IDRC-239e

Chinese-Character Processing for Computerized Bibliographic Information Exchange

Summary Report of an International Workshop Held in Hong Kong, 17–20 December 1984

Proceedings Series

INTERNATIONAL DEVELOPMENT RESEARCH CENTRE • CENTRE DE RECHERCHES POUR LE DEVELOPPEMENT INTERNATIONAL

CANADA

Postal Address: Box 8500, Ottawa, Canada K1G 3H9
Head Office: 60 Queen Street, Ottawa, Canada

Ting, T.C.

IDRC-239e

Chinese-character processing for computerized bibliographic information exchange : summary report of an international workshop held in Hong Kong, 17-20 December 1984. Ottawa, Ont., 1985. 68 p. : ill. (Proceedings series / IDRC)

/Information processing/, /data bases/, /Chinese language/, /information exchange/, /standardization/, /computer programmes/ - /Universal bibliographic control/, /conference report/.

UDC: 681.3:495.1 ISBN: 0-88936-441-9

Microfiche edition available

IDRC-239e

Chinese-Character Processing for Computerized Bibliographic Information Exchange

Summary Report of an International Workshop Held in Hong Kong, 17–20 December 1984

Editor: T.C. Ting

Sponsored by the
International Development Research Centre
and the
Chinese Language Computer Society
and hosted by
the University of Hong Kong

The views presented here are those of the participants and do not necessarily represent those of their institutions or the International Development Research Centre.

ABSTRACT

This publication summarizes the discussions at a 4-day international workshop on Chinese-character processing for computerized bibliographic information exchange held at the University of Hong Kong in December 1984. The workshop covered both general topics concerning Chinese data bases and international information exchange as well as specific issues of character coding, input and output methods, regional and international standardization, and software development. The workshop provided a forum for technical discussions among researchers and systems developers from various countries and it is hoped that it will provide a sound basis for promoting and stimulating international scientific and technical information exchange on Chinese-character processing.

RÉSUMÉ

La présente publication résume les discussions qui ont eu lieu lors d'un colloque international de quatre jours, en décembre 1984, à l'Université de Hong Kong, sur le traitement des données en chinois dans l'optique de l'échange d'informations bibliographiques informatisées. Furent abordés autant les sujets généraux, tels les banques de données chinoises et les échanges internationaux d'information, que les questions particulières, tels le codage des caractères, les méthodes d'entrée et de sortie, la normalisation régionale et internationale et la mise au point du logiciel. Le colloque a suscité, parmi les chercheurs et auteurs de systèmes de divers pays, des échanges techniques. Il est à souhaiter que cette réunion soit un point de départ valable pour promouvoir et encourager les échanges scientifiques et techniques internationaux sur le traitement des données en chinois.

RESUMEN

Esta publicación resume les discusiones de un taller internacional de cuatro dias sobre procesamiento de textos en caracteres chinos para el intercambio de información bibliográfica computerizada, celebrado en la Universidad de Hong Kong en diciembre de 1984. El taller abarcó tanto temas generales sobre las bases de datos chinas y el intercambio internacional de información, como aspectos específicos de codificación de caracteres, métodos de alimentación y salida, normalización regional e internacional y desarrollo de programas. El taller fue un foro para las discusiones técnicas entre investigadores y diseñadores de sistemas provenientes de varios países. Se espera que esto provea una base sólida para fomentar y estimular el intercambio internacional de información científica y técnica sobre el procesamiento de los caracteres chinos.

CONTENTS

FOREWORD

In the past decade, many of the basic problems of software design for computerized systems handling bibliographic and other types of information were resolved. They recurred, however, in somewhat different forms, when one looked beyond most relatively simple Western character sets. Even these difficulties have also been overcome, for the most part, and complex ideographic languages like Chinese have proven amenable to computerized processing. Modular system design and specific technical advances, such as intelligent microprocessor-controlled terminals, have made it possible to isolate and solve these problems. Of course, a good deal of work still remains to be done but, through experimentation, technical problems can be resolved and the appropriate tools constructed.

Data bases are built to be used but they must be shared if they are to provide maximum exposure and efficient utilization of the information resources contained in them. For this, appropriate standards must be developed and tested. It was to promote the open exchange of views on these and related technical topics that this workshop was held.

Some of the basic issues discussed during this workshop, such as those related to standards, may be of interest to information workers who do not have a background in Chinese. This publication therefore contains a brief profile of the Chinese language in Appendix 3 to assist such readers in following some of the technical discussions.

The International Development Research Centre (IDRC) is pleased to have had the opportunity to support this workshop. It is hoped that the results will ultimately lead to better bibliographic control of, and easier access to, Chinese-language literature for scientific, technical, and socioeconomic development purposes.

The Centre would like to express its appreciation to the University of Hong Kong for providing its conference facilities and such a hospitable environment for the participants. It would also like to thank the Chinese Language Computer Society and its members for providing valuable inputs. Special thanks are due to T.C. Ting for organizing and chairing the workshop and to Sally Tan for providing logistic support.

Robert Valantin
Associate Director
Information Sciences Division
International Development Research Centre

OPENING SESSION

T.C. Ting, General Chairman

This international workshop will deal with such a narrow theme that a 4-day period may hardly appear to be justified. However, the theme is important and covers many unresolved issues that require in-depth study. The number of participants at the workshop has been kept small deliberately so that ample time will be available for each participant to express his or her views on the various topics. This is, indeed, a workshop and not a conference and no papers will be presented. Instead, we will together identify and select discussion topics, exchange ideas, and report research results and work experience in a somewhat informal and unstructured manner: in other words, we will have a 4-day brainstorming session. To produce a record of the workshop, the person chairing each session will work with the session recorders to produce a report for the "proceedings."

The scope of the main discussion of each session will be limited to technical aspects of the issues only. However, small group discussions are possible after the main sessions so that specific details can be addressed.

The main thrust of the workshop is to promote international information exchange with particular emphasis on Chinese scientific and technical data bases. The main challenge of the workshop is the technical difficulties of Chinese-character processing. Four general topics have been identified and each will occupy a half-day session. In addition, we will look at and discuss two case studies in one session and have a final closing session for more general discussion. On a less academic note, we will have a demonstration of several computer input systems here at the University of Hong Kong and a demonstration of the MINISIS system at the Hong Kong Productivity Council's offices.

In recent years, almost 400 input and output methods have been proposed for Chinese text material and, although research and development endeavours concerning the processing of these materials have been very active, the problems are far from being resolved. Many independently developed Chinese data bases and information systems exist using different coding methods and formats in different countries. Therefore, international cooperation must be encouraged to promote international information exchange. It is extremely important from the viewpoint of scientific and technological development that we should avoid working in isolation where we may reinvent what has already been invented or develop incompatible systems.

This workshop provides a great opportunity for us to develop general technical guidelines and suggestions that systems developers in different countries can follow. It is an important initial step for informal and informative scientific and technical discussion. These discussions can provide the needed technical solutions for promoting and encouraging international cooperation in bibliographic information exchange.

I look forward to working with all of you for a successful and rewarding workshop. Let me take this opportunity on behalf of all the participants to thank our sponsors, the International Development Research Centre and the Chinese Language Computer Society, and our host institution, the University of Hong Kong, for their support and assistance.

CHINESE DATA BASES FOR INTERNATIONAL INFORMATION EXCHANGE

Chairman: Alan Tucker
Recorders: Andrew Wang, Chorkin Chan, and Wellington Yu

The term "Chinese data base" can be defined as a database containing records for materials that were written in whole or in part in languages using Chinese characters -- that is, Chinese, Japanese, and Korean (CJK) languages -- and entered in the original characters. One of the data bases that has integrated such records is the Research Libraries Information Network (RLIN) developed by the Research Libraries Group (RLG).

RLG is a consortium of more than 30 major research universities and independent research institutions in the USA that provides support for their research and teaching through a variety of programs of which RLIN is one. RLIN is a nation-wide computer network providing support for libraries' technical processing -- acquisitions, cataloguing, and interlibrary loans. It has other features, such as collection analysis, and its database now contains some 16 million bibliographic and authority records.

Three major factors shaped the enhancement of RLIN to support the processing of records for CJK materials: economics, integration, and standardization.

Although 7 of the 10 largest East Asian collections in the USA are in RLG member libraries, East Asian libraries have very little support for equipment purchases. Thus, it was necessary to develop a single terminal to handle all three CJK languages, rather than to have separate terminals for each.

Only about 10% of the terminals on the RLIN network can input and display CJK data, but records retrieved from the data base must be usable by a researcher at any terminal. Moreover, a search for materials on a specific topic, for example, 19th century Chinese agriculture, ought to retrieve all relevant records from the data base, regardless of language. Therefore, changes have had to be made to the USMARC format (MARC = Machine-readable catalogue) to allow for the inclusion and specific identification of all non-Roman character sets (not just CJK), and for the recording of different graphic representations of the same data in parallel fields, e.g., a title in Chinese characters and, in a separate field, in Roman characters.

Although MARC is a uniform system, variants have been developed, for various reasons, by different organizations. For example, the OCLC Online Computer Library Center (OCLC), which provides computerized information services to over 6000 libraries in 10 countries, stores its 12 million unique bibliographic records in OCLCMARC, which is about 99% identical with USMARC, also known as LCMARC (Library of Congress MARC).

OCLC's data base, which is growing at the rate of 1.5 million records per year, does not contain CJK records at the moment; however, these should start being added in about 1 year. The Agricultural Science Information Center (ASIC) in Taipei, China, uses UNIMARC (universal MARC) to store about 16,000 serial records that contain abstracts in both the Chinese and English languages -- this data base also contains about 17,000 terms in a Chinese thesaurus. The University of Hong Kong is now using UKMARC (United Kingdom MARC) but intends to load LCMARC records in 1985 and the Institute of Scientific and Technical Information of China (ISTIC), although only planning its automatic services, will probably adopt UNIMARC.

Standardization is necessary so that exchange of data among major bibliographic institutions in East Asia, such as the several national libraries, is possible. Thus the CJK character set must be compatible with existing national standards. However, for compatibility to be achieved, we need to know how many and which specific characters are involved.

The RLIN system uses roughly 14,000 Chinese characters and, after more than 1 year of operation during which some 50,000 items have been catalogued, a list of only about 20 unavailable characters has been compiled. The system at ASIC uses about 30,000 characters; however, the materials covered by the Center are more specialized than those in the broad, largely nonscientific, coverage of RLIN. The Chinese Character Analysis Group (CCAG) in Taipei, China, has so far identified about 50,000 characters, of which about 30,000 are independent characters, and the remainder are variant forms.

The RLIN character set -- known as the RLIN East Asian Character Code (REACC) -- was developed from five other sets. Four of these are official or nominal standards (China, Japan, Korea, and Taiwan China) and the fifth was the set that had been developed earlier for the Chinese-character terminal on which RLG's CJK terminal was based. REACC is structurally identical to CCCII (Chinese character code for information interchange, the standard developed in Taiwan China) in that it uses three bytes to represent each character, and it maintains the relational coding of CCCII that makes explicit linkages between traditional, simplified, and other variant forms of each character.

