Gaines, B. R. (1983). Office automation in the languages of the world. In Proceedings Graphics Interface '83, pp. 337-344. Canadian Man-Computer Communications Society, Toronto.

OFFICE AUTOMATION IN THE LANGUAGES OF THE WORLD

Brian R. Gaines

CADRE Information Transfer Systems Inc.,
339, Canarctic Drive, Downsview, Ontario M3J 2P9.

ABSTRACT

Computer-based information systems have been restricted to those languages using the Roman alphabet. In non-numeric applications, the limitation has led to an increasing differential in applicability between those countries using Latin languages and the majority which do not. Office automation, database and educational applications in particular have been severely restricted, and these are very important to the developing nations of the world. Low-cost graphics technology now provides the means for information systems to communicate as readily in "exotic" scripts as in Roman characters. Input, display and printing subsystems can be made available that cope with such scripts as Chinese, Devanagari, and Urdu Nastaliq. Problems remain in moving the operating systems, programming languages and application packages across language barriers. This paper is based on experience in developing systems for China, India and Pakistan, and covers both the technological ease of implemention and the remaining problems. It is concluded that Canada has a combination of computing and linguistic knowledge that would make development of "universal language" computer systems simple and rapid as a co-ordinated effort.

KEYWORDS: Office automation, exotic languages, laser phototypesetter, word processing, telex, ideographic computing, Chinese.

INTRODUCTION

Advances in computer technology during the past three decades have made information systems accessible to all. These systems now play major roles in our society: the operation of government, the management of commerce, the development of science and the application of technology would be seriously impeded without them. However, the "all" in the first sentence must be qualified to be "all those whose native language uses the Roman alphabet". For those whose language uses "exotic" characters information systems technology has been accessible only to the extent that they can use Roman characters.

When computer systems were in use largely for numerical data processing this was not a severe limitation. Learning programming in a foreign language is not much worse than learning it at all, and arabic numerals are commonly used in most of the world. However, as information technology has impacted non-numeric data processing, the Roman character limitation has caused an increasingly severe differential in applicability between those countries using Latin languages and the majority which do not. Office automation, database and educational applications in particular become virtually impossible, and these are some of the most important to the developing nations of the world.

Fortunately, graphics technology has also decreased in cost and increased in capability and provides the means for information systems to communicate as readily in the "exotic" scripts as in Roman characters. Input, display and printing subsystems are now available that will cope with languages such as Chinese and Urdu Nastaliq, each requiring some 15,000 or more characters. However, problems remain in translating the operating systems, programming languages and application packages across language barriers. This paper is based on experience in developing information systems for China, India and Pakistan, and covers both the technological ease of implementing such systems and the remaining system problems.

A SYSTEM FOR CHINESE PRINTING

It is technically obvious to someone in the computer graphics industry that current equipment can be used to provide, for example, word processing in any language of the world. However, such systems for exotic languages have been slow in development. To illustrate the technical ease and some of the less obvious practical problems, I will describe a case history of five years ago when the Chinese language was subject to photypesetting for the first time. I was then Technical Director of Monotype, a British Company which had a history of setting languages for the first time. The Arabic and Indian languages had been covered by Monotype some years ago. Chinese had been neglected because the requirement for 5,000 or more characters was beyond the capability of any automatic typesetting system. However, the advent of the laser phototypesetter with characters on computer disk made Chinese feasible and we decided to develop a system for sale to China. The most interesting message from the case history is the ease and speed with which this could be done. The work commenced in October 1978 and by June 1979 complete systems had been developed and installed in Beijing and Shanghai (Gaines 1981).

The Chinese Language

It is impossible to define precisely the number of characters, or ideograms, in the Chinese language. There are over 60,000 ideograms recorded in use during different periods of Chinese history while the modern standard dictionaries used in China list some 13,000 characters currently in use. For the printing of books a face of some 7,000 characters is adequate and for newspapers some 4,500 characters. Chinese typewriters provide about 2,000 characters available in the type case under the print head and about another 2,000 available for insertion as required.

Chairman Mao Zedong in all his writing used a vocabulary of only 3,006 characters. There is a major movement in China to simplify the Chinese language by restricting it to only 3,260 characters but this is a contentious issue (Seybolt & Chiang 1979). For printing purposes, no matter how many characters are provided there will always be the need for more through a good 'sorts' facility since specific jobs require access to nonstandard characters, for example in quoting from an ancient Chinese work. Similarly in database systems it should be noted that the most nonstandard characters are those for personal and place names.

