877; U.S. Patent 1 245 633), and manufactured by Otani Typewriter Company. Sugimoto organized the 2400 kanji used most frequently in government publications on a matrix, from which a typist selected the desired character and applied it to the document curled around the cylinder. It was expensive and required patience to operate, but variations of the design were made and used through the 1970s.

The second input method was the drum method by Yusaku Shinozawa in 1918 (U.S. Patent 1 297 020). In this approach, the list of Chinese characters is curled around a drum, as shown in Fig. 1. Users rotated the hand wheel with their left hand, moved the needle to the desired position, and pushed the key with their right hand to select a character. Oki Electric Industry and Toshiba commercialized kanji drum typewriters in the 1950s.

The third input method made a keyboard practical by using multistage shift technology. In 1954, Yomiuri Shimbun and the Defense Agency developed a kanji teletype. In 1958, Asahi Shimbun and the Shinko



(a)

(b)

Fig. 1. *Toshiba's Japanese character typewriter (courtesy of Toshiba Corporation).*

MechatroTech Co., Ltd., Tokyo, Japan, commercialized the kanji teletype. The kanji teletype adopted the multistage shift method in which 12 letters were assigned to each key. The kanji teletype handled 2304 letters with a keyboard of 24 × 8 keys. Users selected the letter position key from 12 positions with their left hand and selected the character families with their right hand. A skilled operator could input 70–100 letters per minute. Newspaper companies used the kanji teletype for tele-typesetting system (TTS) through which stringers sent news articles to the central office.

The fourth input method was character alignment. It updated the Sugimoto typewriter by using a pen-touch tablet. The letters, a few millimeters in size, are arranged in a 60 × 35 matrix on a tablet. Users touched the letter on the tablet with a pen for selection and the tablet generated the character code electrically. This method was easier than the multistage shift method, but the input rate was 30–70 letters per minute, which was much slower. There are several types of code generation, including electromagnetic coupling by Hitachi and Toshiba; capacitive coupling by Pentel, Oki, and Fujitsu; and photoelectric coupling by Mitsubishi. Later, a piezoelectric-sheet method and a pressure-sensitive touch-sheet method were developed.

The fifth was the two-stroke method. In this case, a writer inputs the letter by two strokes, based on a character encoding scheme that assigns two alphabetic characters to each kanji and kana character. In 1972, Akira and Tadasi Kawakami of Rainputto Company first presented the two-stroke method. Their keyboard was unique in arrangement in order to achieve right-and-left hand keying, in which typing activity was divided relatively equally among all fingers. All characters are represented as an associated combination of two kana characters. For example, 品 is composed from three kana characters of 口. So 品 is represented as 口 (ro) and 三 (mi) which means three. Skilled operators memorized these two-kana character combinations and could input 100–125 characters per minute. An enthusiastic group strongly advocated the two-stroke method, but it fell into decline because of the greater ease of learning and the good performance ultimately achieved through kana-to-kanji conversion.

The difficulties of these five input methods for nonprofessional users supported the growing belief of some that Japan should give up using kanji. But most Japanese people wanted to input Japanese text with kanji and kana as easily as other people input their text using an alphabet.

III. SOLVING THE PROBLEMS

Although the mix of kanji and kana adds complexity to a text, it also significantly and ironically facilitated the creation of Japanese word processors. First, the presence of kana provided a set of characters familiar to educated Japanese that could be used to spell the sounds of kanji as a means of input. Second, it facilitated the automatic segmentation of inputted text by word processors or computer software.

The first person who undertook to develop kana-to-kanji conversion technology was Toshihiko Kurihara of Kyushu University. He developed the fundamental kana-to-kanji conversion technique in 1967 and proposed the kana sentence segmentation method, the word dictionary matching method, and other fundamental techniques as described in Japanese patent Sho-42-3241.

The first big problem of kana-to-kanji conversion was achieving automatic-word segmentation. There are spaces between words in English written sentences, but there are no spaces to express the pause between words or word groups in Japanese written sentences. A word group is called a bunsetsu. A bunsetsu corresponds to an English phrasal unit such as a noun phrase or a prepositional phrase. Since Japanese syntactic analysis is usually done after bunsetsu identification, identifying a bunsetsu is important for analyzing Japanese sentences. The world's first word-processor product for the Japanese language, Toshiba's JW-10, supported a bunsetsu-input method based on automatic word segmentation.