Different types of multibyte character representations -- two bytes or three, fixed-length or variable -- have cost and benefit implications. The change from two to three bytes for each character increases the cost of transmission in proportion to the number of elements within the records that require three bytes. However, bulk transfer of bibliographic data among, for example, national libraries currently takes place on the extremely inexpensive medium of tape and will probably continue to do so for some time. Nonetheless, protocols for packaging and transmitting multibyte data is an important issue, although separable from the underlying code.

Two of the coding systems now in use, CCCII and REACC, are identical in structure. That is, in each system three bytes are used to represent each character. Basically, the first (i.e., high-order) byte identifies one of 94 "planes," where each plane consists of a 94 x 94 matrix (c.f. ISO 2022), while the second and third bytes are the x and y coordinates of a position on that plane. Characters are arranged on the planes such that, if a given Chinese traditional character, say, C_T, is coded on plane 21 as 21/37/25 (in hexadecimal), then a simplified form of that character, C_S, would be coded as 27/37/25, another variant form, C_{V1}, at 2D/37/25, and so on in increments of six. There is thus a predictable relationship between the codes for various forms of any character.

The systems differ, however, in their content, in that not all the characters present in CCCII are known to REACC. The former now consists of nearly 50,000 characters, of which some 22,000 are already available in machine-readable form as part of the Chinese character data base (CCDB), whereas the latter now recognizes about 25,000 characters including the 14,000 now available on an RLG CJK terminal. As RLG and CCAG continue to collaborate and exchange data, it is hoped that within 12-18 months (i.e., early 1986) the two systems may become identical in content as well as in structure.

In comparing two- and three-byte coding schemes, RLIN's use of a three-byte code can perhaps be justified in two ways. First, RLIN uses the relational information contained in the longer code to permit a user to search for a word and retrieve instances of its use in which only the exact form of the character specified in the search argument occurred, or, alternatively, instances in which any form of the character occurred. Although the three-byte code could be mapped to a two-byte value for processing purposes, the three-byte code makes this sort of manipulation more efficient. Second, within the context of the RLIN data base, where most records do not contain multibyte characters and where CJK records constitute a very small proportion of an extremely large (about 20-gigabyte) file, the overhead is insignificant in terms of storage costs. In data bases in Asia, however, the same argument might not be valid.

The time has come for the adoption of a standard character set that could be used for exchange of bibliographic data between nations, each of which has its own national standard. Such a character set would facilitate and simplify exchange by reducing the number of different code translations that each participating agency has to perform. However, certain countries may be unwilling to subscribe to a standard using a three-byte code because of the perceived extra costs involved. It has been suggested that the control codes included in the coding schemes of some existing standards may be inadequate for some purposes and any proposal for a new standard might need to start with an examination of that part of the coding system. Exchange of CJK data within the USA will soon begin, when the Library of Congress starts to distribute CJK records via its tape distribution service, and there is some concern that the encoding should be standardized, probably by proposing -- as a first step -- the adoption of REACC as a U.S. standard.

CHINESE CHARACTER INPUT AND OUTPUT METHODS

Chairman: C.Y. Suen
Recorders: Lawrence Tam and Victor Li

Two basic approaches to the topic of input and output of Chinese characters are possible: one is methodological, which is concerned with the structure and phonetics of the topic, and the other is holistic, which is related to the economics of the issue. However, input and output must be considered separately -- indeed, it is possible to use different input and output systems -- and the middle system, the internal codes, forms a bridge.

Because there are hundreds of input and output methods (see Figs. 1 and 2), our main concern is how to input and output fast and at low cost. This is a subject of intensive research.

Input Methods

For character input (Fig. 1), there are basically two methods -- sound and shape of character -- but if these two are combined, a third is created. Although sound is used as a method of input, the sound of character is sometimes difficult to identify. Therefore, spelling of the sound, for example, using Pinyin, either in full or using the initial and final letter of the sound can be used. Of the two, the second is more economical.

Phonetic symbols can be used to represent sounds of Chinese characters, and quite a variety of systems exists, e.g., JIFH, IPA, Pinyin, Yale, Suen 1979, and Suen 1983 system (Suen 1979, 1983a). However, there are problems related to the use of these systems. In Chinese, there are about 1200 syllable-tone combinations representing about 50,000 Chinese characters. Therefore, there is no one-to-one mapping between the shape

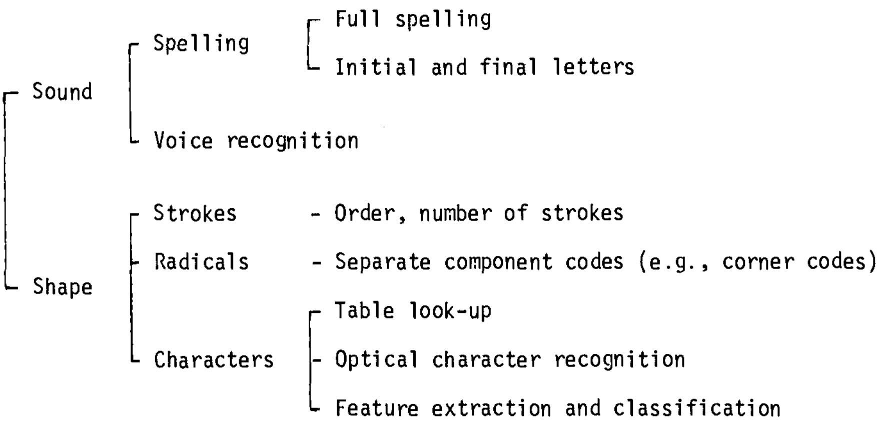
Sound
Spelling
Full spelling
Initial and final letters
Voice recognition
Shape
Strokes
- Order, number of strokes
Radicals
- Separate component codes (e.g., corner codes)
Characters
Table look-up
- Optical character recognition
Feature extraction and classification

Fig 1. **Classification of Chinese-character input methods.**

and the sound of characters. For comparison, it is difficult to spell out or represent sounds represented by the Roman letter e. Equally, because many characters have the same sound, purely sound input will produce frequent duplications. One way of resolving this is to display all the potential characters on the screen, and the user then selects the correct one. This is, however, a slow process. Another way to speed up the process is to use the initial and final key strokes of the character's pronunciation. For example, xiang (香) requires five key strokes, but if the initial and final letters (x and g) are used to represent the character, there is a saving of three key strokes.

Another input method is voice recognition. Because Chinese has only about 400 syllables, a relatively small number, voice recognition by computer is feasible. If tones are added, the chance of "collision" or duplication becomes much smaller although there would then be about 1200 syllables. Even with voice input, there are problems: differences in individual voices and the mood of the speaker, noise in the environment, and talking speed of the speaker. However, if the system is programed to recognize a particular voice, the machine becomes the slave of that human and voice input can be a success.

Problems still occur because many characters have the same sound: for example, ji and ch, rj and rc, and dz and ts are difficult pairs for the machine to recognize (Suen 1983b). Context should make recognition more specific. For example, (大) would be easier to distinguish in the combination (大人) than alone because of the context.

Equally, there are differences between single characters and words. Research on isolated utterance and continuous speech is quite different because segmentation of continuous speech is difficult.

Japanese is a much easier language to segment for voice recognition because only about 10 vowels are involved. If Japanese is pronounced syllable by syllable, it is quite possible to input Japanese vocally but this slows down the input process and stressing the sound is tedious for the operator.

The second main method of input is by the shape of the character. Again, there are three methods: by classification of stroke, by radical, or by the character itself. As an extreme example, the word (籲) has 32 strokes but, on the average, each Chinese character has only 8-12 strokes. There are many ways of breaking up the Chinese character, e.g., from top to bottom or from left to right, etc., depending on the structure of the character and these may be combined.

Chinese is represented by radicals, about 700 radicals represent the entire language; however, 15 basic structures would probably be sufficient (Suen and Huang 1984). With these 15 basic structures, the 4600 most frequently used Chinese characters can be accommodated. Also, with these 15 structures, a maximum of four key strokes are needed to input a character. However, when frequency of occurrence is considered, only 2.1 key strokes are required to input a character.

The last method of input is by whole character. This may be done by "table-look-up" and by character recognition. IPX is an example of table look-up. In optical character recognition (OCR), the purpose is to let the machine do the work. However, a primary concern with this is differences among fonts, particularly on the style of characters, and the differences in printed and handwritten characters. For example, there are two different forms for each of the following three characters:

(回回), (来来), (青青).

Standardization of fonts is important. The OCR reader at Concordia University can recognize 3000 characters from a single font (Wang and Suen 1984). Some work on font development has been done, or been proposed, but the requirements for OCR and human reading are different. Therefore, more research is needed.

Some efforts have been made to compare systematically the pros and cons of the various input methods. Each method of input has its strong and weak points: for example, if radicals are used, a different type of keyboard is needed to that required for stroke analysis. However, research in Japan has shown that the most mechanically satisfactory method may not be user friendly. As word processors have been introduced into offices, the problems of fatigue and occupational disorders have appeared -- in some cases, months after operators started using the system for long hours. The way we use our brains must be studied because the right celebral hemisphere is used for pattern recognition, which is an element of input for Chinese, Japanese, and Korean (CJK) languages. Therefore, design of input methods that are right hemisphere dominant will create user-friendly systems for professional operators. However, professional operators must be distinguished from casual users. Professional operators can be intensively trained and different types of input should be designed for different types of users. Tests by the Institute of Information Industries (III) have found the radical input method more acceptable to some users (Chen and Gong 1984). It may be more important to look at factors other than the speed of input alone. The end product may well be determined by the market place -- if a manufacturer can sell a large number of machines designed for casual users, against small numbers for professional data-entry people, they will design for the larger market and this generally means a keyboard that is small, simple, and easy for casual users. Perhaps the equipment should be adaptable to both types of users and a holistic approach for the design should be adopted.

To sum up, there are hundreds of methods for input. The questions are which is the best, the fastest, and the most accurate? What costs are involved? What memory is required? And, finally, which method is most user friendly? As computers become more powerful, it is possible that personal computers may be able to process Chinese. Equally, it is possible that input methods will be matched to the user, or that individual users can select his or her own input method. These are areas of development that require further research, as do the fields of intonation and connected speech for spoken input and studies of the basic structures of Chinese characters.

Output Methods

Two basic types of output can be considered, character and voice (Fig. 2). Character output can be further divided into display and printout for each of which the requirements are somewhat different.

Voice output has several potential uses, for example, in airports for messages and in computer-aided instruction (CAI). For bibliographic material or other bulk-data entry, an audio system will relieve occupational disorders by giving immediate feedback for input checking. Audio output is even being used in supermarket checkouts and has potential in typesetting as a form of proofreading.

The 400 or so syllables in Chinese can be computer synthesized for audio output. Speech synthesis is done by generation of syllables through digitization, linear predictive coding, or Fourier transform, which offers considerable savings in terms of memory. It is possible to combine format frequency, synthesis of sound, and digitization to produce spoken Mandarin so that voice can be a realistic form of output.