The calligraphy of Chinese characters was greatly simplified in China after the liberation in a move to aid literacy. A standard form of phonetic romanization of the the Chinese language, Pin Yin, was also introduced and is widely used in China for shop and street names; however, it has yet to have any major impact on the printing industry. The direction of setting of Chinese text was also changed to correspond to the Western format of horizontal, left-to-right reading, rather than the original Chinese vertical setting from right to left. The simplified characters and horizontal setting are primarily used in China, Singapore and Malaysia, but the original characters and vertical setting are still used in Hong Kong, Taiwan, and, with extensions, in Japan.

Because of the large number of characters required, Chinese text is still primarily hand-set using hot metal techniques in the printing industry. The operators work within an alcove of type cases containing the characters needed for the text they are preparing. The arrangement of the characters and the number available is a feat of organization that minimizes the effort in hand picking for a particular text. The configuration required for rapid setting of newspapers can be quite monumental with operators literally skating from case to case to achieve high speeds. On routine textual material skilled operators can achieve continuous throughput of up to 1,000 characters an hour. As a rough guide in translating these figures for comparison with Latin languages a three to one ratio has been found appropriate in translated material, i.e. one Chinese character requires about three English ones on average. Thus a comparable setting rate for English would be about 3,000 characters an hour, or just under one a second. This compares unfavourably with the setting rates for skilled keyboard operators with Roman texts.

The key subsystems in a phototypesetting system for Chinese are: keyboards for text acquisition and editing; visual displays for editing; printers for proof-

ing; high-quality phototypesetters for final composition. The key additional requirements that the Chinese language places upon these all stem from the number and complexity of the ideographic characters.

A Phototypesetter for Chinese

The phototypesetter itself presented no problems. Indeed the initial impetus for this development came from the realization that in its "Lasercomp" laser-based, digital phototypesetter Monotype had available the technology to set Chinese text with the full repertoire of characters and typesfaces required and with the quality also essential. The Lasercomp is essentially a high resolution raster scan plotter in which a Helium-Neon laster is used to form an image on photographic material. The laser beam is one thousanth of an inch in diameter and deflected horizontally by a spinning mirror system. Vertical traverse is through the movement of the film material in synchronization with the mirror. Essentially one can place any number of dot 1 thou in diameter anywhere on film material 58 picas or 100 picas wide (10 inches or 17 inches wide) and of virtually indefinite length. Hence one can build up images of any form at a resolution below that of the eye.

The Lasercomp is controlled by a Computer Automation 1si4 computer with special microcode for graphics. Character founts are held in digital form on CDC 80 MB discs and there are no intrinsic limitations to the number of different characters the Lasercomp can handle. The machines used in China have a capacity of some 60,000 Chinese characters, sufficient to store the range of ideograms required in a number of different faces.

A Chinese Keyboard

Whilst the typesetter presented no problems the same was not true for the keyboard technology. In China itself there are over 200 known keyboard designs under development and evaluation and other designs for ideographic keyboards are in use or under development in Japan. Some use a large bank of keys from 700 to 2,000 or more with multiple shifts so that each key represents from four to twelve characters. Others use multiple key depressions for a single character representing it: in phonetic form; as the sequence of strokes making up the written character; as a set of shapes, or radicals, making up the character, and so on. All these different forms of keyboard have their merits and demerits, their proponents and opponents, and there are many hundreds if not thousands of papers and articles describing them. This is quite apart from system that use the Romanized forms of the Chinese characters and translate them back to the ideograms.

It would take a separate paper to even begin to describe the plethora of Chinese keyboards, the underlying approaches, the relationships between them, applications studies, and so on. It is more relevant here to note some of the logic a commercial manufacturer must apply in selecting a suitable keyboard for any language. The primary criterion is not a technical one of speed, accuracy, speed of learning, or cost of manufacture. It is one of acceptance to the purchasers - what is the standard which the users, or the country of use, has adopted or will adopted. For example, the QWERTY keyboard is the de facto standard in the UK and USA despite its severe and well documented deficiencies. Unfortunately no such standard yet exists in China. However, it became apparent in 1978 that a keyboard system developed in the Machine Translation Unit at the Chinese University of Hong Kong had a good chance of at least serving as a model for a high speed multiple key per character keyboard in China. This keyboard had been given much publicity in China and the China Printing Corporation was interested in giving it practical trials for use in book setting.