Researchers had been implementing kana-to-kanji conversion software on a mainframe computer experimentally, in a batch processing mode, but Toshiba developed an interactive system, which worked promptly even on a computer with very small memory compared to that of today's computers. The user input text using kana keys on the keyboard and pressed the "translation key" at the end of each bunsetsu.

A dictionary for the kana-to-kanji conversion needed to be developed because no digitized dictionaries were available at that time. Japanese language dictionaries available on the market contained only basic words or difficult words. Moreover, they did not contain many words frequently used in business documents or letters, e.g., proper nouns, such as names of people, companies, and places. These words were naturally necessary when making ordinary documents. By collecting words from high school textbooks, various standard guidebooks, business documents, and newspaper glossaries, a total of 62 000 words were selected and manually put into the dictionary for the JW-10. The morphological analysis and proper noun analysis programs each needed its own dictionary to analyze the given kana phrase group. Each dictionary has a certain set of kana entries with its kanji string, grammatical properties, and frequencies of use of the word. In these programs, the dictionaries are consulted to avoid ungrammatical expressions. Using the dictionaries, less than satisfactory conversion accuracy was initially achieved, but it was continually improved over time.

The second big problem of kana-to-kanji conversion is the homonym. A Japanese word, which is normally represented by one or more kanji, can

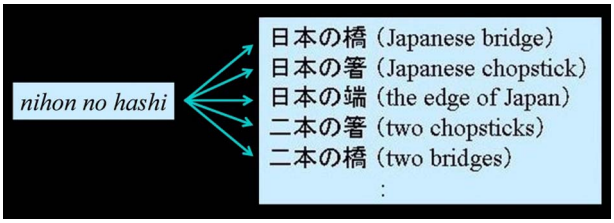


Fig. 2. *Example of homonyms (courtesy of Toshiba Corporation).*

be spelled phonetically using either kana or the Roman alphabet. However, when this is done, there are typically several homonyms that cannot be uniquely determined by morphological analysis.

Fig. 2 shows an example of homonyms. The word nihon, as spelled with Roman letters, has only one pronunciation but two meanings: 日本, or Japan, and 二本, or two. Similarly, hashi has several meanings: 橋, bridge; 箸, chopstick; 端, edge; 梯, step, etc. The result of kana-to-kanji conversion has many candidates because of such homonym combinations.

IV. PRODUCT DEVELOPMENTS

Toshiba started the automatic kana-to-kanji conversion project in 1971. Kenichi Mori's research group solved the homonym problem using usage frequency information and interactive learning. In text editing, a user selects one of the homonyms displayed in the order of the usage frequency information. This information was based on the usage frequency in newspapers. A user's selection is automatically learned and used to update the usage frequency information: using the algorithm of last used homonym is the first out next time. Use of the last-in–first-out algorithm for selection of the homonym was crucial to the success of the conversion technology.

In September 1978, Toshiba exhibited the JW-10, the first Japanese language word processor, with the following hardware specifications:

- central processing unit (CPU): about 0.2 MIPS;
- memory: 64 kB;
- storage: hard disc drive: 10 MB, floppy disc drive: 250 kB;
- keyboard: JIS layout (almost the same as QWERTY layout);
- monitor: 24 × 24 dot matrix, 12-in raster-scan CRT;
- printer: 24 × 24 dot matrix, 24-pin wire-dot impact printer;
- size: 1115 mm (W) × 1022 mm (H) × 760 mm (D).

The biggest problem in hardware at that time concerned printers. There were very expensive kanji line printers for mainframe computers, but they were too costly, large, and noisy to be utilized in offices.