The requirements of screen display and printed output differ somewhat. Although a 16 x 16 matrix will produce acceptable hardcopy, 24 x 24 would be better for the screen. For example, the resolution from 16 x 16 is not detailed enough for such words as (轟). Both display and printout are developing and advanced hardware will soon reduce the problems -- therefore, it

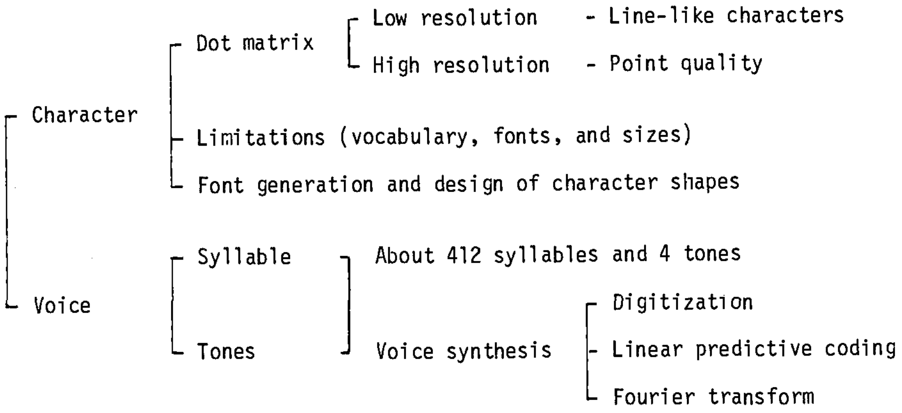

Fig. 2. Classification of Chinese-character output methods.

would be unwise to set standards prematurely. Equally, however, it is not necessary to accept what is now available: technology should serve human needs, not vice versa. If, for Chinese-character output, a larger screen is needed to produce greater resolution, it should be developed. However, we should not be totally committed to the cathode-ray tube (CRT), which could well be phased out in as little as 5 years, with "liquid crystal" screens as a replacement.

Printout of characters can be done on a 16 x 16 dot matrix but, for high quality, a 24 x 24 or even 48 x 48 matrix is needed. The higher number matrix eliminates "zig-zag" effects created if slanted, italic-type, output is created. However, dot density (number of dots per unit area) will also have an effect on resolution. Laser printers have a very high resolution, over 300 dots per inch, compared with more readily available systems that have about 200 dots per inch. The shape of the dots -- they are now usually eliptic -- will also affect the resolution of the output.

Although output of 5000 or 50,000 characters through a dot matrix dictionary can be considered, the dictionary. Therefore, a data base plus a program to synthesize characters from basic components will have to be installed to create new characters, i.e., (植字).

However, if generation of a character such as (植字) is allowed, it will create problems in information exchange. The Chinese character code for information interchange (CCCII) already has some sort of standard proposed but this is for internal representation only. Character generation in Chinese is complicated but characters built up in the data base will form a useful tool in developing new characters. However, creating a new character for a entity or concept could create problems: for example, in naming the newly discovered subatomic J-particle, what character should be used and who should control the selection of the character?

At the moment, there are four character sets for CJK languages, i.e., GB 2312-80 in China, JIS C6226 in Japan, KSC 82 in Korea, and CCCII in Taiwan China. CCCII has a set of nearly 50,000 characters that would probably satisfy all present needs. However, the Chinese language is dynamic and character creation is, and should be, open; although stylistic (typographic) variations should not be treated as new characters.

Linked to an inexpensive printer, some new software developed by the Nippon Electrical Company (NEC) will produce output by components. The strokes of the character are produced with elliptic and circular mathematic functions that specify the position and the size of the components and result in highly satisfactory characters. Internally, the output of Chinese words is represented by components using binary tree structure. To produce the total of 34 strokes and 1000 components, 100k bytes of memory are used and, at the moment, about 20,000 characters can be built. To produce an additional 10,000 characters, another 28k bytes would be used.

Conclusion

Several topics related to input and output of Chinese language material require much more study. They can be grouped as: methods, memory, data structure, retrieval, speed, cost, and efficiency. For data input at least, the ideal solution would be to have several input methods available so that operators can choose whatever is best for them in their applications.

ISSUES OF MIXING CHARACTER SETS

Chairman: Hisao Yamada
Recorders: L.-B. Kan and Wellington Yu

As well as a large set of Chinese characters, texts in nonalphabetic writing systems such as Chinese, Japanese, and Korean (CJK) contain alphanumerics and other characters. When such a text is stored or transmitted through a communication medium, the representation of characters, called character codes, is necessarily a mixture of alphanumeric codes and CJK character codes. Structural definition of these codes is separated: generally alphanumerics first, then the CJK characters. Therefore, a text of mixed modes results in less uniform structure and less consistency than texts with alphanumerics only.

Alphanumeric characters have a fairly well agreed code structure whose framework has been standardized by the International Organization for Standardization (ISO). It is usually represented as shown in Fig. 1, where each character consists of seven bits, and columns are represented by the most significant three bits (eight columns) and rows are represented by the least significant four bits (16 rows).

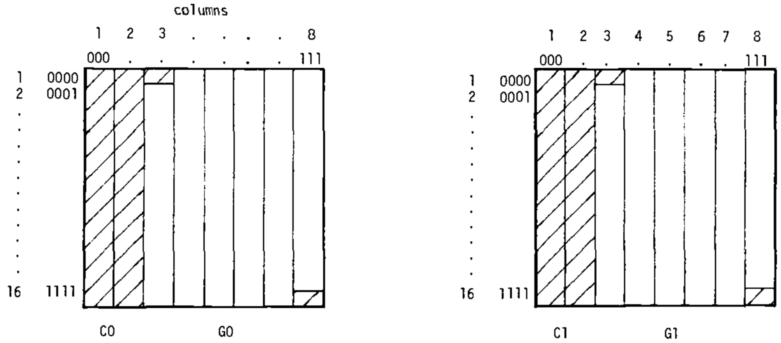

Fig. 1. Seven-bit, one-byte code (shaded areas represent the control characters).

All the 32 code points of the first two columns (16 x 2) plus those in the first row of the third column and the last row of the eighth column are assigned as control characters (C0, a total of 34 code points). The remaining 94 code points are assigned to the graphic characters (G0). When a group of these 128 (34 + 94) code points (C0 + G0) is not sufficient, two control codes, SI (shift-in) and SO (shift-out), are used to switch into and out of a second group (C1 + G1) of 128 code points (right side of Fig. 1). Not all of the 128 code points are usable because of the control code conflict (e.g., SI and SO in C1 are the same as SI and SO in C0). Therefore, only 94 code points in G1 are readily usable for graphic characters in the second group. As long as we stay in one alphanumeric language, this arrangement of G0 and G1 is usually sufficient. However, if we take the aggregate of all alphanumeric languages, including Arabic, Devanagari, Thai, etc., then a much larger set is needed.

For such a case, the above control code C0 includes an escape code, which, together with the following character sequence, will make up a very flexible escape sequence (ISO 2032). Therefore, it allows one to get into any other separately defined character code table (with appropriate restrictions).

Chinese Character Codes

The character code set-up that is used for an alphanumeric writing system is not able to represent a character set such as Chinese because it contains at least 50,000 items. Therefore, any writing systems that include a subset of Chinese characters, such as Japanese and Korean, must use a multiple-byte representation code.

For example, the Japanese code system is defined as follows. By taking the 94 graphic codes of G0, and using one each for row and for column, we have a total of 8836 code points (94 x 94): this which is illustrated in Fig. 2. Therefore, by using two seven-bit bytes, we are able to represent up to 8836 characters. The Japan Industrial Standard (JIS C6226) character set for Japanese uses such a representation. In addition, Japan also used two other similarly defined code

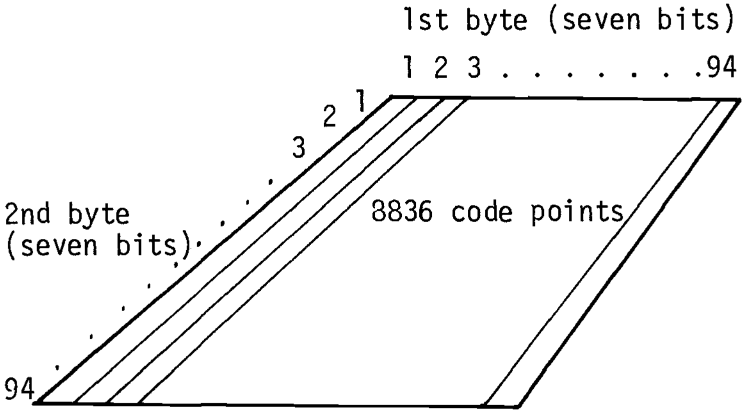

Fig. 2. JIS C6226 code plane.

systems: CO 59 codes for newspaper and communication services and CO 77 codes for Kyodo Tsushin. Other systems with similar structures are the GB 2312 standard and the Tung Yung code for Chinese characters and KIPS for Hangul and Chinese characters.

The most comprehensive code structure for Chinese characters is the Chinese character codes for information interchange (CCCII) proposed by the Chinese Character Analysis Group (CCAG). It uses three seven-bit code spaces for characters, as shown in Fig. 3. However, it is more convenient to display this code plan in a layer structure (Fig. 4) to show the interrelationships among character codes.

The CCCII code system has 53,019 code points per layer (8836 x 6) and 15 layers are potentially available for characters. The 16th layer consists of the remaining four planes, which are reserved for special characters such as graphic symbols, nonalphabetic Western languages, etc.

The appealing feature of the CCCII system is that one layer (the top) contains the standard traditional forms of characters. The 2nd-15th layers are used for variant forms and, as it stands now, the first six layers are for Chinese character variants, including GB 2312-80 in the second layer, and remaining layers (7-15) are for other standards such as JIS C6226, and, in the future, KIPS, etc.

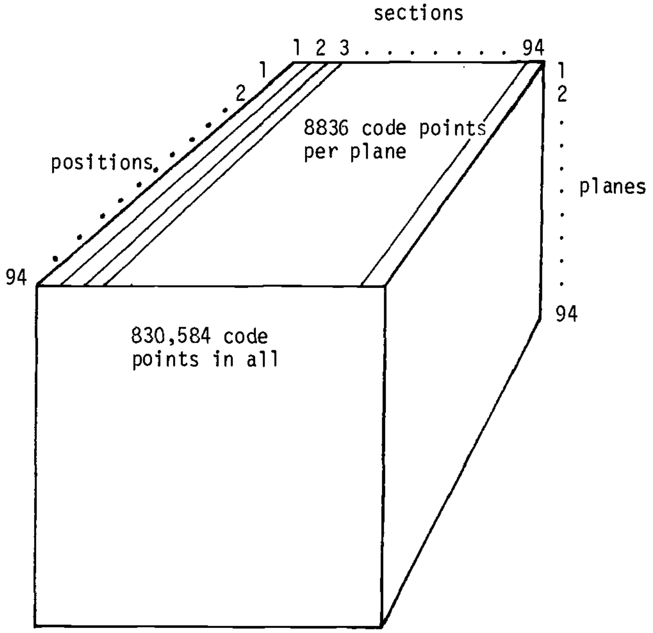

Fig. 3. Code point space for CCCII.