We developed a version of this keyboard incorporating Roman characters also together with typographic commands and put it into production rapidly as a variant on a standard large keyboard normally used for mathematics typesetting. The Chinese keyboard in the version developed had 238 keys arranged on a matrix of 14 rows by 17 columns. This was fitted in the centre of the 17 rows by 26 columns of the Monotype LD400 keyboard, and typesetting function keys, Latin and Greek alphabets, and mathematical keys were fitted around it. For ergonomic reasons the central region of the Chinese keyboard was designed as 10 rows of 11 columns each representing one symbol. These are arranged so that the first row has the one-stroke Chinese character for 1, 一, at the centre and consists of single stroke characters or

radicals. The second row has the two-stroke Chinese character for 2, 二, at the centre and consists of double-stroke characters or radicals. The remaining rows of the central keyboard follow this same logic. The keys to the left of this central region represent components of characters that occur only on the left hand side of an ideogram; those to the right of it occur only on the right hand side; those above occur only in the top part of an ideogram; and those below it occur only in the bottom part. The keys outside the central region each represent two symbols but there is no need for a shift key to distinguish between them since they can never be used as alternatives for one another.

This form of keyboard is able to represent over 13,000 Chinese characters by sequences of only 238 keys without the use of shift keys and with a layout that is very logical, ergonomically well-organized and easy to learn. The actual sequence of keys to depress for a character is derived from the stroke sequence that would be used to draw it and hence is easy for a person trained in Chinese writing to master. Keyboards have been designed with very few keys that allow ideograms to be entered as their actual stroke sequences. However, the average number of key depressions required then becomes over six and text preparation is unacceptably slow. The coding structure adopted by Monotype may be regarded as an enhancement of such basic schemes in which common key sequences are compressed into single key strokes. The average number of key strokes in standard book work is three with a maximum of nine (later research at Shanghai Printing Research Institute has improved these figures by reducing the maximum to four).

It is worth emphasizing here that the keyboard essentially gives just an unambiguous key sequence for each character. It does not precisely encode the shape or structure of the character. The sequences produced by the keyboard are decoded by computers elsewhere in the system to identify which Chinese characters they represent. Thus the Lasercomp calls its characters by number from its disc and uses a code structure unrelated to that of the keyboard. This is significant in that the entire system was designed to be keyboard independent in view of the probable future variability of keyboard standards in China.

A Chinese Text Editing Terminal

With a suitable keyboard for text preparation and a suitable phototypesetter for text output, it remained only to develop a text editing terminal for Chinese for the typesetting system to be complete. This proved surprisingly easy using modern microprocessor and display technology. The ideographic editing terminal was manufactured from modular computer components available as standard units. It consists of two Zilog Z80 microprocessors, one with 64 Kbytes of store used as a text editor, and the other with 256 Kbytes of store used as a character store for the display. Twin 315 Kbyte minifloppy disc units on the editor unit are used for program and data storage respectively, and a similar unit is used on the character store to hold the rarer characters beyond the most common 5,000. Input to the terminal may be through keyboard, paper tape, minifloppy disc or serial line, and output may be through paper tape, minifloppy disc, or serial line.

The editor operates with Chinese, Latin, and Greek text, mathematical and typesetting symbols, in the normal fashion allowing text files to be displayed, searched, modified, split, merged, and so on. The display screen has a resolution of 256 by 256 picture elements which allows 8 rows of 14 characters to be displayed at a time together with special areas for operator interaction, search strings, status information and error messages (all of which are in Chinese). It was found possible to represent the Chinese characters adequately with a resolution of 16 rows by 14 columns of dots. Some characters have to be distorted but operators found the screen easy to read at this resolution. A proofing printer is also provided with each editing terminal and this consists essentially of an 8-wire matrix printer driven in the "graphics" mode where the wires are directly controlled from the same character store used to drive the display. The standard CP/M operating system was used for the Z80s and the editor programs were written in a mixture of macro assembler and C. Modem communication protocols were also written for the editing computers enabling text to be transmitted over telephone links with full error checking and correction.