Toshiba, therefore, developed a small and inexpensive dot matrix kanji printer in 1975. A dot matrix printer has a print head that holds several pins in a line. These pins, or wires, are driven by electromagnetic coils to strike the ink ribbon to print characters. To express correct kanji composed of many strokes and to avoid any incorrect characters, it is necessary to compose character patterns with at least 24 black dots available both horizontally and vertically. Therefore, the print head needs to have 24 pins, or more than twice that of conventional print heads then used for English letters. By using 24 fine wires, each 0.2 mm in diameter, and arranging them in a staggered pattern on the print head, Toshiba's engineers succeeded in realizing a print head as small as those

Fig. 3. *Prototype of Japanese-language word processor JW-10, Toshiba, 1978 (courtesy of Toshiba Corporation).*

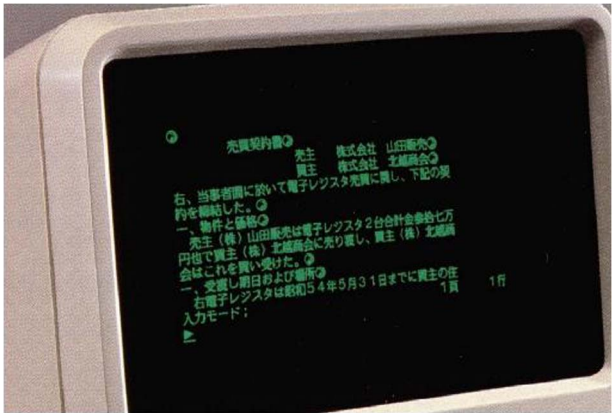


Fig. 4. *Chinese character display of JW-10 (1978) (courtesy of Toshiba Corporation).*

for the Roman alphabet, which could print characters of 3.85 mm^2.

The JW-10 kanji printer was a pioneer for 24-pin dot matrix printers, although it was embedded in a Japanese word processor. Clearly the increased pin count for printing kanji characters also permitted superior print quality for Roman letters and graphics in comparison with the conventional 9-pin printers then used for Western alphabets. By the mid-1980s, other manufacturers had increased the pin count from 9 to 18 or 24 pins for home printers. Also, 24-pin printers were becoming popular worldwide owing to their high print quality. Before inkjet printers appeared, 24-pin printers were the most commonly used for personal computers (PCs).

In 1978, Toshiba exhibited the first Japanese word processor, the JW-10, equipped with the printer described above and a 12-in dot-matrix CRT that provided a what you see is what you get (WYSIWYG) display. In 1979, it began shipping the product shown in Fig. 3. People could directly edit Japanese documents on the character display shown in Fig. 4.

In 1977, Sharp Corporation demonstrated a prototype of kana-to-kanji conversion at the Business Show Tokyo exhibition, but it abandoned development of a commercial system. Instead, two years later, Sharp released its Japanese word processor, the WD-3000 Shoin. It had a kanji pen-touch panel instead of a keyboard but lacked a kana-to-kanji conversion function. The user selected kanji characters from the full-kanji matrix on the panel.

Other major Japanese electronics companies subsequently marketed Japanese-language word processors. As a result, the kana-to-kanji conversion technology became the standard Japanese input method after 1980.

Some Japanese word-processor keyboards featured special alignments. In 1980, Fujitsu Limited released its Japanese word processor OASYS (Office Automation SYStem) with a special, non-JIS keyboard named the thumb shift. The kana-to-kanji conversion of this system was different from the automatic-word-segmentation method. To avoid grammatical analysis and automatic word segmentation, the system required the user to press a special key at the end of every word with his or her thumb. Professional operators initially preferred this system because its accuracy in selecting the correct kanji was superior, but this superiority diminished in time as Japanese language analyses improved.

NEC introduced its M System keyboard, which also used a thumb shift, in 1983. Named after its inventor, Morita Masasuke, the keyboard was designed on his insight that it could be divided into 15 vowel keys for the left hand and 15 consonant keys for the right hand, using the on'yomi or Chinese pronunciation of kanji. He based the keyboard on the kana Gojūonzu, the table of 50 sounds, and used thumb shifting to combine vowels and consonants. The simplicity and logic of the arrangement enabled some users to input text at two times the rate of kana-to-kanji input methods.

Although the mass adoption of computers with western-style keyboards ultimately drove corporations from offering specialized keyboards, enthusiasts who appreciate their input rate and national origins continue to order custom-made thumb-shift keyboards.

