A Multi-Lingual Coding System for Chinese, Japanese and Korean (CJK) Data Processing

Jack K.T. Huang, C.T. Chang, C.C. Hsieh,

C.C. Yang and S.S. Tseng

Chinese Character Analysis Group,
Council for Cultural Planning and Development
Taipei, Taiwan, R.O.C.

Reprinted from
ROC-Japan Symposium on
Information Management and Exchange:
Present and Future
November 19-20, 1986, Taipei, Taiwan, ROC

A Multi-Lingual Coding System for Chinese, Japanese and Korean (CJK) Data Processing

Jack K.T. Huang, C.T. Chang, C.C. Hsieh,

C.C. Yang and S.S. Tseng

Chinese Character Analysis Group,
Council for Cultural Planning and Development
Taipei, Taiwan, R.O.C.

ABSTRACT

Information technical developments have made it possible to develop comprehensive databases and to make them available to users worldwide. As the degrees of influences from the diverse sources of the East Asian countries impact the international affairs are increasing upwardly. The processing of information from such East Asian sources for such a wide audience high-lights the importance or cross cultural factors in international communication. Transfer and interchange of vernacular character information of East Asian sources over time and across geographical space become one of the significant factors. A review of the Chinese Character Code for Information Interchange (CCCII) is presented as a background for this Symposium on Chinese, Japanese and Korean (CJK) data processing and information interchange.

1. Introduction to CJK Multi-lingual Environment

The multi-lingual environment is based on CCCII coding system [1] and CCDB [2], [6] which issued in 1980, 1982 and 1985 respectively. CCCII is a three 7-bit-byte code based on ISO-646 and -2022, the graphic area of this system has 94*94*94, totally 821,748 coding positions and 94 control codes for data processing. The first byte, B1, denotes the positions

within a section, the second byte, B2, denotes the sections of a plane, and the third byte, B3, denotes the planes, as shown in Figure 1. CCDB [2], [6] is a data base for Chinese characters tools. The software structure and index mechanism of CCDB are shown in Figure 2-3 respectively. Under those structures of CCCII and CCDB, it is not only adquately used for more than 70 thousands Chinese characters, but also adquately used for the CJK languages which use a quite number of Chinese characters or Chinese-likely characters, such as in Japan and Korea.

In fact, as the degrees of influences from the diverse sources of the East Asian countries impact the international affairs are increasing upwardly. The processing of information from such East Asian sources for such a wide audiance highlight the importance or cross cultural factors in international communication. Transfer and interchange of vernacular character information of East Asian sources over time and across geographical space become one of the significant factors. Therefore, we need a multilingual coding system for international information processing system [3]. The RLIN East Asian Character Code (REACC) uses CCCII for its basic structure and pattern of coding completely [4], with the CCCII and CCDB provide an excellent environment for the work. The advantages are explained as follows:

(1) The implementation of CCCII is shown in Figure 4, 94 planes devide into 16 layers, each layer has 6 planes except layer 16 has only 4 planes (33,000 Chinese regular and variant characters all are included). The first section of every plane is reserved for CCCII control code which will explain later, section 2 to 15 of the first plane of each layer are reserved for user's area except layer 13, 14, 15, and 16. Please see Figure 5 and 6.

(2) The first layer allocates the regular form of Chinese characters. The second layer through layer 12 allocate the variant forms of Chinese characters and the simplified forms are treated as variant forms which are allocated in layer 2 (6,763 Chinese characters of GB 2312-80 all are included in layer 1-12).

(3) We treat part of the 6,349 Japanese Kanji of JIS C 6226 as variant forms of Chinese characters that are not matching the regular form of Chinese characters and allocate them in layer 13. The other symbols of JIS C 6226 then allocated in the section 2 to 15 of layer 13.

By the mechanism of CCDB, the Japanese Kanji can be coded in CCCII structure in such a way, and we had done it already.

(4) Layer 14 is reserved for Korean characters (2,058 Hanja out of 2,392 characters of KIPS are included, or KSC-5619 if it is so demanded by users) in the same way as allocate the Japanese Kanji in layer 13.