SYSTEM EXPERIENCE

With the keyboard, visual display editing terminal and proofing printer, and phototypesetter it was possible to put together for the first time complete phototypesetting systems for Chinese. In demonstrations in Hong Kong in December 1978, 3 months from the start of the project, we had shown only a prototype keyboard working into a Lasercomp with only 300 characters. These demonstrations led to invitations for full-scale demonstrations of book production in China the following year. By June 1979 two systems had been installed in China, one in the Xin Hua printing works in Beijing and the other in a newly built printing factory in Shanghai. These systems each consisted of ten keyboards, three editing terminals, two lasercomps, one character digitizer and one film processor.

Each of the two installations was suitable for complete book production. Each had provision for the preparation, storage and transmission of text either on paper tape or on floppy disc, or through any combination of the two. In use up to ten operators could key in text directly to tape or disc. The output could be proofed through the line printers or directly through the Lasercomps. Tapes or discs could be corrected from the proofs or screens at one of the three editing terminals and the resultant tape or disc used to produce the final output on paper or film at one of the two lasercomps.

Arrangements were made for the Shanghai Printing Research Institute to draw and digitize type faces. This Institute was responsible for drawing the simplified faces adopted in China after the liberation and had the master drawings. In May 1979 two Monotype character digitizers were installed in Shanghai and by August some 5,000 characters had been digitized for the Lasercomp in the face used for the majority of books in China. Bold and italic versions of this face were also digitized so that complete books could be set and compared directly with hot metal equivalents.

By the end of July 1979 the systems were in full operation and Chinese keyboard and system operators were being trained. Six engineers from the Chinese Printing Corporation also came over to the Monotype works in the UK for 6 weeks training on the maintenance of the system. Some twenty Monotype engineers, training staff and demonstrators were also on tours of duty in China during the period from May to October 1979 for periods of two to eight weeks. Conditions in China were difficult at times - when the installation engineers arrived in Shanghai the temperature was 90 F, the relative humidity 90%, and the building for the equipment not yet completed - air conditioning equipment was rushed out and the schedule maintained.

By the end of August 1979 experimental production of books had commenced together with a very wide range of other demonstration material such as newspaper pages, complex tabular and mathematical work, Chinese music, and so on. A 155 page book of French fairy stories translated in Chinese was set in Shanghai and a 182 page book in Beijing. A 12 page booklet was also set and printed in Beijing which contained Chairman Mao's famous speech to the music workers in which he calls upon them to "take that which is best in the West and make it Chinese". This seemed a particularly apposite handout to the many thousands of visitors who flocked to the full system demonstrations in October and November of 1979.

Apart from seeing the book production systems these visitors also had the opportunity to see how modern information technology, "made Chinese", might be used for communications and database systems in China. One demonstration which aroused great interest was the communication of Chinese text over normal telephone lines between Beijing and Shanghai. Modems operating at 1200 baud had been installed in the printing works at each end and it was possible to communicate text from an editing terminal in Beijing to one in Shanghai, or vice versa, and even transmit text directly to the Lasercomps. An important practical application of such communication is in the Chinese newspaper industry where newspaper producers could now keyboard text at one location and, within minutes, be producing printing plates at a number of remote locations. This is particularly significant in China which is a vast territory with difficult physical communications.

It is fascinating to note that the Chinese have moved directly in a very short time from the use of the "zeroth" generation of hot metal printing equipment to

the "fourth" generation of digital laser phototypesetters without ever using the intervening generations of filmsetters, CRT photypesetters, and so on.

GENERAL APPLICATIONS

The demonstrations of the use of the Chinese language in the printing industry described above may be taken in a far wider context to show that it is now possible to develop "Information Industries" around languages other than those based on Roman type faces. The significance of this in the spread of computer technology to the majority of the world's population who do not use Roman character sets cannot be overestimated. Clearly some such languages have already been tackled, Arabic in particular. However many of the languages of the Indian and Asian continents have been regarded as beyond computer use currently. This is clearly no longer so. It is now feasible to design "Universal Language" systems that operate in all the languages of the world and whose cost is not very much greater than those operating in the English language alone.

