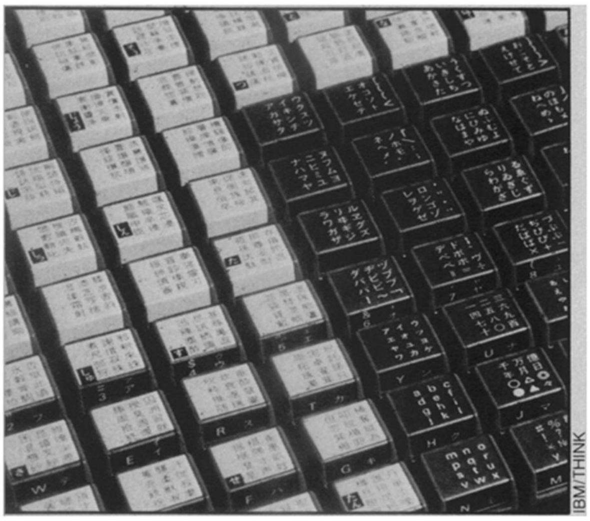
Computers with Character



IBM/THINK

Chinese calligraphy is exquisite and eloquent but it causes major problems for today's computers

BY ROBERT J. TROTTER

How do you take a language with 50,000 characters and make it compatible with a conventional computer keyboard of 50 characters so that a country with one billion people can catch up with 20th century technology?

The problem, of course, is the Chinese written language. Its characters are beautiful, but there are just too many of them to be handled efficiently by existing computers. The problem began 6,000 years ago when the Chinese language originated with pictures scratched on bones and shells. These pictographs gradually evolved into ideographs (symbols used to represent words and ideas) and finally into the current square Han characters. Over the years, Chinese writing followed a pattern of divergence that led to the eventual development of 50,000 characters. Indo-European languages, on the other hand, followed a pattern of convergence. Hundreds of ancient hieroglyphs were eventually funneled by the Egyptians, Hebrews, Greeks and Romans into the 26 standard characters of the Latin alphabet. Today, 10,000 Han characters are in circulation and more are being coined. To be "newspaper literate" one must know at least 2,000 characters; to be college-educated, 5,000.

As beautiful and as time-honored as the Han characters may be, they do present certain obstacles to a technological society. In the 1920s, for instance, the Chinese intellectual Lu Xun complained that Chinese characters are too complicated to educate the masses. Students, he said, spend too much time learning to write Han characters and practicing calligraphy—at the expense of their education in science and mathematics. For a country that is trying to upgrade itself technologically, this is serious. Chinese typewriting and typesetting, for example, are complicated, expensive and time-consuming. But perhaps the greatest problem is with computers. Fortunately, computers may also hold the eventual solution to that problem.

The best answer would be an optic scanner or some other device that could read the 10,000 Chinese characters (or Korean or Japanese, which are based on Chinese). But this is not possible with current computer technology, so linguistics and computer experts have been working on various complex keyboard arrangements and coding schemes that allow the Chinese, Koreans and Japanese to use computers in their own languages.

One example is the IBM information processing system developed for Japan and Taiwan. After ten years of R&D, IBM researchers developed a system that can do everything that can be done with English-language equipment. Software is the key to its success, explains IBM's Charles Swift. Conventional data processing uses one byte, or eight bits, of coded information. But that yields only 256 character combinations. This system can handle thousands of characters because it processes things in terms of two bytes instead of one. A special keyboard and a character generator were the other necessary elements in its design.

The keyboard is large—large enough to hold more than 2,000 characters. Each key has 12 characters on it, and at the side there is an additional 12-digit keyboard. If the operator presses key number 12, for example, then selects a key on the large board, the 12th character on that key is displayed. It takes an operator about six weeks of training to get to a top speed of 60 to 75 characters per minute. Top speed on a Japanese typewriter is about 35 characters per minute.

High-resolution allows for accurate video display of the complex characters, and an ink jet printer can produce 37 characters per second at the terminal. Laser technology allows the system to print 10,000 lines per minute.