After the JW-10 appeared, Toshiba developed smaller products. These included the desktop-type JW-1 in 1982 and the laptop-type Rupo JW90F in 1985, shown in Fig. 5. For the Rupo Miwako Doi developed a new user interface for novice users based on

Fig. 5. *Toshiba's small-sized product, Rupo, sold at a price slightly less than ¥100 000, 1985 (courtesy of Toshiba Corporation).*

Fig. 6. *Toshiba's last Japanese word processor, the Rupo JW-G7000, 1999–2000 (courtesy of Toshiba Corporation).*

Toshiba's user experiments. Word processors at that time used an overlay input method. If a user made a mistake, the corrected input letters overlaid the originals and deleted them. In order to preserve the user's input letters, the Rupo adopted Doi's insert input method instead. Today, this technique is the most popular one for PCs, cellular phones and smartphones, and personal digital assistants.

Japanese word processors spread not just to offices but also to homes. The annual production grew to about 3 000 000 units in 1990 and 27% of Japanese homes possessed one. In time, Japanese word processors offered an increasing range of functions: desktop publishing, spread sheet, graph generator, communication facility, optical character recognition (OCR), machine translation, etc.

V. TRANSITION TO COMPUTERS AND CELLPHONES

During the 1990s, Japanese word processors became obsolete because of the growing popularity of PCs and word-processing software that incorporate kana-to-kanji conversion using automatic word segmentation (Fig. 6). Programs like Microsoft Word and Justsystems' Ichitaro (1985) made this conversion a basic function of PCs. By 1999, major Japanese companies had exited the Japanese word-processor business.

By 2010, kana-to-kanji conversion systems had been installed on tens of millions of PCs and also on 100 million mobile phones in Japan. This information technology is pervasive. Cloud-type input methods based on access to the Internet are also available, such as the Google Japanese input Cloud application program interface (API). Users can input Japanese on a PC that does not contain its own Japanese IME by using a cloud-type input method.

Extensive use by Japanese of the Internet has altered their preference for Japanese language input methods on their PCs. Specifically, most Japanese now use Roman alphabet characters rather than kana for inputting Japanese text. The underlying conversion technologies are the same and software programs work equally well with either of the two input methods. Interestingly, the original Toshiba JW-10 word processor was designed with both capabilities, but at that time the vast majority of users preferred kana input. The growing preference for Roman character input on PCs may also be influenced by the fact that all children learn the Roman alphabet in school.

On the other hand, many cellphone users prefer to use kana for input because it generally requires fewer keystrokes than when using the Roman alphabet. Five kana characters are assigned to each key in a manner similar to the assignment of three Roman alphabet characters to each key. Assume that you want to input 日本, nihon, in kana characters (にほん). You push key 5, which is assigned to なにぬねの, two times for に; key 6, which is assigned to はひふへほ, five times for ほ; and key 0, which is assigned to わをん, three times for ん. Therefore, the total number of strokes is ten. In the case of Roman alphabet characters, the total number of keystrokes is 12.

The Japanese word processor was not merely an adaptation of any existing system, but a highly innovative system realized by entirely new software and hardware. It combined modern information technology with the traditional Japanese culture of kanji and kana characters, and revolutionized the availability of information technology in all aspects of Japanese life, including offices, factories, schools, and homes. It contributed to economic growth, personal development, and enhanced living standards throughout Japan. ■

Miwako Doi

Haitao Lei

ABOUT THE AUTHORS

Miwako Doi (Fellow, IEEE) received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1977, 1979, and 2002, respectively.

She is a Chief Fellow of the Toshiba Research and Development Corporate Center. She was one of the Japanese word-processor researchers.

Dr. Doi is a Fellow of the Information Processing Society of Japan (IPSJ) and the Institute of Electronics Information and Communication Engineers (IEICE). She has been awarded 11 other honors, most notably IPSJ's and IEICE's Achievement Awards and a National Commendation for Invention.

Haitao Lei received the B.S. degree in electronic engineering from Zhejiang University, Beijing, China, in 1984 and the Ph.D. degree in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1992.

He is a Resident Officer to Toshiba (China) Co., Ltd.; Vice President of Toshiba (China) Co., Ltd.; and Director of the Research & Development Center of Toshiba (China). He participated in the Joint Pinyin-to-Chinese Project of Dalian University of Technology and Toshiba in 1996.