A) 33

Seminar On Library Automation and Information Network 1988
June 6-12, 1988
20 Chung Shan S. Road, Taipei, 10040. Taiwan, R.O.C.

The Recent Developments and Perspectives of CCCII

by
C.T. Chang, Jack Kai-tung Huang, C.C.Hsieh, and C.C. Yang

Chinese Character Analysis Group
Council for Cultural Planning and Development
Taipei, Taiwan, R.O.C.
June 9, 1988

1. Historical remarks of Chinese Character Code for Information Interchange (CCCII).

The Chinese Character Analysis Group (CCAG) was formed on December 25, 1979. The purpose of this research work were outlined as: (1) Categorizing the Chinese characters, (2) Identification of the Chinese punctuations, (3) Identification of the correct character shapes, (4) Pronunciation of the characters, (5) Organization and deduction of the character structures, (6) Indexing methods (7) Research and Development of the methods to teach Chinese, (8) Establishing the Chinese Character Code for Information Interchange, and (9) Eastablishing a comprehensive Chinese Characte Data Base. It's main objective is to provide research information for the applications of related professionals internationally. Different experts were assigned to undertake different subjects.

Chinese is an ideographical language, and there is no one in the world can say definitely how many Chinese characters are in existence today. However, in the COMPREHENSIVE CHINESE DICTIONARY by Dr. Chi-yun Chang, there are 49,889 characters, both the orthographic and the simplifed forms are included. Certain Chinese characters have one or more variants and/or simplified forms. They usually have exactly the same pronunciation and meaning, but are different in their stroke images. Usually, they are interchangeable in writing. But, when they are used as identifiers to name persons, places or things, they are considered as different characters, and should not be interchanged. The relationship between the orthographic form of a character and its variant and/or simplified forms, can in many cases have important implications. It is a very important consideration during the design of the overall coding structure. Simplified form of the Chinese characters are part of the variant forms of the Chinese characters. As previously defined, they are all collected, coded, and allocated in the second layer in this coding scheme. Up to now, there were 74,000 Chinese characters collected and 53,000 of them were coded and published following the coding scheme.

In April of 1980, the CCAG published the first edition of the Chinese Chararacter Code for Information Interchange which contained the complete coding structure of CCCII. It remains the same as today. It includes 4,808 Most Frequently Used Character set and 41 phonetic and tonal symbols [1].

2. The arrangement of the CCCII Chinese characters.

All orthographic forms of the Chinese characters, which are estimated to be about 50,000 and hence could be placed within the first layer. They are classified into four sets based on their usage frequency (provided by the Ministry of Education, The Republic of China) as follow:

(1) Most Frequently Used Character Set: contains 4,808 characters.
(2) Next Frequently Used Character Set: contains 17,032 characters.
(3) Rarely Used Character Set: contains 20,583 characters.
(4) Suplemental Character Set: unknown number.

The order of Chinese characters for each character set mentioned above is arranged according to the established RADICAL and STROKE COUNT sequence. The three sorting keys are defined as follows:

(1) First sorting key: The natural sequence of RADICALS of the Kang Shi Dictionaary RADICAL system. It contains 214 RADICALS ordered in ascending STROKE-COUNT sequence.

(2) Second sorting key: The ascending order of the STROKE-COUNT excluding the RADICAL of the character.

(3) Third sorting key: The precedence of the STROKES excluding the RADICAL of the character. The arrangement of the precedence of STROKES is defined as:

(a) A dot (dean), weight is l.

(b) A horizontal stroke (herng), weight is 2.

(c) A vertical stroke (jyr), weight is 3.

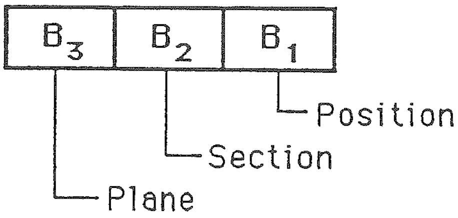
(d) A stroke down from the upper right to lower left (piee), weight is 4.

(e) A stroke down from the upper left to lower right (nay), weight is 5.

3. Code structure of the CCCII.

The CCCII and its communication system is based on the ISO 646 communication 7-bit coding standard. It utilizes three 7-bit bytes to represent a given Chinese character, figure 1. Its technique of code extension and the identifying location of the escape sequence are based on the ISO 2022.