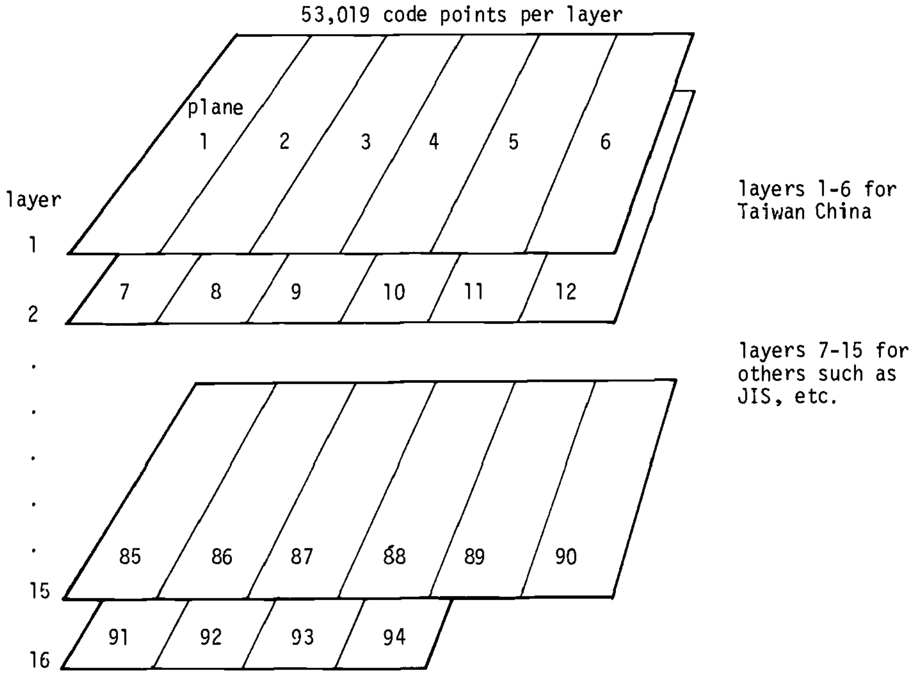
53,019 code points per layer
plane
1
2
3
4
5
6
layer
1
2
7
8
9
10
11
12
15
16
85
86
87
88
89
90
91
92
93
94
layers 1-6 for Taiwan China
layers 7-15 for others such as JIS, etc.

Fig. 4. CCCII layers.

Furthermore, the variants of a particular standard character in the lower layers are placed directly below the standard character in the same coordinates on lower planes. Therefore, a sort of symmetry exists among the layers and this framework can accommodate the coding system in other geographic regions just by interchanging the layers, e.g., for Japanese use, JIS C6226 could be brought up to the first layer and the coordinates in the layer rearranged into the JIS code sequence. Thus, if the three-byte code structure of a traditional character is B_3 B_2 B_1 then the code structure for the variant character will be B_3' B_2 B_1 with only the leading byte (B_3') being different. Hence the relationship between the characters can be established easily.

The disadvantage of this plan in reality is that JIS C6226 and GB 2312-80 are already established and in use, and B2B1 of their characters are different from B2B2 of the variants in the proposed CCCII, destroying such would-be nice vertical correspondences, unless all code systems are reworked from scratch, which is unlikely.

International Information Exchange and Transmission Codes

With the increasing need for international information exchange and the developing technological feasibility to support it, we are rapidly approaching the point of establishing a worldwide information interchange network for bibliographic data bases.

To define codes so that more than 8836 characters can be used, more than two seven-bit bytes are necessary (Fig. 2). However, when transmitting a large quantity of information, it would be wasteful to use three-byte codes for all characters to be transmitted. This fact is particularly true for transmission within the USA, where the proportion of Chinese characters intermixed with alphanumeric information is, on the average, small and only one byte is needed for alphanumerics.

Even within the CJK character sphere, by far the majority of characters are from the basic set, which is well within the two-byte representable code set size of 8536. For example, 5 years ago in Japan, the Japanese Diet Library (JDL), Japan Information Center for Science and Technology (JICST), Nippon Keizai Newspaper (NKS), Nippon Telegraph and Telephone Corporation (NTT), and Japan Patent Information Center (JPIC) standardized the specification of a public Kanzi (Chinese character) terminal for common use for public information services. This was based on the two seven-bit byte JIS codes (Anon. 1979, 1980).

With the rapid spread of office automation, aided by the commercial production of Japanese word processors and the initiation of digital communication services, transmission of various documents through electronic transmission media is becoming increasingly popular, and Japan Electronic Industry Development Association (JEIDA) and Japan Business Machine Manufacturer's Association (JBMA) are working on various industry-wide standardizations (JBMA 1983; JEIDA 1983, 1984). They have also standardized on the two seven-bit JIS character codes for transmission.

In the short-term, the worldwide demand for information transmission is expected to increase exponentially and it is desirable that the cost of transmission be kept down. The engineering cost of transmission must and will come down but, at the same time, the transmission code structure must be designed to keep the cost down by being efficient. Coding theory indicates that the variable length code is inevitable for an efficient code. That is, no matter what code system is used, the bulk of transmission must be in short codes and only infrequently appearing characters should be transmitted by shifting into longer code structures.

CCCII's approach is different and takes three seven-bit codes and maps them one-to-one onto 16-bit codes by radix 94 conversion. This allows the representation of more than 50,000 characters in two eight-bit codes, with a minimal loss of efficiency.

Regional Standard Codes and Worldwide Information Exchange

It goes without saying that if the whole world got together and worked out a comprehensive standard coding system, the problems would be simplified, and also that the code could be somewhat more efficient. However, because of differences in regional requirements, a unified, single-code system for the entire world would not be the most efficient even for code-space definition, let alone for text storage and transmission.

As it stands now, several different coding schemes may be in use within a specific region, and even within a maker community. Because regional wisdom has been exercised, even if somewhat loosely, these regional codes are generally fairly efficient, although not optimal, at least for text storage. As soon as information interchange is required, however, the lack of standards will present a problem even if it is restricted to within a region, because different machines cannot communicate directly and code conversions (CC) are needed (Fig. 5).

At present, more than a dozen manufacturers in Japan are commercially supplying more than 50 different models of Japanese word processing systems -- most as stand-alone word processors. Within a single model set, documents can be interchanged through floppy disks, for example. However, different products of certain manufacturers do not necessarily use the same character codes, or file formats, and such direct information exchange is not always possible even within a single manufacturer's products. Instead, a conversion-code process has to be employed (Fig. 5, "Manufacturer A Community," upper left).

JEIDA has adopted the JIS character code as the standard and also established a document format code for interchange (JEIDA 1983, 1984), and all manufacturers are, in principle, to implement a code converter to the JEIDA standard. In the future, the JEIDA standard will be made, with appropriate modifications if necessary, into a JIS standard, and it will serve as the Regional Code for Japan. This situation is illustrated in the left half of Fig. 5.

Now, if a Regional Code is standardized from the outset in a certain region, code converters will not be

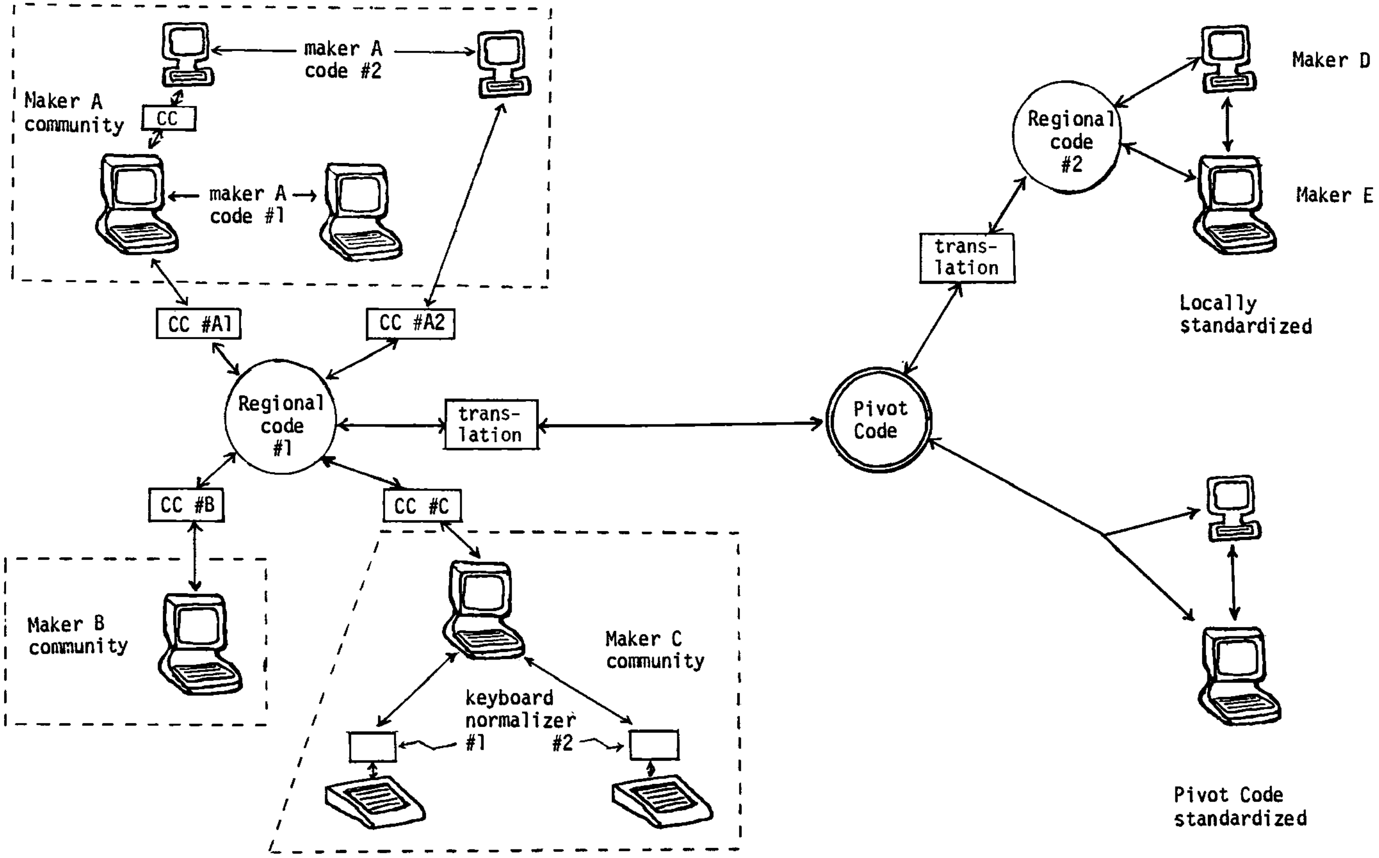

Fig. 5. Information interchange scheme.