(5) Plane 85, the first plane of layer 15 is reserved for characters of minority Chinese language such as Mongolian, Tibetan, Machurian, Moslem, etc..

(6) Plane 86 through plane 94 are reserved for the other languages of the world. There are adquately enough coding spaces and convenience mechanisms for an universal language coding system.

(7) The control codes of CCCII include two part, the first part is used for the extension techniques by means of escape sequence [5]. The second part is used for edition or other applications of text files. These code keep B3 = blank, B2 = 33 as control code indicators, practically, only B1 is used. We have assigned 27 control codes as shown in Figure 7, it can be increased by practic requirement. Those codes are grouped as follows:

(a) Text Separators [5].

(b) Typesetting Effectors [5].

(c) Code-Format Switchers [5].

2. The Multi-lingual Coding System for CJK Data Processing

Chinese Character Data Base (CCDB) [6], shown as Figure 2, is a data base together with necessary software that was developed based on the structure of CCCII in order to make the application of CCCII easy and efficient. CCDB has three main functions:

(1) To provide Chinese character constructions, phonetic transcription, dot matrix configurations of Chinese characters, various correlated input methods and various existed corresponding Chinese coding systems [6], please see Figure 8.

(2) To provide controlling and searching mechanism of transcription table between corresponding lexicographic character fonts and phonetic symbols.

(3) To provide the interchanging function among various existed corresponding multi-bytes coding systems. For example, from 3-byte CCCII to 2-byte JIS C 6226 or from 2-byte JIS C 6226 to 3-byte CCCII [5], etc.

The CCDB file structure is mainly organized by two portions: data files and indexing files. The data files subdivided into 16 layers. All files are grouped into seven file-groups as described below:

(1) G1 (file of layer 1): designates the data files of regular form of Chinese characters.

(2) G2 (files of layer 2): designates the data and searching files of simplified form of variant forms of Chinese characters.

(3) G3 (files of layer 3-12): designates the data and searching files of other variant forms of Chinese characters appearing in layer 1. Currently, only five selected variant forms are verically allocated from layer 3 to 7.

(4) G4 (files of layer 13): the data and searching files of Japanese Kana and 6,349 Kanji of JIS C 6226 are allocated in layer 13.

(5) G5: inverted lists of Chinese lexicographic fonts, phonetic symbols, and the other attributes of Chinese characters.

(6) G6: Indexed tables of various Chinese character input methods, e.g., 3-Corner code, Dragon input method, etc.

(7) G7: Cross-Reference Table of various existed relevant coding systems.

The Radix of Chinese characters stored in CCDB is called R94, a 16-bit condensed code of CCCII. The file groups separated by a layer pointer LP. To calculate R94 and LP, CCDB provides formula as:

R94 = [(B3-33) mod 6]*94*94+(B2−34)*94+(B1−33), and

LP = (B3−33)/6.

The searching mechanism of CCDB uses three different addressing modes: (See Figure 3)

(1) Direct addressing: If LP = 0, the R94 code is the storing address of the regular form of Chinese characters in layer 1 in the environment of CCCII.

(2) Indirect addressing: If LP = 1, the R94 code is the address of the storing address

of the simplified form of Chinese characters in layer 2. If Lp = 12, R94 code is the indirect address of Japanese Kana and Kanji of JIS C 6226 allocated in layer 13 in the environment of CCCII. If Lp = 13, the R94 code is the indirect address of KIPS or KSC-5619 Korean characters (Hangul and Hanja) which are allocated in the environment of CCCII.

(3) Indirect-displacement addressing: If Lp = 2 to 11, the R94 code is the address of a Base, denoted by B, and the storing address of the variant forms of Chinese characters in layer 3 through layer 12 in the environment of CCCII are calculated by the formula m = B+LP−2 (please see Figure 3), where Lp must less than or equal to the upper boundary of Lp (L). The upper boundary of Lp and B are different from character to character, so it should be pre-determinated when the character was stored.

The above mentioned three addressing modes are also applicable to currently reserved 85-94 planes of CCCII. The CCDB provides this mechanism for multi-lingual data processing as an universal coding system as long as those language were coded within the scope of CCCII. One important and common feature of CCCII and CCDB we would like to point out that the both mentioned systems are open systems in nature, subject to be increment of appending of relevant searching and data files without further modification of the system.