Telex

For any language the only requirements for a telex system are a keyboard and a printer interfaced together through a serial communications link. Our keyboard for Chinese was designed with an RS232 output for convenience of interfacing to computers and hence was intrinsically capable of operating over any serial communications link. Telex systems operating in any language of the world, in which when one types in Chinese it comes out as Chinese, in Arabic as Arabic, and so on, are now technically feasible. The commercial and political significance of these may be noted by the fact that currently messages to the Chinese Ambassador in London have to be translated into English for telex transmission and then re-translated into Chinese on receipt. The laboriousness of such procedures and the scope for error is restrictive of all forms of communication.

Typewriters

Clearly similar considerations as to telex apply to typewriters. In general the output has to be of higher quality but ink-jet printers, variable pitch matrix printers, and low-cost laser printers make the printing of arbitrary characters at correspondence quality technically feasible on a cost-effective basis.

Word Processing

The editing terminal in the MITS system is a basic word processor. In any word processor only the keyboard, screen and printer have to be extended to cope with a variety of languages. The text processing software can remain the same for all applications.

Computer Data Entry

The telex and word processing applications already described give a suitable technology for data entry and retrieval from computer systems. In our studies of Chinese database systems we have used the microprocessor in the word processing system to encode fields of Chinese characters into ASCII sequences that appear as character strings to standard database software. Similarly it has been possible to recode the messages from the computer supporting the database as equivalent Chinese messages. This recoding is by no means language "translation" in the full sense of the word but just a table lookup requiring knowledge of the system generating the messages. Since this system is a computer, however, the 'language' used is simple, formalized and algorithmically generated so that it is generally easy to recode it.

Thus the modification of the input system of a database designed for use with Latin languages to operate with any arbitrary language whatsoever is straightforward and very cost-effective in comparison with rewriting the database software itself. The technique is essentially one of syntactic signal conditioning in the terminal equipment. In cost terms once the terminal has the keyboard, display and printer capabilities, the marginal cost of code conversion is only a few percent. Thus the "intelligent" terminal can display its intelligence in languages other than those based on Roman characters.

Computer Programming

It is natural to consider the extension of the data entry techniques described above to programming languages. If one

contemplates the problems of programming in English and then adds to these the problems of using a foreign language in an entirely different script then the difficulty of developing indigenous information industries in Asia and Africa can be appreciated. It would be attractive to use syntactic signal conditioning as for databases to recode Basic, Cobol, and so on, programs in Chinese into a semantically equivalent English program that could be compiled or interpreted using standard computer software.

We funded a detailed study of an ideographic front-end processor for accessing English language computer systems (Witten & Ng 1981) which showed that, whilst most language constructs could be translated directly as for the database, a few posed major difficulties. The problems arose mainly with string data where the length and position of items within a character string had no direct relation to the data structure after the translation. This forces some form of semantic pre-processing for operations involving string manipulation. However, it is certainly possible to offer substantial subsets, sufficient for many applications, through the simple syntactic recoding.

Computer Systems

The editing terminal described is basically a "personal" computer and there are no problems in making its data processing capabilities also available in Chinese or other lanaguages. Such a development could be of great importance to third world countries where the use of small computers in light industry, education, communications, and so on, is probably of far greater significance than the availability of very large computing systems. There are no intrinsic problems to making small computer systems that operate completely in the orthoography of such languages as Chinese, Japanese, Devanagari and Urdu.

CONCLUSIONS

The argument put forward in this paper is:

1) Computer systems are increasingly significant to the operations of government, industry and commerce in all nations;

2) Computer systems development has taken place largely in Western nations using the English language;

3) This had led to the adoption of the Latin alphabet as the standard communications code between computer and user;

4) This presents problems in applying computer technology in countries where the everyday language does not use the Latin alphabet;

5) For numeric applications this is not a very serious problem;

6) For computer programming the need to use a foreign language to write programs is a restriction on the widespread dissemination of programming knowledge and personal computing, but not too serious in professional applications;

7) For non-numeric applications, particularly those involving personal and place names, the Latin alphabet is a severe restriction -- yet record systems in government departments and medical services are some of the most important to developing countries;

8) Office automation in particular is becoming key to the operation of commerce in the West but the developments are based on Latin languages -- the use of communicating word processors would be an important aid to communications in countries with poor postal services;

9) Current microcomputer, graphics and printer technology is capable of providing low-cost computer terminals and systems operating in any of the languages of the world, including any combination of them;

10) The case history of Chinese printing given in this paper shows that the technical feasibility can be made actual, simply and rapidly;

11) The residual problems of developing software systems such as databases, compilers and accounting packages in each language of the world are very much more severe, probably impossible if tackled piecemeal;

12) It is suggested in this paper that much can be achieved through the use of purely syntactic transformations in front-end systems translating an exotic character set and language to serve as input to packages operating in the Latin character set and designed for the English language;

13) This would be simple to do if all dialog with the system was table-driven from separate well-defined text data structures -- this is already done in Canada for some packages designed to operate in both French and English and is a useful design discipline for any package.