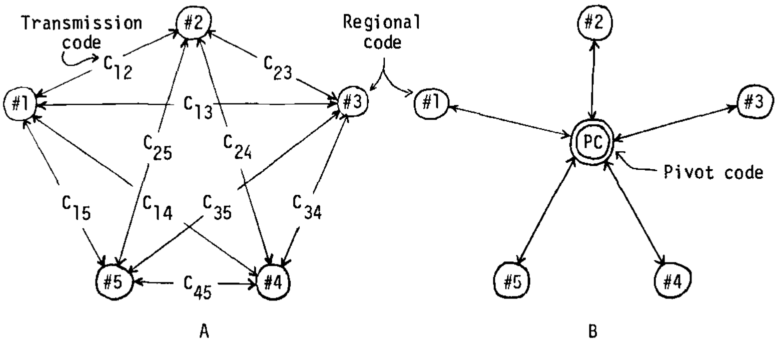

Fig. 6. Code translation schemes: (A) Direct translation (delta or complete graph translation), n(n - 1) translators are required; and (B) Indirect translation (star or pivot translation), 2n translators are required.

needed for code transmission and information exchange within the region (Fig. 5, upper right).

Once such regionally optimized codes are established in each geographic region, the next task is the establishment of the worldwide information-interchange facilities. Because the regional codes are optimized to serve the specific requirement of the community within the region, it cannot be hoped that regional codes would be identical. Hence, for worldwide information interchange, code translation among different regional codes is a necessity.

Two basic modes are possible for such translations (Fig. 6). Direct translation between individual regional codes requires a total of n(n - 1) common transmission codes for n regions, although not all of them will necessarily be different, to optimize interregional transmission (Fig. 6a). Alternatively, if a common transmission code, which may be viewed as a pivot code (PC), is used, then only 2n translators are needed. Where more than three regions are involved (i.e., $n > 3$), the pivot code system is less complex than direct translation. However, the pivot code is a compromise and will be less efficient for transmission.

All things being equal, it appears advisable to use the pivot code approach to keep the commonality of information transmitted for all regions. Otherwise, a single region must be prepared to process n - 1 different interchange codes. Fig. 5 is based on the common code concept for worldwide interchange and for code conversions within a single region, as seen in Region 1.

Tasks to be Undertaken

Writing systems that utilize CJK characters, and possibly other writing symbols in other regions, definitely require a larger character code space than those that are based on the Roman alphabet. Each region using such a writing system must have developed a character code system that adequately satisfies the regional need, and the mixing of different character sets is not a problem within the region. However, regional standards are not yet established for all regions.

Establishment of such regional standards is best left to the individual regions so that convenient and reasonably efficient code systems that satisfy regional requirements are developed. However, development of the pivot code for the worldwide information interchange is an altogether different matter. The pivot code should achieve an overall efficient transmission (and possibly storage for common data base) for the entire world. Yet no one region alone has sufficient knowledge of all the requirements of all regions for transmission (and storage) for the present and for future development. Therefore, any "standard" pivot code developed by a single region, purportedly for the entire world, cannot be expected to be optimal. Thus, it is strongly desirable that efforts to develop the pivot code should involve worldwide cooperation and that this cooperative effort should be present from the beginning.

Although we cannot go into the technical details of such a code system here, it appears that a system that covers the bulk transmission by two seven-bit codes (8836 graphic characters, c.f., Fig. 2) of the most frequently used characters would be sufficient. For less frequently used characters, we would simply shift into other space by the use of the escape sequence provision of the ISO standard. If this turns out to be

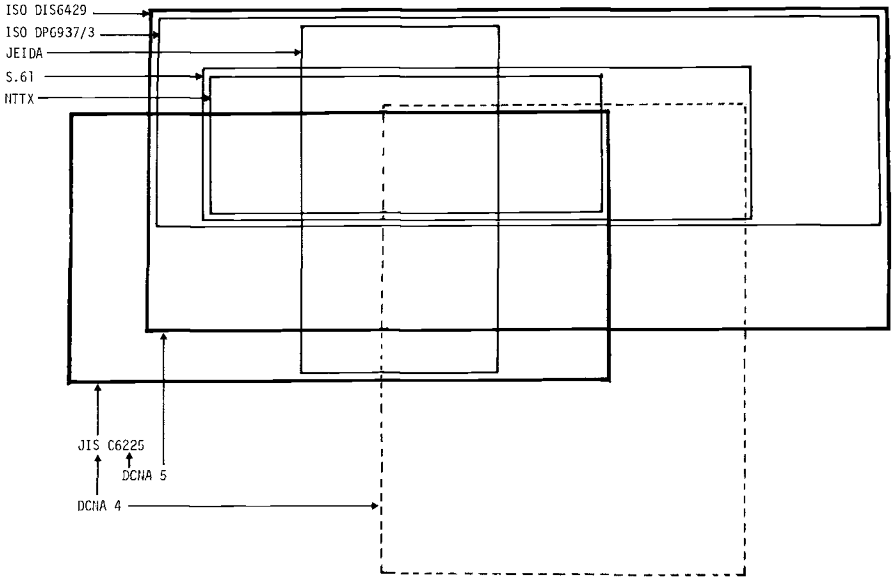

Fig. 7. Diversity and overlap of existing control codes standards.

inadequate for reasons that are not now apparent, the CCCII approach of two eight-bit representation will certainly give sufficient code space. Therefore, it is expected that the pivot code will be in a two-byte format for transmission efficiency, rather than a three-byte format.

The more serious problem is that of control code standardization. ISO provides a standard for the control codes of alphanumeric text information; however, CJK text transmission requires additional control codes simply because it employs a more complex character set. In addition, at present, there is no ISO standard for the transmission of document files, which consist of more detailed formatting of bulk text than the ISO standard addresses so far. The details are not important for the present discussion, but the overall divergence of some specific standards can be represented by Fig. 7.

It is important to realize that control codes for text format and document file format should be as uniform among the regional codes as possible and that this should be implemented as soon as possible. This is necessary because, unlike the character codes, there is a much lower diversity of regional requirements and, more importantly, because nonuniform regional control codes would make code conversion a much more complex task for the transmission system. This is especially true if character codes and control codes of one region "invade" the opposite code categories of another region, thus reducing the availability of code points in a straightforward clean coding system. In addition, such an overlap between character and control codes demands more sophisticated communication protocols, which further reduce the physical efficiency of transmission.

Conclusion

The foregoing analysis draws attention to the following focal points:

- Mixing character sets will not create serious problems as long as a regional standard exists.

- The different needs of regional communities must be recognized and the regional autonomies in decisions must be respected; nevertheless, each region must have a regional standard.

- For worldwide information interchange, a pivot system must be adopted for efficient and clean operation.

- The longer a decision on standardization is delayed, the less clean and more complex the standards will become, resulting in an overall system that is more complicated, less efficient, and much more expensive in the long run. Therefore, it is imperative that efforts to produce worldwide standardization start at once.

- Policymakers of the regional governments must be made aware of the need for standardization because of the trade-off between short-term investment now and long-term financial drain resulting from the lack of such investment.

APPLICATIONS SOFTWARE DEVELOPMENT FOR CHINESE BIBLIOGRAPHIC DATA BASES

Chairman: Kyu-Soo Kim
Recorders: C.C. Hsieh and C.Y. Suen

Although the general functions of a bibliographic data base system are commonly understood, different user demands have resulted in a variety of coding methods, processing algorithms, and systems and equipment-applications software that have been developed independently by various organizations in different countries. This situation may be improved if cooperation in development and exchange of software is encouraged. This is particularly important for those organizations that are at the planning stage of systems development. They must decide whether to develop their own applications packages or to purchase existing software.

Unfortunately, most Western countries that already have developed data bases handle only a small percentage of Chinese, Japanese, and Korean (CJK) characters in their total data base operations. Because CJK languages are not a major concern in the Western world, most software houses do not pay any special attention to their needs and most existing systems attempt to fit the CJK records by using the available software, which was designed to handle Roman text materials. These packages may not satisfy the needs of those countries where CJK characters constitute the majority of the data base collections. For example, the Research Libraries Information Network's (RLIN) data base has about 70,000 CJK records in a total of 16 million records. The OCLC Online Computer Library Center (OCLC) in the USA has now started to develop CJK records, but at present the number of items in the collection is extremely small. In Canada, most collections are in Roman-alphabet languages.

In Asian countries, the demands are different.

One of the existing systems in Korea handles about 60% Roman characters, 10-20% Hangul (however, this is increasing rapidly), with the remainder in Kanji and Kana. The University of Hong Kong system has about 9000 Roman-script titles and 4000 CJK titles. Thailand has a system that collects about 90% of its information in Thai and 10% in English. The Agricultural Science and Technology Information Management System (ASTIMS) handles both Chinese and English publications and is growing by about 12,000 titles per year. In Japan, various organizations have large collections of Japanese titles and are already providing comprehensive cataloguing and data base services. China is now beginning to create a union catalogue that will include 7000 Chinese-language titles in science and technology.

As mentioned earlier, different users have developed a variety of different systems according to their applications and type of equipment. These systems are generally not homogeneous in handling CJK in terms of number of characters and type of character. RLIN has a CJK system that is totally integrated in a general system that provides different screen formats for different applications. All the software packages were designed in-house and are running on IBM systems software and equipment. The information is transmitted as one whole screen at a time. Each character in RLIN's system has a three-byte control code that is used as an index to raise the dot matrix character pattern for display on the screen. OCLC is currently using an IBM personal computer-based M300 Chinese keyboard. The University of Tokyo's bibliographic information service centre, which serves regional centres and local libraries, uses different software packages produced by software houses for each application. Singapore's National Library has just completed a survey and the consultant concerned recommended the use of an Australian system. The Institute for Scientific and Technical Information of China (ISTIC) uses a system based on a Japanese computer (TK-70) with Chinese input and output subsystems for Chinese-character processing. An identical system was being used in Korea but it has been replaced by an IBM 4341 mainframe with newly developed software to run on the IBM hardware. This new system is called STAIRS (Storage and Information Retrieval System), which uses a two-byte code system to handle CJK data, and was developed cooperatively with IBM Korea. Hangul characters can be handled successfully, but the Chinese and Japanese character-information handling is still in the developmental stage.

Sorting of Chinese text is difficult, and is complicated by different codes used in different systems, for example, CCCII, GB code, JIS, KIPS, etc. However, information has to be sorted. In Taiwan China, most systems use sorting by radicals and by the number of strokes -- the University of Hong Kong follows this system. In China, records are sorted by a Pinyin pronunciation system. There is no optimal collating sequence available today.

It is generally agreed that international cooperation in software development is highly desirable to permit exchange of collection information among participating countries. Coding systems, user's query commands, and the choice of computer languages must be standardized. Use of a high-level language is recommended for the development of applications software so that it can be easily transferable to other systems. It is also highly desirable that a common, standard format MARC (Machine-readable catalogue) should be used. At present, many countries have created their own MARC format. However, a proposed universal format (UNIMARC) exists and is being used in Taiwan China and Korea. China has not decided yet on which MARC format to use but it wishes to create its own and, in the USA, several MARC formats already exist in several different systems.

CHINESE DATA BASE SYSTEMS -- CASE STUDIES

Chairman: Wan-Jiun Wu
Recorders: Andrew H. Wang and Victor Li

Two systems were presented and discussed during this session: The Agricultural Science and Technology Information System and The Computerized Chinese Mechanical Engineering Abstracts.

The Agricultural Science and Technology Information Management System -- Presented by Wan-Jiun Wu

Over the last 30 years, as agricultural education has become widely available and qualified agricultural researchers and facilities improved, agricultural science in Taiwan China has reached an advanced stage of development. The problem that now urgently needs a solution is how to handle the large volume of Chinese agricultural literature generated so that it can be used to the greatest possible extent so as to assure the success of agricultural research through the availability of accumulated research results, and to allow for the exchange of information with other countries.