3. Conclusion

Since CCCII provides the largest coding spaces (821,748 characters) available for data processing purpose. The 3-byte structure of CCCII allows for the multi-lingual data processing. The internal logic of 3-byte CCCII makes it possible for the linkage of regular, simplified, variant forms of Chinese characters, and others such as Japanese Kana and Kanji, Korean Hangul and Hanja, and almost all alphabetic characters to be processed under one unique coding system. To avoid of multiple character sets in processing of multi-lingual environment. invoke escape sequences between them, and thus arbtrarily partitioning a unitary coding system.

The CCCII offers a single, coherent coding system and extends the applications limit for multilingual data processing in all fields. Transfer and interchange of vernacular character information of East Asian sources over time and across geographical space become reality.

4 Reference

[1] The Chinese Character Analysis Group, "Symbol and Character Table of Chinese Character Code for Information Interchange", Taipei, Taiwan, R.O.C., May, 1983.

[2] The Chinese Character Analysis Group, "The design of a Cross-Reference Database for Chinese Character Indexing", Taipei, Taiwan, R.O.C., Apr. 1980.

[3] (a) OCLC Newsletter, "Libraries testing CJK350", Dublin, Ohio, U.S.A., June 1986, Page 11.

(b) OCLC Newsletter, "CJK350 field testing completed", Dublin, Ohio, U.S.A., August, 1986, Page 8.

[4] Karen Smith-Yoshimura & Alan Tucker, "RLIN East Asian Character Code and RLIN CJK Thesaurus", 2nd Asian-Pacific Conference on Library Science, Seoul, Korea, May 20, 1985.

[5] IFLA Pre-Conference Seminar on Automated System to Access for Multilingual and Multiscripts Library Materials: Problems and Solutions, "An Universal Coding System for Multi-lingual Environment". Nihon Daigaku Kaikan, Tokyo, Japan, August 21-22, 1986.

[6] The Chinese Character Analysis Group, "Chinese Character Data Base," 2nd Edit., Taipei, Taiwan, R.O.C., May 1985.