Finally, let me remark that Canada seems extremely well positioned to undertake systems development of "universal language" systems as described here. The need for systems to operate in the two national languages has already created an awareness of the problems described. The graphics technologies necessary to cope with exotic languages are well developed in Canada. There is in the population a mixture of people of all nations who can provide locally the knowledge on the idiosyncracies of each language. To make computing universally available to all nations in all languages is a fascinating challenge. It can be seen as both a service to the world and as offering major commercial opportunities. There are already many individual research groups and companies tackling specialist parts of this problem. I suggest it is one where the co-ordination of effort would lead to very rapid progress.

ACKNOWLEDGEMENTS

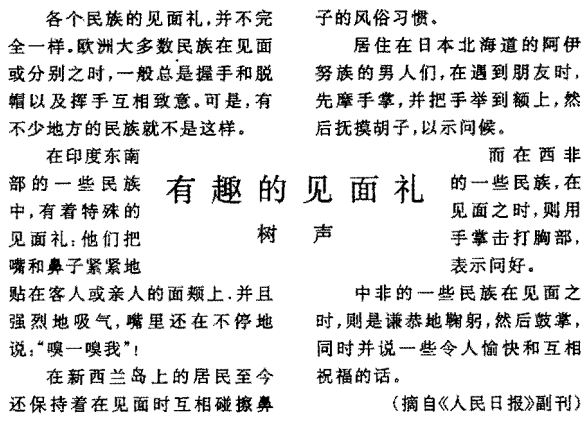
The Chinese system development was undertaken jointly between Monotype Corporation and the China Printing Corporation. It involved many people in both organizations. The Machine Translation Unit of the Chinese University of Hong Kong provided the initial keyboard design.

REFERENCES

Gaines, B.R. (1981) "Unravelling the Chinese typesetting puzzle." **Penrose International Review of the Graphic Arts** 73, 25-40.

Seybolt, P.J. & Chiang, G.K. (1979) **Language Reform in China.** New York: M.E. Sharpe.

Witten, I.H. & Ng, Y.H. (1981) "An ideographic language front end processor for accessing English language systems." **Computer Journal** 24(1) 62-70.


有趣的见面礼
树声

各个民族的见面礼，并不完全一样。欧洲大多数民族在见面或分别之时，一般总是握手和脱帽以及挥手互相致意。可是，有不少地方的民族就不是这样。

在印度东南部的一些民族中，有着特殊的见面礼：他们把嘴和鼻子紧紧地贴在客人或亲人的面颊上，并且强烈地吸气，嘴里还在不停地说："嗅一嗅我"！

在新西兰岛上的居民至今还保持着在见面时互相碰擦鼻子的风俗习惯。

居住在日本北海道的阿伊努族的男人们，在遇到朋友时，先摩手掌，并把手举到额上，然后抚摸胡子，以示问候。

而在西非的一些民族，在见面之时，则用手掌击打胸部，表示问好。

中非的一些民族在见面之时，则是谦恭地鞠躬，然后鼓掌，同时并说一些令人愉快和互相祝福的话。

（摘自《人民日报》副刊）

Figure 1 Sample of Lasercomp Output

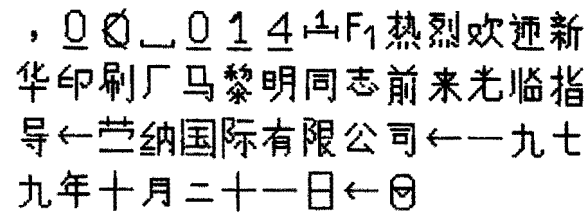


Figure 2 Sample of Line Printer Output

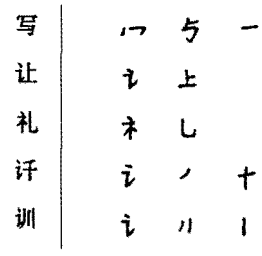


Figure 3 Character and Keyboard Sequence