Wang Laboratories of Lowell, Mass., has taken a different approach to the Oriental character problem. Last year they began marketing the Ideographic Word Processing System that can operate in conventional (Mandarin) Chinese, simplified Chinese or Japanese.

Instead of displaying thousands of characters on a keyboard, the Wang system uses a coding system. In this way, a minimum number of keys can be used to generate 10,000 characters. Each character has a six-digit identification number based on the shape of the character. According to the company, "Users familiar with the basic shape and structure of Chinese characters can be trained to use the method [called the three corner coding method] quickly and easily. In fact, an operator need use only 297 character elements and 15 rules to be fully proficient on the entire system." The system can be learned in two weeks. If offers editing capabilities such as insertion, replacement and deletion of characters, lines, paragraphs or entire sections of text. Standard disk storage can range up to 137.5 million characters.

A less complicated system for entering Chinese characters may be the one developed at Cornell University by Paul L. King, with grant money from the National Cash Register Corp. King says a Chinese-speaking person with the equivalent of a junior high school education can learn to operate this system in about half an hour and enter characters at a rate of 50 per minute — nearly five times as fast as a highly skilled person can operate a Chinese typewriter. The system is easy to operate because it uses a 12-digit keyboard to enter the thousands of characters. Each digit describes a basic shape used in Chinese characters in one of the four quadrants into which all the characters are divided. By selecting up to four keys, an operator can identify an entire character. Because of the complexity of the characters, however, the same four digits can sometimes produce ten or more characters — all similar enough to have been described by the same digits but very different in meaning. When this happens, the system uses linguistic rules to automatically select the correct character. And if the automatic selection process is not specific enough, the computer displays the remaining choices and the operator makes a manual selection. The system contains about 2,500 words, and additional sets of 500 special vocabulary words are being developed.

And if an operator doesn't want to go to the trouble of learning one of these systems, there may eventually be a system for on-line recognition of handwritten Chinese characters. E. F. Yhap and E. C. Greanias of the IBM Thomas J. Watson Research Center in Yorktown Heights, N. Y., describe this still experimental process in the May IBM JOURNAL OF RESEARCH AND DEVELOPMENT. It consists of a specially designed tablet that produces a pattern of varying electromagnetic signals that are picked up by the pen and then fed through a five-part recognition program. Preliminary tests of the system, which recognizes 2,249 symbols, have found it to have a recognition rate of 97.8 percent.

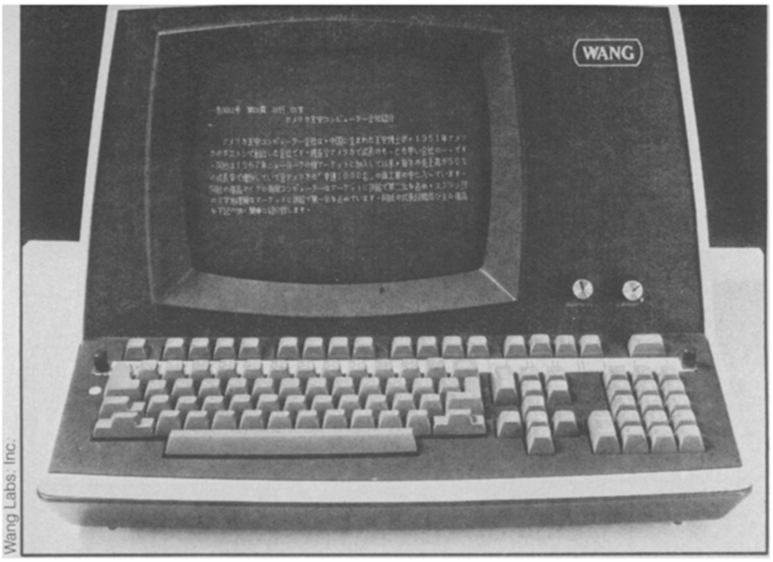
"On-line tablet recognitior.," say the researchers, "offers considerable promise as a natural data entry device for casual users of information systems."

H. C. Tien offers what he considers to be an even more natural data entry system. Instead of working with character-recognition systems, he wants to computerize China's language by alphabetizing it in such a way that it will work with existing, Latin alphabet computers — and after 20 years, he thinks he's done it.