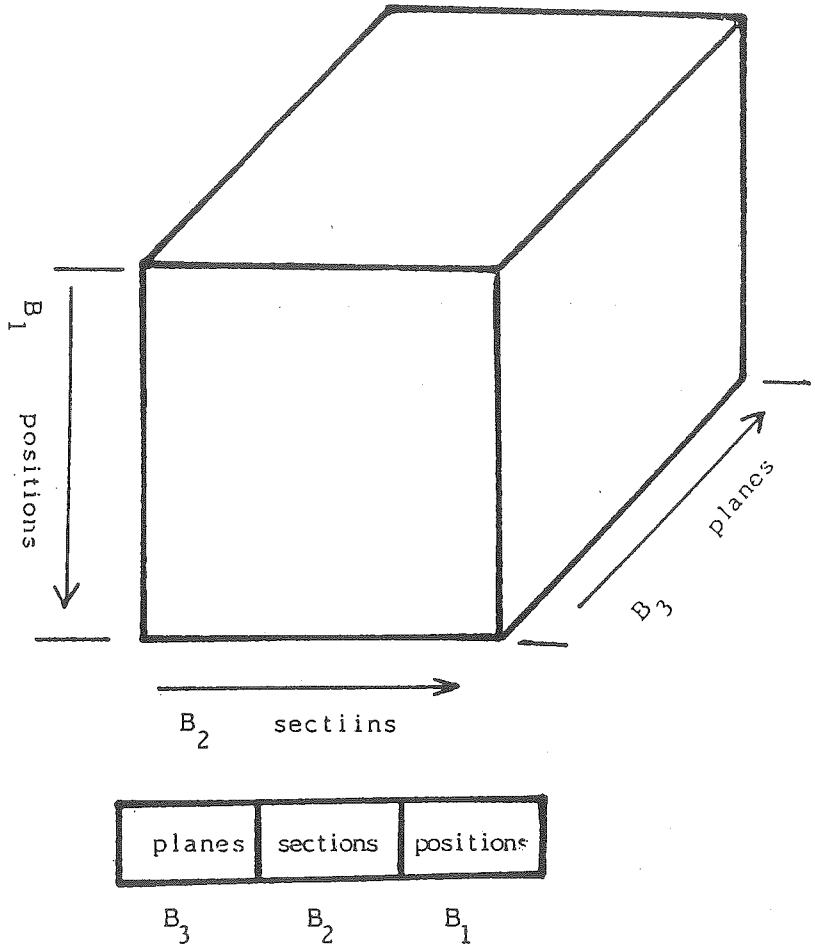


Figure 1. Coding structure of CCCII

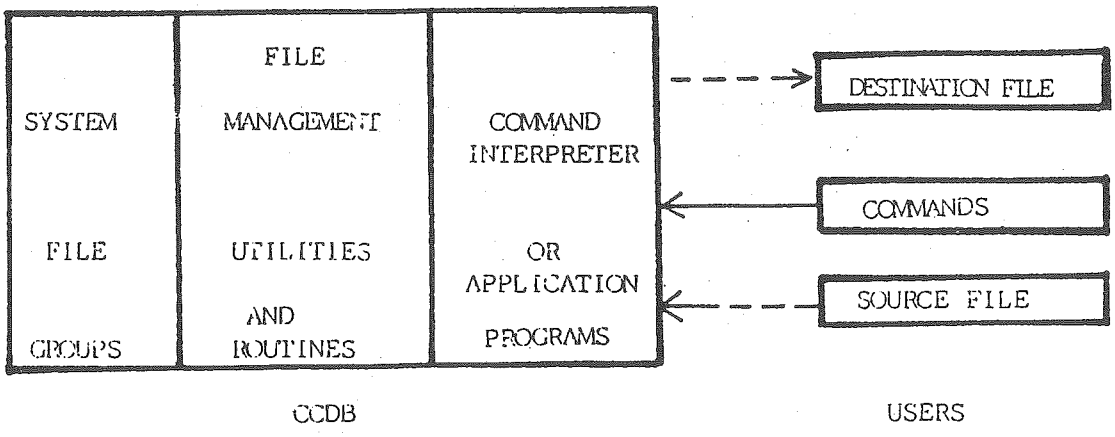
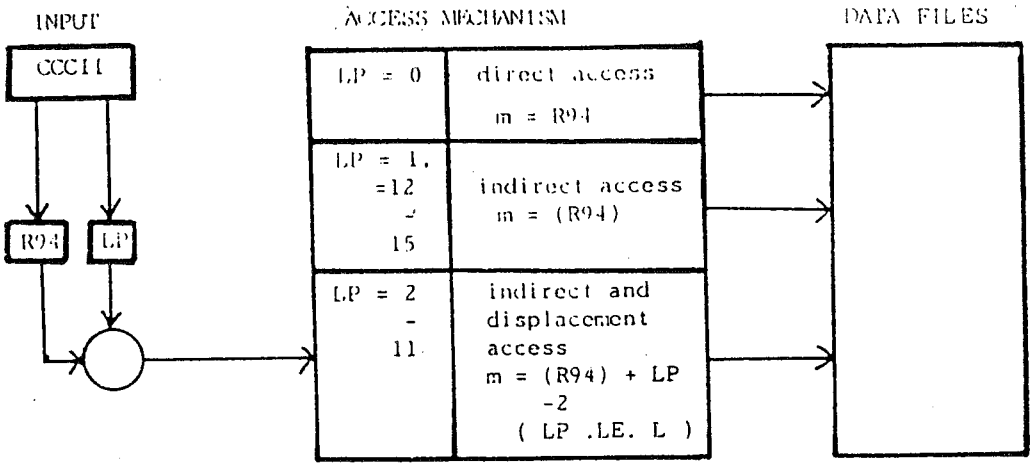


Figure 2. Software Structure of CCDB



LP: Layer indicator, from 0 to 15
L: The upper boundary of LP
m: The effective address of the character to be access
R94: Condensed code of CCCII