The economic development plan for the next 10 years is to move from a labour-intensive phase to a technology-intensive phase and thus from a less-developed state to a developed one. The success of the plan will depend on optimun use of raw materials and energy resources, and on the early creation of an information management system. Of these, development of better information-processing techniques is the most urgent task for development.

In January 1978, the Agricultural Science Information Center (ASIC), following the guidelines for creating a national agricultural science and technology information service system, completed the Agricultural

Science and Technology Information Management System (ASTIMS) -- which took 4 years to design and implement.

Objectives

The objectives of ASTIMS are to create data bases on agricultural personnel, research projects, research reports, and agricultural literature; to develop a network of agricultural libraries; to introduce international standards for information processing; to provide access to international data bases; to develop an agricultural thesaurus and classification scheme; and to pave the way for computer processing of agricultural information.

ASIC databases

ASTIMS now includes four data bases, in various stages of development. These data bases are described in the following paragraphs.

The agricultural thesaurus is a controlled vocabulary file of about 17,000 terms. The AGRI-THESAURUS is bilingual, with Chinese and English versions. It acts as the indexing tool for the other data bases within ASTIMS and also provides subject access to these data bases.

Files for Agricultural Science and Technology Personnel (FASTEP) contains information about the country's agricultural labour force from organizations involved in policy-making, administration, research, education, and extension work. The data base now includes 14,883 persons. There are 59 access points available for retrieval and retrieval results may be browsed in a number of ways. In addition, the system can provide summary statistics of one organization or of the whole data base concerning personnel information. For instance, summaries are possible by age brackets, degrees earned, and so forth. Access points include personal name, age, school, graduation date, degree, major field, country where degree was obtained, affiliation, expertise, and AGRI-THESAURUS keywords.

Files for Agricultural Science and Technology Research Projects (FASTEJ) contains research projects sponsored by the Council for Agricultural Planning and Development from 1981 to 1983 and now contains 4310 research projects. Aside from retrieval and browsing, the system can also provide funding reports and analysis. Access points include supervising agency, executing agency, supervisors, performers, executing year, and AGRI-THESAURUS keywords.

The ASIC-MARC Bibliographic System data base contains serial articles, research reports, monographs, and serials. The communication format of the cataloguing subsystem is based on the UNIMARC format whereas that of the authority control subsystem is based on the LCMARC format. Access points to the bibliographic system include personal and corporate names, conference name, title, series title, author/title, ISBN, ISSN, ASIC number, location symbols, and AGRI-THESAURUS keywords. Retrieval can also be limited by year of publication, type of record, target audience, language, and so forth. Boolean logic can be applied in all cases. Searching the cataloguing subsystem retrieves truncated entries, and various record display formats are possible. Searching the authority subsystem retrieves authority headings such as names, series, and keywords, their postings, and cross-references. The cataloguing subsystem also contains abstracts of serial articles and research reports.

Services being offered by ASIC and plans for development

At the moment, ASIC provides four major services:

° Online access to ASIC's data bases and to international data bases such as AGRICOLA and CAB through DIALOG and ORBIT;

° Thesaurus construction;

° Design of management information systems; and

° Information processing.

Future plans are to:

° Continue development and maintenance of a national agricultural information system;

- Develop a network of the country's agricultural libraries that can support shared cataloguing, reference requirements, interlibrary loans, circulation, and acquisition functions and can access ASIC data bases online;

- Develop a communication system for agricultural extension workers and farmers; and

- Initiate a forecasting system for agricultural production and marketing.

Some specific developments in the following areas have to be pursued by ASIC and information industry personnel:

- Analysis of word and language structure to develop computer manipulation of textual matter for indexing, classification, abstracting, and retrieval purposes;

- Bibliographic and numeric data base management;

- Library automation; and

- Planning for the information system in agriculture in the country.

Computerized Chinese Mechanical Engineering Abstracts -- Presented by Ai-Lan Huang

The Chinese Mechanical Engineering Abstracts (CMEA) started publication in 1966, but the present title was not adopted until 1982. This abstract journal is published in 12 monthly issues and an annual subject index. Each monthly issue contains about 900 citations from articles published in over 500 Chinese journals and publications. The main coverage of CMEA includes basic theory, design, materials, measurement, and management in mechanical, electrical, and electronic engineering.

To expand international scientific and technical exchange and speed up the Four Modernizations of

China, a computer-readable form of CMEA has been proposed. This will include editing the monthly abstract journal and annual subject index, recording the tapes, and creating a CMEA online data base for retrieval.

The record format of the CMEA is based on the national standard GB 2901 for information interchange, which is similar to ISO 2709. The Chinese characters in CMEA are input by Zhi Bingyi codes, but are stored in the data base as telegraphic codes, and output as Chinese characters. Any ASCII characters are input in simplified form.

The Scientific and Technical Information Institute (STII) of the Ministry of Machine Building Industry has an HP 3000 (III) computer with the HPCIOS Chinese input and output package, MINISIS database management system, and some utility programs. On the basis of this equipment, the programmers at STII created the Chinese-character information system of CMEA. In addition, they have

- Created a pilot data base in which 20 different document records are stored. Its data were input and can be retrieved by telegraph codes through MINISIS, and the Chinese-character search results are printed out using HPCIOS.

- Edited the annual subject index of CMEA on the HP 3000 in 1983. This involved processing 10,000 records, including sorting (four times), correcting, and deleting. The index is arranged in phonetic order of the Chinese-character Pinyin.

- Created the "subject-term number/telegraph code/ Chinese-character Pinyin of the subject term" file using the corresponding file of the CMEA thesaurus of about 10,000 terms. Instead of the subject terms themselves, the subject-term numbers can be used to index and input the document records. This is not only more convenient than using the telegraph codes, but also reduces data-entry errors.

- Created the "Zhi Bingyi code/telegraph code/ Chinese character" corresponding file. Zhi Bingyi codes can be used easily by a person who has taken only a short training course to input the Chinese-character data directly without looking for the codes of words.

ln addition to these developments, we are now editing the first experimental issue of the computerized Chinese-character CMEA. ln this issue, 886 document records will be input to form a new pilot data base that can also be provided for online retrieval. Through the use of MlNlSlS, document data have been input using Zhi Bingyi codes. They can be retrieved by telegraph codes and the searched document in Chinese-characters will be displayed on the terminal screen or printed out.

Computerized production of the Chinese character CMEA will begin formally in the latter half of 1985. The online CMEA data base will be available for service about the same time.

CONCLUDING SESSION

Chairman: T.C. Ting
Recorders: Allan M. Tucker and Wellington Yu

Contrary to some expectations, this 4-day workshop has not been long enough to explore many major issues fully. We have had a very full schedule that has produced active, informative discussion. It was, indeed, an exceedingly productive and fruitful workshop. However, we did not solve all the problems that were identified and instead, perhaps, we raised more questions than answers. It is obvious that the issues of international information exchange of Chinese-character material require serious cooperative work before solutions can be proposed. Clearly, we have a long way to go and the goal of free international information exchange is not just around the corner, nor can it be accomplished soon.

We now have a better understanding of the need for standardization, which would certainly reduce the difficulties faced in exchanging Chinese, Japanese, and Korean (CJK) bibliographic data among different countries. We must establish a sound technical framework and basis for standardization before local and regional developments make it too difficult to accomplish this. Information and library professionals and information technologists and researchers must be aware of regional, as well as international, standardization efforts in relevant areas -- including documentation format, character coding, software and hardware, communications, and input and output methods and technologies.

Information technology is a new field and is still developing rapidly. New products and new methods are being introduced at a fast pace and many of them are based only on ad hoc research results. Therefore, flexibility in systems design must be maintained to take advantage of new technological advancements. Equally, the needs of users are dynamic and they are affected by the availability of new information systems and new products.

During our discussions, we have tried to identify some areas where changes could influence future development. Choice or flexibility of input method for a user is an important consideration and technology may permit reliable voice input in the not-too-distant future -- this would be a highly desirable development. The possibility of expanding the character set is necessary to accommodate new scientific discoveries and must be maintained. Font design is one area where research is needed. New fonts should have eye-legibility, but must also consider machine recognition for efficient bulk-data input. Advances in OCR (Optical character recognition) may reduce the degree of human effort needed for inputting large numbers of records such as are required by large data base systems: human needs must not be ignored in designing input systems.

Problems in output of Chinese-characters are relatively easier to deal with than those of input. A variety of high quality, flexible printers and display devices is becoming available and, in the near future, the capability to produce various fonts in different sizes and styles should be available. The output devices will be able to handle mixtures of fonts and non-Chinese-character sets as well as graphic materials. Voice output devices are already available and are desirable in various contexts.

Applications software development cannot be considered in isolation. They are affected by the coding systems, data formats, systems software, and hardware equipment. Most existing software packages in the Western world have been developed for alphabetic languages and very few have been designed specifically for CJK applications. International cooperation in software development and technical exchange for CJK applications is highly desirable not only to promote international standardization, but also because it can be economically advantageous to all participating organizations.

Many existing large-scale computer systems do not have software that allows for input, manipulation, or output of CJK materials. However, recent developments in microcomputer-based systems are changing this picture. A variety of CJK packages and features has been proposed and is being developed for these highly flexible microcomputer systems.

This workshop has opened a new forum for technical exchange and has shown that this type of meeting is extremely valuable and must be repeated. Subsequent workshops should perhaps be more limited as to topic but among the more important areas for discussion are developing multilingual software systems and software exchanges. Support for future meetings must be sought from international organizations and from government agencies of participating countries.

The purpose of promoting international information exchange in science and technology is to stimulate research and development to the benefit of all people in the world. Although many unsolved technical problems still exist, we believe that the technical issues can be resolved. However, we will require management commitment for resources and support for seeking solutions. International groups such as this workshop can produce positive results and can demonstrate to the various governments the importance and benefit of international information exchange. We must hope that they will make resource commitments toward the common goal.