Tien, an M.D., is editor of the Michigan Institute for Psychosynthesis in Lansing. The institute's goal, he says, is "the union of Eastern and Western medicines, thoughts and languages." And the quickest way of doing this, he believes, is by alphabetizing the Chinese language. The system he developed, he says, "has the power to transcribe common spoken Chinese (Putonghua) and to facilitate and accelerate printing, typing, telegraphy, computer input-output, indexing, library cataloging, scientific reprints ...."

Tien is following a long tradition of attempts to reform and simplify the Chinese language. The most recent attempts include the simplification of Han characters, the teaching of Putonghua (the Beijing dialect) throughout the country and the introduction in 1958 of the Pinyin System — a Chinese phonetic alphabet.

Although based on the Latin alphabet, the Pinyin system resists computerization because of the many homophones (words that sound the same but have different meanings) in the Chinese language. Many Han characters, for example, are spelled the same way in Pinyin. When spoken, these words are differentiated by tone. When written in Pinyin, they are differ-

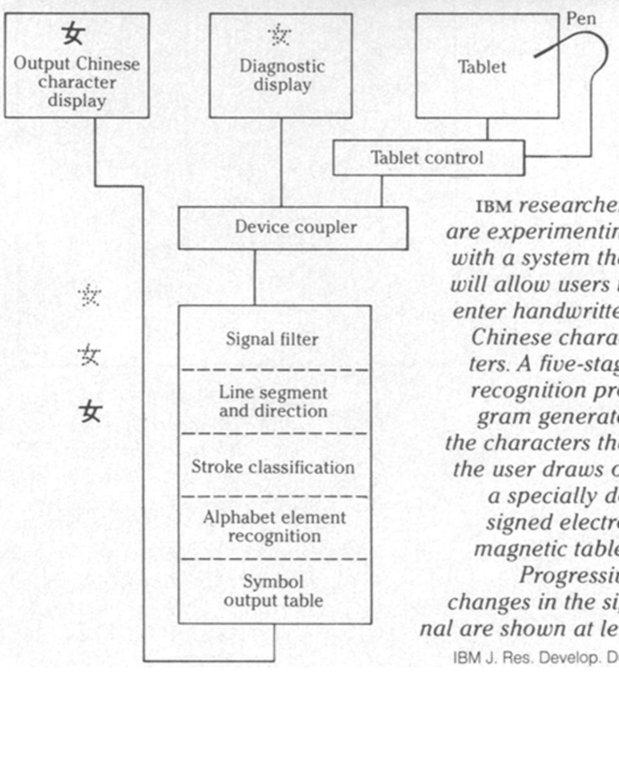


Wang Labs. Inc.

The Wang system uses the three corner coding system, based on the shape of Chinese characters, to generate 10,000 characters with a minimum number of keys.

THINK/IBM

The IBM system (above) with 12 characters to the key and an additional 12-digit keyboard can generate more than 2,000 characters.



IBM researchers are experimenting with a system that will allow users to enter handwritten Chinese characters. A five-stage recognition program generates the characters that the user draws on a specially designed electromagnetic tablet. Progressive changes in the signal are shown at left.

IBM J. Res. Develop. Dev.

entiated by diacritical marks (an extra burden to computers). But there are only four tonal or diacritical marks in Pinyin, and in some cases these marks have to separate more than four similar-sounding characters. The word "ma," for instance, has 18 homophones; "li" has at least 81, and "yi" has 126.

In order to eliminate the diacritical marks and solve the homophone problem, Tien uses a letter-doubling technique and a set of 189 suffixes. Most Chinese syllables consist of a consonant and a vowel, so letter doubling can yield four distinct spellings (ba, bba, bbaa and baa), which correspond to the four tones.

"We are now one step closer in the computerization of Chinese characters," says Tien, "but another obstacle remains." Most Chinese characters consist of two parts, or radicals. One is spoken, the other is silent (it may signify such things as the gender of the word). A character pronounced ba, for example, may be made up of the radical pronounced ba and one of a number of silent radicals that change the meaning of the word. Tien uses silent suffixes made up of letters or letter combinations to represent these silent Han radicals.