Figure 3. The Access Mechanism of CCDB

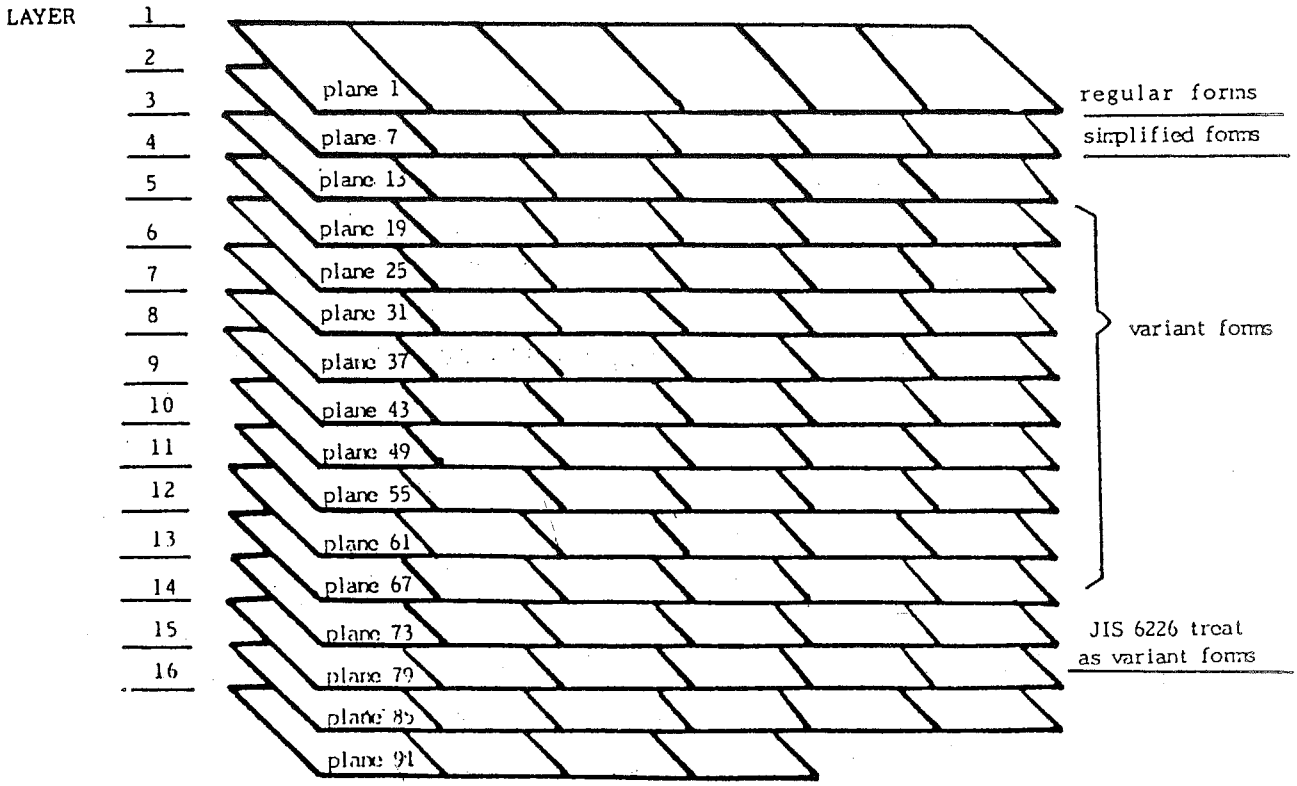


Figure 4. Structure and Allocation of CCCII

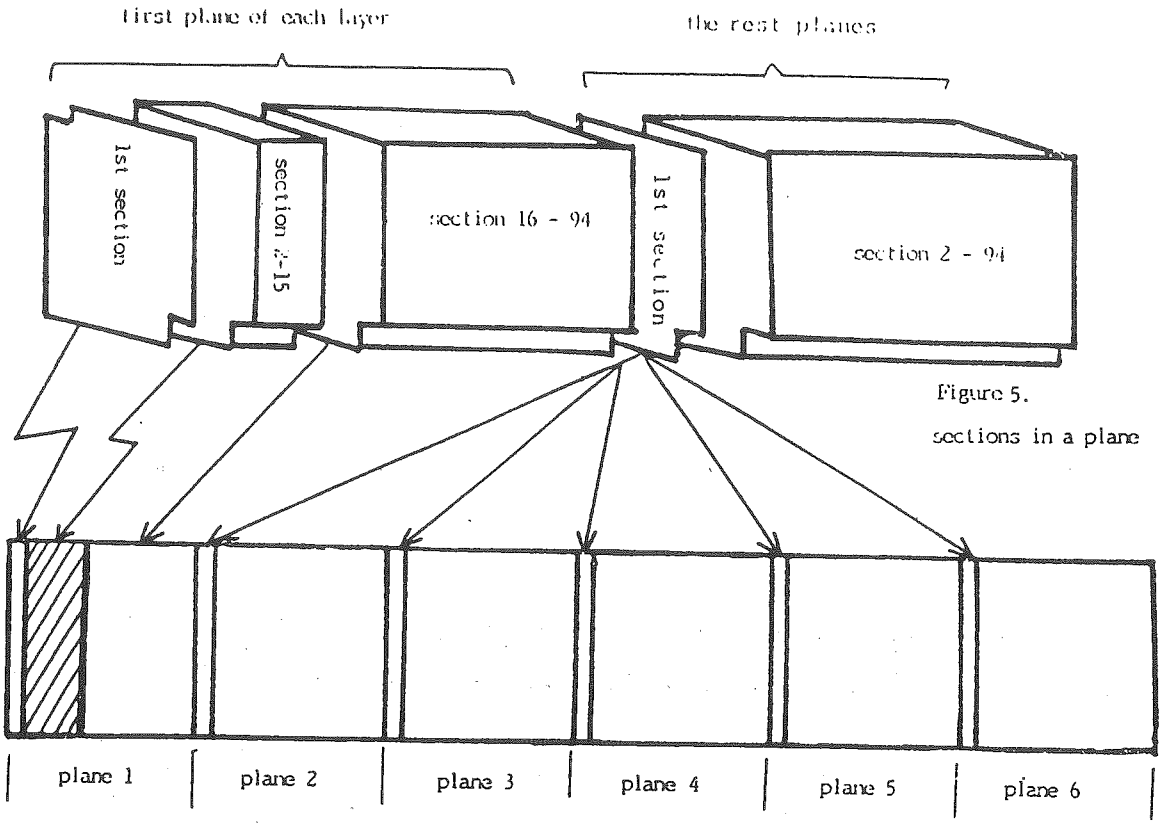
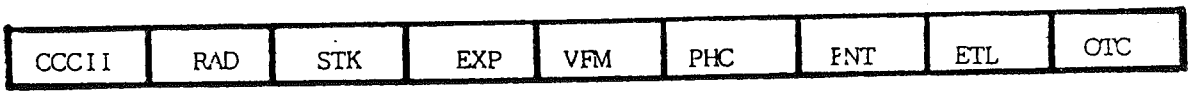


Figure 5.
sections in a plane

Figure 5 & 6. A Layer is Compsed of 6 planes
the 1st section of each plane is recerved for the control codes of CCCII
section 2 – 15 of the 1st plane of layer 1—12 are reserved for user area

R/C	2	3	4	5	6	7
0			DWO	SCH		PSL
1			DST	SLN		PSR
2			DPR	CHS		PSU
3			DBL	CSR		SSL
4			DSE	CFS		SSU
5			DCH	CFR		
6			DPA	CDR		
7				CHL		HOM
8			DLN	HLR		
9			DPG			
10			DVO			
11			DCU			
12			DTT			
13						
14						
15						

Figure 7. Control Codes of CCCII



CCCII: Chinese Character Code for Information Interchange
RAD: Radical which a character belongs to
STK: Total stroke count
EXP: Component expression of a character
VFM: Variant forms of a character
PHC: Phonetic sysbols of a character's pronounciation
FNT: Dot matrix representation of the FONT of a character
ETL: External or Indexing code of a character
OTC: Codes of different coding system

Figure 8. The Attributes Provide by CCDB