APPENDIX 1: PARTICIPANTS AND IDRC STAFF

Chorkin CHAN, Director, Computer Studies, University of Hong Kong, Hong Kong

Richard T. CHENG, Eminent Professor and Chairman, Computer Science Department, Old Dominion University, Norfolk, VA 23508, USA

J.C. GUAN, Director, Library of Domestic Literature, Institute of Scientific and Technical Information of China, P.O. Box 640, Beijing, China

C.C. HSIEH, Research Fellow, Institute of Information Science, Academia Sinica, Nankong, Taipei, Taiwan 115, China

Ai-Lan HUANG, Engineer, Document Division, Scientific and Technical Information Institute, Ministry of Machine Building Industry, 22 Bai Wan Zhang Street, Beijing, China

Lai-Bing KAN, University Librarian, University of Hong Kong Libraries, Pokfulam Road, Hong Kong

Kyu-Soo KIM, Director, Computer Systems Laboratory, Korean Institute for Economics and Technology, P.O. Box 205 Cheongryang, Seoul, Korea

Victor LI, Computer Systems Analyst, Automated Systems Office, Library of Congress, Washington, DC 20540, USA

Y.A. LIAN, Chief Engineer, Institute of Scientific and Technical Information of China, P.O. Box 640, Beijing, China

Weina LIU, Engineer, Beijing Institute of Information for Management, SSTCC, P.O. Box 2828, Beijing, China

Shiu-Chang LOH, Professor and Head, Computer Science Department, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Ronghua LU, Deputy Director, Computer Service Division, Institute of Scientific and Technical Information of China, P.O. Box 640, Beijing, China

Ching Y. SUEN, Professor, Department of Computer Science, Concordia University, 1455 de Maisonneuve West, Montreal, Quebec, Canada H3G 1M8

Lawrence Wai-Hong TAM, Associate Librarian (Technical Services), Hong Kong Polytechnic, Yuk Choi Road, Hung Hom, Kowloon, Hong Kong

T.C. TING, Professor and Head, Computer Science Department, Worcester Polytechnic Institute, Worcester, MA 01609, USA

Alan M. TUCKER, Associate Director, Library Systems Development, The Research Libraries Group, Inc., Jordan Quadrangle, Stanford, CA 94305, USA

Andrew H. WANG, Manager, Online Systems Products and Services, OCLC Online Computer Library Center, 6565 Frantz Road, Dublin, OH 43017, USA

Huimin WANG, Engineer, Computer Services Division, Institute of Scientific and Technical Information of China, P.O. Box 640, Beijing, China

WARAWUDH Kruasin, Manager, Computer System Development and Techniques Division, Krung-Thai Bank, Bangkok, Thailand

Wan-Jiun WU, Director, Agricultural Science Information Center, 14 Wen Chou Street, Taipei, Taiwan 106, China

Hisao YAMADA, Professor, Department of Information Science, Faculty of Science, University of Tokyo, Hongo, Tokyo 113, Japan

Wellington C.P. YU, Senior Engineer and Manager, DASD III Microcode Development, IBM Corporation, 5600 Cottle Road, San Jose, CA 95193, USA

IDRC

Gilbert CROOME, Senior Technical Editor, Scientific Editing Unit, Communications Division, International Development Research Centre, P.O. Box 8500, Ottawa, Ontario, Canada K1G 3H9

Richard LEE, Head, MINISIS Future Systems, Information Sciences Division, International Development Research Centre, P.O. Box 8500, Ottawa, Ontario, Canada K1G 3H9

Michael SHERWOOD, MINISIS Project Advisor, Information Sciences Division, International Development Research Centre, c/o Southeast Asian Regional Centre for Graduate Study and Research in Agriculture, College, Laguna, Philippines

Sally TAN, Program Assistant, Information Sciences Division, International Development Research Centre, Tanglin P.O. Box 101, Singapore 9124, Republic of Singapore

Robert VALANTIN, Associate Director, Information Sciences Division, International Development Research Centre, P.O. Box 8500, Ottawa, Ontario, Canada K1G 3H9

APPENDIX 2: ACRONYMS AND ABBREVIATIONS

AGRICOLA	Agricultural Online Access (U.S. Department of Agriculture)[1]
AGRI-THESAURUS	Controlled vocabulary agricultural thesaurus (ASIC)
ASCII	American Standard Code for Information Interchange
ASIC	Agricultural Science Information Center, Taipei, Taiwan, China
ASTIMS	Agricultural Science and Technology Information Management System (ASIC)
CAB	Commonwealth Agricultural Bureaux, Farnham Royal, U.K.
CAI	Computer-aided instruction
CC	Code conversion, code converter
CCAG	Chinese Character Analysis Group, Taiwan, China
CCDB	Chinese character data base
CCCII	Chinese character code for information interchange (CCAG)
CJK	Chinese, Japanese, and Korean languages
CMEA	Chinese Mechanical Engineering Abstracts

[1] Entry in parentheses is parent organization.

CRT	Cathode-ray tube
DIALOG	Registered trade mark of Lockheed Information Systems on-line information retrieval system
FASTEJ	Files for agricultural science and technology research projects (ASIC)
FASTEP	Files for agricultural science and technology personnel (ASIC)
GB	Chinese national standard code (GB 2312-80: Chinese graphic character set for information interchange - primary set)
HP	Hewlett-Packard Company, USA
HPCIOS	Hewlett-Packard Chinese Input and Output System
IBM	International Business Machines
IDRC	International Development Research Centre, Ottawa, Canada
III	Institute of Information Industries, Taiwan, China
IPA	International phonetic alphabet
IPX	IPX Ideographix Inc.
ISBN	International standard book number
ISSN	International standard serial number
ISO	International Organization for Standardization, Geneva, Switzerland
ISTIC	Institute of Scientific and Technical Information of China, Beijing, China
JBMA	Japan Business Machine Manufacturer's Association, Tokyo, Japan

JDL	Japanese Diet Library, Tokyo, Japan
JEIDA	Japan Electronic Industry Development Association, Tokyo, Japan
JICST	Japan Information Center of Science and Technology, Tokyo, Japan
JIFH	Ju In Fu Hau
JIS	Japan industrial standard (JIS C6226: Code of the Japanese graphic character set for information interchange)
JPIC	Japan Patent Information Center, Tokyo, Japan
KIET	Korean Institute for Economics and Technology, Seoul, Korea
KIPS	Korean Information Processing System
KSC	Korean standard code
KWIC	Keyword-in-context
LC	Library of Congress, Washington, DC, USA
LCMARC	Library of Congress (USA) MARC
MARC	Machine-readable catalogue
MINISIS	Interactive minicomputer system for information retrieval and library management (IDRC)
NEC	Nippon Electric Company, Tokyo, Japan
NKS	Nippon Keizai Newspaper, Tokyo, Japan
NTT	Nippon Telegraph and Telephone Corporation, Tokyo, Japan

OCLC	OCLC Online Computer Library Center, Dublin, OH, USA
OCLCMARC	Online Computer Library Center MARC (OCLC)
OCR	Optical character recognition
ORBIT	On-line retrieval of bibliographic information, time-shared
PC	Pivot code
REACC	RLIN East Asian Character Code (RLG)
RLG	Research Libraries Group, Stanford, CA, USA
RLIN	Research libraries information network (RLG)
SI	Shift-in control code
SO	Shift-out control code
STAIRS	Storage and Information Retrieval System
STC	Standard Telegraphic Code
STII	Scientific and Technical Information Institute, Ministry of Machine Building Industry, Beijing, China
UKMARC	United Kingdom MARC
UNIMARC	Universal MARC
USMARC	United States MARC

APPENDIX 3: BACKGROUND INFORMATION[1]

Profile of the Chinese Language

Chinese is a nonalphabetic language that uses a large number of characters or ideograms. It has been treated by many as a monosyllabic language but is actually not. Although each Chinese character is still pronounced as a single syllable and individual characters still maintain an independent meaning, most Chinese words are made up of two or more characters. The formation of Chinese words is much like the formation of the English words "railroad" and "blackboard" -- i.e., they are made up of elements with meanings of their own.

Mandarin, the official spoken language, has only 405 possible syllables with which to pronounce the 5000 written characters in common use. Each syllable, however, is pronounced with one of four "tones" or inflections: high/level, rising, dipping, and falling. The same syllable, pronounced with a different tone, takes on a different meaning (Table 1) and, in effect, becomes a new syllable.

Table 1. Example of effect of tone on syllable meaning.

Syllable[a]	Tone	Character	Meaning
mā	1	媽	mother
má	2	麻	hemp
mǎ	3	馬	horse
mà	4	罵	to scold

[a]Note the tone marks over the vowel.

[1]Adapted from Koach (1984).

Table 2. Example of effect of combination of syllables in reducing ambiguity.

Character	Pronunciation	Meaning
27	bei	--
41	jing	--
北京	běijīng	Beijing (the capital)
背景	bèijǐng	background

Thus, taking tone into account, the total number of Mandarin syllables increases from the original 405 to about 1300. Even so, the number of homophones in Chinese is still very high. For example, the syllable ji, pronounced with the fourth tone, is used for at least 30 characters and, if tonal inflection is not taken into account, the syllable ji is used for 123 characters.

Formation of words with two or more characters reduces the phonetic ambiguity of the language. For example, there are 27 Chinese characters pronounced bei, and 41 characters pronounced jing, but only two words pronounced beijing (Table 2) and these are pronounced differently because of tonal differences.

Pronunciation of the same character varies widely depending on the dialect and regional accent. With Mandarin being promoted as the official language -- it is taught in schools and used in mass media -- many Chinese are bilingual, retaining their native dialect to converse with family and friends but using Mandarin to speak with others.

How many characters are there?

A vocabulary of 4800 characters will satisfy 95% of all teaching, writing, newspaper printing, word processing, and popular cataloging. Larger character sets of 9000-15,000 are used for business data processing (e.g., utility billing, telephone, etc.) and libraries. Census applications and other general data-processing applications require about 22,000 characters. The largest dictionaries contain as many as 60,000 characters but many of these are obscure.

Structure of Chinese characters

Chinese characters are actually composed out of a relatively limited set of about 200-500 components. The number of components is indefinite because there is no standard set of components and no standard way of dividing a character into components.

For each character, one component is designated the radical. This is the semantic component -- the component that has something to do with the meaning of the character. For example, the characters for iron (鐵) and lead (鉛) both contain the metal radical (金). The character meaning to speak (說) and the character for language (話) both contain the speech radical (言).

Other components in the character, if present, may give some indication of the character's pronunciation (a phonetic component), or may have no particular significance. It is important to stress that radical and phonetic components may bear only a remote resemblance to the character's meaning or pronunciation.

Radicals range in complexity from 一 (the number one) to 龍 (dragon) and it is often difficult to guess which component of the character is its radical because radical components do not always function as radicals. One character may have several components, each of which is on the list of radicals, but only one of which serves as the radical under which that particular character is classified. For example, the character 矮 (short) contains the three components 矢 , 禾 , and 女 , each of which is among the list of radicals. In this case, only the first component is the official radical for that particular character.

Components themselves are composed of a limited set of brush strokes (Table 3). The total vocabulary of brush strokes is a little larger, but they are variations or combinations of the basic seven strokes. Each character has a certain number of brush strokes, which is exploited for character sorting. Also, because there

Table 3. The fundamental strokes.

	Stroke	Chinese name	English name
1	丶	diǎn	dot
2	一	héng	horizontal
3	丨	shù	vertical
4	丿	piě	downward to left
5	㇏	nà	downward to right
6	㇀	tí	rising
7	亅	shùgōu	straight hook

is a proper way of drawing the characters, i.e., a standard sequence of strokes to follow, stroke sequence can also be exploited for sorting.

Simplified versus traditional characters

Simplification of Chinese characters began in 1955 as part of an effort to improve literacy. By 1964, 2238 characters (the majority of frequently used characters) had been officially simplified. For example, in its traditional form, the word "Chinese" is written: 中國話 . In simplified form, the same word is written: 中国话 .

Simplified characters are used in China and are accepted in Singapore with some variations. Even though simplified characters are the standard, traditional characters may still appear on title pages and other places where a formal appearance is desired. In Taiwan China, the traditional character forms are still the official written style and the simplified forms are not used in print.