This, in essence, is what Tien calls the Pinxxiee System. It includes the phonetic syllables of the Pinyin System (which is being learned by all school children in China), the letter-doubling system for the tones and silent suffixes for the silent radicals. "It is obvious," he asserts, "that every character may now be uniquely and equivalently transformed for programing into the computer." The Pinxxiee system, he continues, can lead to the alphabetization of all Han characters with existing computer technology without waiting for further development of pattern recognition or image processing techniques and equipment.

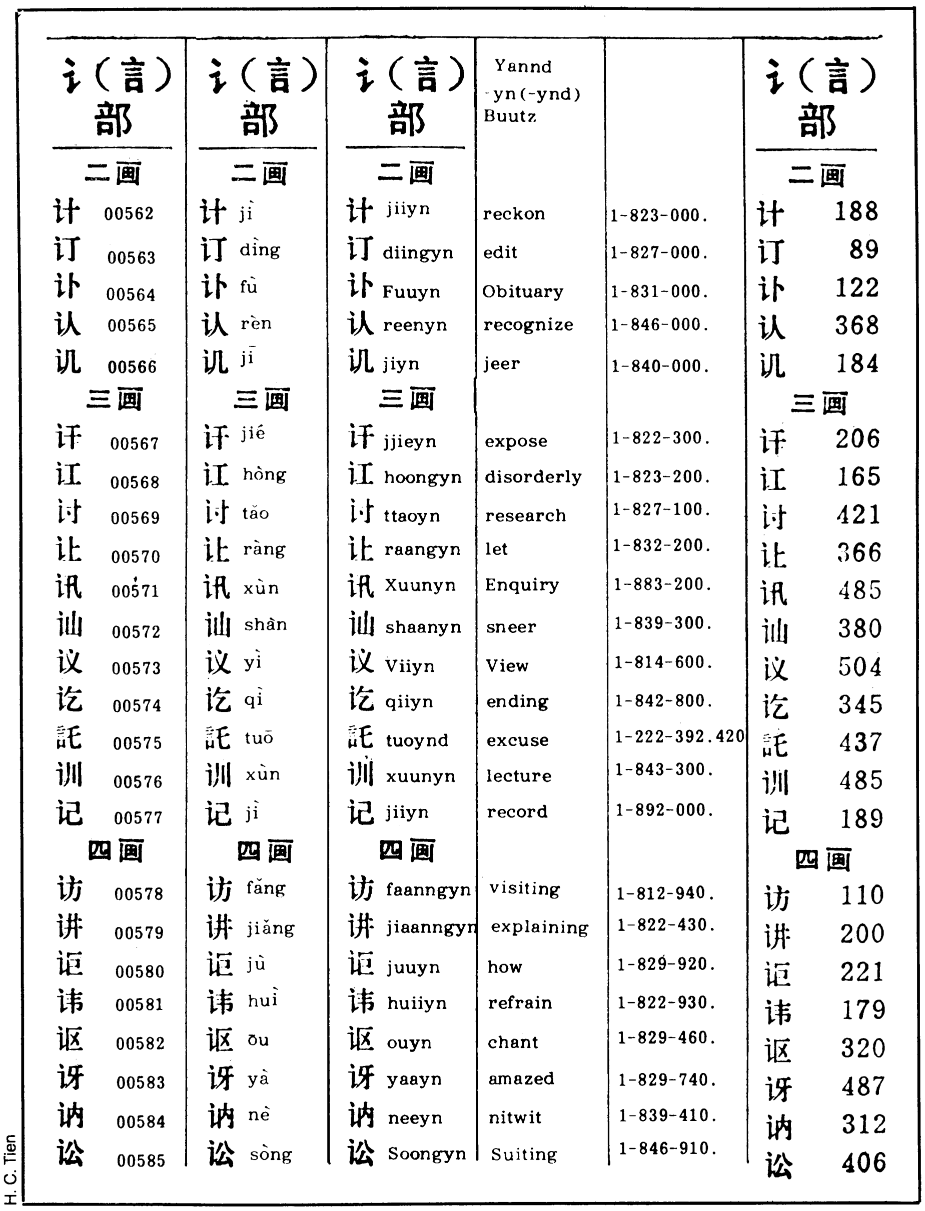
There is, however, still one major problem: convincing the Chinese to use the Pinxxiee. Cultural pride, respect for the ancestors who contributed to the heritage of the Chinese script and the force of habit of thousands of years are involved. Tien admits this won't be an easy problem to solve, but he is working on it. In the case of the silent suffixes, for example, he has attempted, when possible, to use Latin letters that at least look something like the original Han radicals.

Tien is also working through formal scientific channels and the Chinese government. He described the Pinxxiee system last fall in Hong Kong at an international computer conference and has been dealing for several years with the Chinese Ministry of Education. He has published a two-volume English-Han-Pinyin-Pinxxiee dictionary that contains almost 12,000 computer-coded words. "This is just the beginning," he says, and quotes Lu Xun: "Shall we sacrifice ourselves for the ideographs, or shall we sacrifice the ideographs for ourselves? All but the insane can answer this immediately." □

The Origin of Pinxxiee

DERIVATIONS	1ST TONE	2ND TONE	3RD TONE	4TH TONE
Pinxxiee	bavj	bbaty	bbaaty	baatv
Letter-doubling Technique	ba	bba	bbaa	baa
Pinyin	ba	ba	ba	ba
Han Character	疤	拔	把	坝
English	scar	pluck	handle	dike

H. C. Tien



讠(言)部	讠(言)部	讠(言)部	Yannd -yn(-ynd) Buutz		讠(言)部
二画	二画	二画			二画
计 00562	计 jì	计 jiiyn	reckon	1-823-000.	计 188
订 00563	订 dìng	订 diingyn	edit	1-827-000.	订 89
讣 00564	讣 fù	讣 Fuuyn	Obituary	1-831-000.	讣 122
认 00565	认 rèn	认 reenyn	recognize	1-846-000.	认 368
讥 00566	讥 jī	讥 jiyn	jeer	1-840-000.	讥 184
三画	三画	三画			三画
讦 00567	讦 jié	讦 jjieyn	expose	1-822-300.	讦 206
讧 00568	讧 hòng	讧 hoongyn	disorderly	1-823-200.	讧 165
讨 00569	讨 tǎo	讨 ttaoyn	research	1-827-100.	讨 421
让 00570	让 ràng	让 raangyn	let	1-832-200.	让 366
讯 00571	讯 xùn	讯 Xuunyn	Enquiry	1-883-200.	讯 485
讪 00572	讪 shàn	讪 shaanyn	sneer	1-839-300.	讪 380
议 00573	议 yì	议 Viiyn	View	1-814-600.	议 504
讫 00574	讫 qì	讫 qiiyn	ending	1-842-800.	讫 345
託 00575	託 tuō	託 tuoynd	excuse	1-222-392.420	託 437
训 00576	训 xùn	训 xuunyn	lecture	1-843-300.	训 485
记 00577	记 jì	记 jiiyn	record	1-892-000.	记 189
四画	四画	四画			四画
访 00578	访 fǎng	访 faanngyn	visiting	1-812-940.	访 110
讲 00579	讲 jiǎng	讲 jiaanngyn	explaining	1-822-430.	讲 200
讵 00580	讵 jù	讵 juuyn	how	1-829-920.	讵 221
讳 00581	讳 huì	讳 huiiyn	refrain	1-822-930.	讳 179
讴 00582	讴 ōu	讴 ouyn	chant	1-829-460.	讴 320
讶 00583	讶 yà	讶 yaayn	amazed	1-829-740.	讶 487
讷 00584	讷 nè	讷 neeyn	nitwit	1-839-410.	讷 312
讼 00585	讼 sòng	讼 Soongyn	Suiting	1-846-910.	讼 406

H. C. Tien

The Pinxxiee system devised by Tien uses letter-doubling and silent suffixes to alphabetize the Han characters. Tien's computer-coded Han-Pinyin-Pinxxiee dictionary contains almost 12,000 words.