Phonetic alphabets

The earliest and, until recently, the most widely used romanization system is Wade Giles. It is still used in many Western language libraries for transcription of Chinese titles, authors, and place names. However, it is not a particularly good or easy system to learn and Pinyin, which is a great improvement, has been introduced as a replacement.

Pinyin has been in use since 1958 when the Pinyin program was first announced and in January 1979, it became the official romanization system for Chinese characters. Thus, names spelled like Teng Hsiao-ping under the Wade Giles system became Deng Xiaoping under the Pinyin system. In 1981, the Information Science Committee of the International Organization for Standardization (ISO) adopted a resolution to make Pinyin the international standard for spelling of Chinese characters.

Pinyin is used in China as an educational tool in primary schools as an aid for learning characters and their pronunciation. Dictionaries use it for character pronunciation and, in some cases, for alphabetical arrangement of entries. Although the Pinyin spelling of characters has been standardized, there is little agreement on how to group character spellings into words. The idea of replacing characters with Pinyin has proven unworkable.

Pinyin is not used on Taiwan China which has adopted the National Phonetic Alphabet. This is sometimes referred to as bopomofo -- which is how the first four letters of the alphabet are pronounced. Because the alphabet is non-Roman, Wade Giles is still being used for Western audiences.

Special Issues in Processing Chinese Data

Sorting

Chinese characters are sorted a variety of ways. The common sorting sequences are phonetic spelling, radical component, number of strokes, stroke sequence, and four corner code.

In phonetic spelling, characters are sorted by how they are spelled -- in China using Pinyin and in

Table 4. Example of sorting by Pinyin, tone, and number of strokes.

Character	Pinyin	Tone	Number of strokes
哥	gē	1	10
土	tǔ	3	3
吐	tǔ	3	6

Taiwan China using the National Phonetic Alphabet. For characters with the same spelling, the subsort is by the four tones. For characters with the same pronunciation and tone, a further subsort is by the total number of strokes (Table 4). It is possible for one character to have different pronunciations, in which case no unique collating sequence exists.

Radical component is the traditional way of sorting characters. The traditional list of 214 radicals used in several popular dictionaries is sequenced by number of strokes in the radical; however, there are other lists of radicals. Characters with the same radical are sorted by number of residual strokes, i.e., the number of strokes besides the radical component. When the three characters in Table 4 are sorted on the basis of radical component, a new sequence is created (Table 5). Because the selection of radical for a given character is not always fixed, alternative collating sequences do exist.

Characters can be sorted by their total numbers of strokes with two alternative secondary sorts, by stroke sequence or by radical.

Stroke sequence can be used as a basis of sorting because characters are drawn with a definite stroke order. By defining a set of fundamental strokes and assigning numbers to each stroke, characters can be sorted by the resulting sequence of stroke numbers. This sorting method is typically used as a secondary sort only.

Table 5. Example of sorting by radical and residual strokes.

Character	Radical	Number of residual strokes
吐	口	3
哥	口	7
土	土	0

The four corner code or initial stroke method of sorting is the same as the stroke sequence method except that only the initial stroke is considered. The first five strokes in Table 3 form one common ordered set of fundamental strokes. Thus, a character whose first stroke is a dot is sorted before a character whose first stroke is a horizontal line. Based on this concept, the four corner code was developed to consider the stroke patterns of the four corners of a given character. This code plus the number of strokes has been used for sorting. When the same three characters of Table 4, plus a fourth for this example, are sorted by initial stroke and secondarily by total number of strokes, another sequence is developed (Table 6).

Different libraries use different sorting methods

Table 6. Example of sorting by initial stroke and number of strokes.

Character	Initial stroke		Number of strokes
土	一	(2)	3
哥	一	(2)	10
上	丨	(3)	3
吐	丨	(3)	6

for their catalogues. For example, the Beijing and Singapore library catalogues are organized by Pinyin; the National Central Library, Taipei, China, uses a three-level sort, stroke count, radical, and stroke sequence; and public libraries in Hong Kong use a two-level sort, number of strokes and initial stroke.

Standards

The Standard Telegraphic Code (STC) is a Chinese standard that grew out of the traditional telegraphic code used in transmitting telegraphs and cables. It is still used for that purpose, and is now also used for computer data entry and as an internal code for storing Chinese data. STC is a four-digit code from 0000 to 9999. Each character is also assigned an alternative code of three Roman letters, ranging from AAA to OUP. Character arrangement is roughly by radical. STC has diverged quite a bit from the traditional telegraphic code, which is still used outside China. The ministry responsible for maintaining STC, the Ministry of Posts and Telecommunications, revised the code so that it would be in line with the character simplifications.

GB 2312-80 is the Chinese national standard for information interchange of Chinese characters. It is modeled after the Japanese Industrial Standard JIS C6226. GB 2312-80 is a two-byte code with each byte having seven bits. The binary values for each byte range from hex 21 to 7E, which correspond with the printable ASCII character set (ASCII-7 less the control codes, the DEL character, and the SP character). GB 2312-80 is represented by a two-dimensional table with 94 rows and 94 columns and, therefore, has 8836 possible positions.

GB 2312-80 contains general characters (punctuation and special symbols); Arabic numerals; Roman, Greek, and Russian alphabets; Japanese kana; vowels with tone marks needed for Pinyin; the National Phonetic Alphabet (bopomofo); and Chinese characters. The 6763 Chinese characters are divided into two groups. The first contains 3755 of the most commonly used characters, arranged alphabetically by the Pinyin spelling with a secondary sort by stroke sequence. The second group of 3008 less frequently used characters is arranged by radical component with secondary sort by number of strokes and tertiary sort by stroke sequence. Because the two levels of GB 2312-80 are sorted differently, the code cannot be used for sorting. GB 2312-80 will be expanded to include more characters in the near future -- this is possible because only 7445 of the 8836 possible positions are being used.

Chinese character code for information interchange (CCCII) has been proposed as a standard for Chinese information exchange among data-processing systems and within message-transmission systems. It is not being promoted as a desirable internal coding system although it may be used as such.

CCCII is a three-byte code. As with GB 2312-80, each byte has seven-bits, and corresponds to 94 printable ASCII characters. CCCII defines a three-dimensional 94 x 94 x 94 code space with enough capacity to include all Chinese characters, as well as Tibetan, Manchurian, Mongolian, and other languages. Subsets of CCCII may be defined and mapped mathematically to a two-byte code. Variant forms of the same character are given special code assignments. Because they share two bytes with the standard form and vary only in the first byte, sorting and searching is easy. Simplified characters are included in CCCII, but they are treated as variant forms.

CCCII is still in the process of development and, so far (through volume 2), 33,544 characters have been defined. The ultimate goal is to include all Chinese characters (up to 80,000).

The Chinese Character Analysis Group (CCAG) is responsible for CCCII. They maintain a CCCII database that includes the character's code; bit patterns at 32 x 32 resolution in two different font styles (kai shu and Sung ti); pronunciation by Wade Giles, Pinyin, and National Phonetic Alphabet; the stroke count; the three-corner code; and the GB 2312-80 code. Chinese characters in CCCII are arranged by radical, then by initial stroke.

CCAG has taken a comprehensive approach to the coding problem but it is difficult to say how much acceptance CCCII will achieve; however, the CCCII file is available on tape.

Treatment of traditional and simplified forms

Traditional and simplified forms of characters can be treated in three ways:

- The computer terminal handles only one form -- either simplified or traditional but not both.

- The terminal treats the two forms as different fonts. Internal coding is same for both forms of the character. A font selection switch on the keyboard or font-indicator codes embedded in the text determine which form of the character is displayed.

- The traditional form is treated as a separate character from the simplified form. Each form of the character has its own unique code.

In the character simplification process, several traditional characters have been consolidated into a single simplified character. For example, the traditional characters:

臺 (stage) 檯 (table) 颱 (typhoon),

all of which are pronounced tai, are represented by a single simplified character 台 whose meaning now must be derived from context.

Mapping from traditional-character data to simplified-character data is easy but entails a loss of information. The reverse, mapping from simplified to traditional, is a problem and the safest solution for such ambiguities is perhaps to simply leave the character in its simplified form.

APPENDIX 4: REFERENCES AND BIBLIOGRAPHY

Anonymous. 1979. Study report on the public Kanzi terminal, volume 1. [In Japanese.] Tokyo, Japan, Japanese Diet Library, Japan Information Center for Science and Technology, Nippon Keizei Newspaper, Nippon Telegraph and Telephone Corporation, and Japan Patent Information Center. Joint publication. 51 p.

__________ 1980. Study report on the public Kanzi terminal, volume 2. [In Japanese.] Tokyo, Japan, Japanese Diet Library, Japan Information Center for Science and Technology, Nippon Keizei Newspaper, Nippon Telegraph and Telephone Corporation, and Japan Patent Information Center. Joint publication. 109 p.

Chen, C.-K. and Gong, R.-W. 1984. Evaluation of Chinese input methods. Computer Processing of Chinese and Oriental Languages, 1 (4), 236-247.

JBMA (Japan Business Machine Manufacturer's Association). 1983. A study on engineering development factors for the spread of Japanese word processors, volume 1. [In Japanese.] Tokyo, Japan, JBMA. 297 p.

__________ 1974. A study on engineering development factors for the spread of Japanese word processors, volume 2. [In Japanese.] Tokyo, Japan, JBMA. 283 p.

JEIDA (Japan Electronic Industry Development Association). 1983. Summary of presentations by the technical exchange delegates on the Kanzi information processing systems. [In Japanese.] Tokyo, Japan, JEIDA. 122 p.

Koach, D. 1984. Processing Chinese characters with MINISIS: Consultant's report to the International

Development Research Centre. Ottawa, Canada, IDRC. Centre File 3-A-84-4017 (restricted distribution).

Suen, C.Y. 1979. Computational analysis of Mandarin. Boston, MA, USA, Birkhauser. 160 p.

____________ 1983a. Computer aided design of Mandarin phonetic system. Proceedings of the 8th annual convention of the Chinese-American Academic and Professional Association, New York, NY, USA. 37-38.

____________ 1983b. Recognition of Kanji characters. Proceedings of the International Conference on Text Processing with a Large Character Set. 429-435.

____________ 1985. Character recognition by computer and applications. In Young, T.Y. and Fu, K.S., eds., Handbook of patterns recognition and image processing. Orlando, FL, USA, Academic Press Inc.

Suen, C.Y. and Huang, E.-M. 1984. Computational analysis of the structural compositions of frequently used Chinese characters. Computer Processing of Chinese and Oriental Languages, 1 (3), 163-176.

Suen, C.Y. and Stein, S.B. 1985. Synthesis of speech by computers and chips. In De Mori, R. and Suen, C.Y., eds., New systems and architectures for automatic speech recognition and synthesis. New York, NY, USA, Springer-Verlag.

Wang, Q.R. and Suen, C.Y. 1984. Analysis and design of decision tree based on entropy reduction and its application to large character set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6 (4), 406-417.

Xie, K.Z. and Suen, C.Y. 1985. Computer generation of quality Chinese characters in multi-fonts and different sizes. Proceedings of the International Conference on Chinese Computing, San Francisco. D-2.1 - D-2.12.