Figure 1 Three 7-bit Byte extension--- ISO 2022

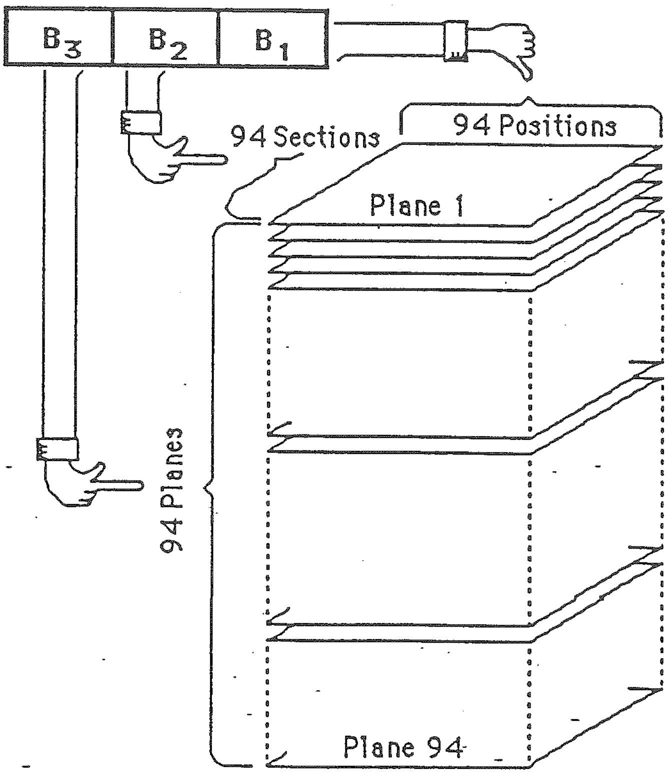


This protocal provides a three dimension space, i.e., 94x94x94 coding positions, for the Chinese characters used in computing and its development of the escape sequence. This method could encompass more than 50,000 Chinese characters. Meanwhile, it is also suggested that the same multiple byte coding scheme of the GRAPHIC SYMBOLS should be used for the symbols of the escape sequence.

It is planned that all the current Chinese characters, including all VARIANT and SIMPLIFIED characters, will be coded with a three byte code. The frequency of usage and the classification of characters are used as the guiding principles.

The code, consisting of three 7-bit byte vectors, forms a finite three dimensional coding space, i.e., it provides a total of 94x94x94 positions according to the ISO 2022. It is called a space of 94 PLANES, with 94 SECTIONS in each PLANE, and 94 POSITIONS in each SECTION, figure 2.

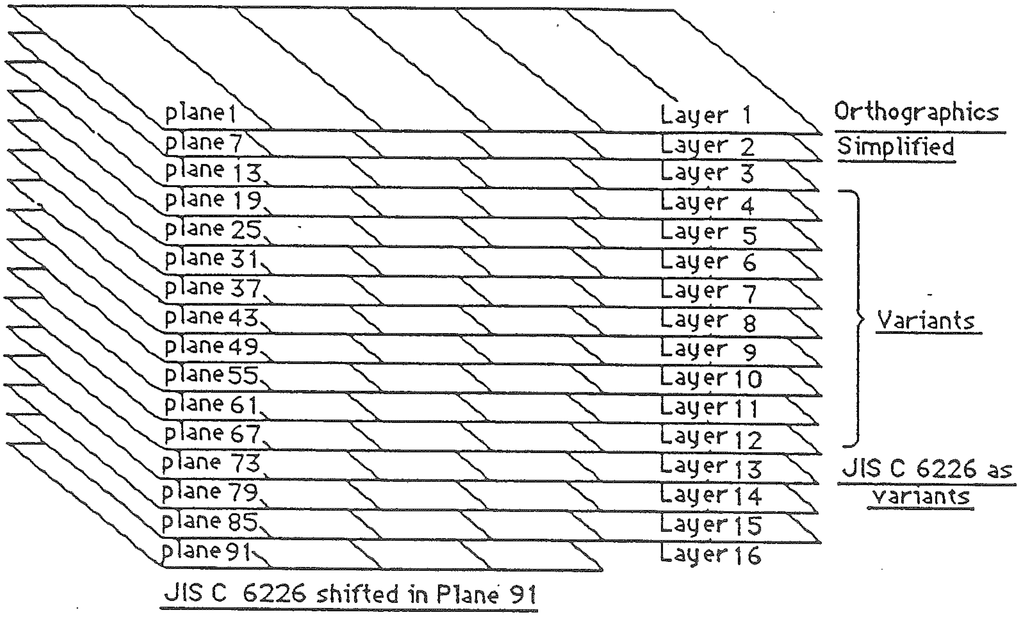
Figure 2 Three dimensional (94x94x94) Coding Structure



by C.T. Chang, Jack Kai-tung Huang, C.C.Hsieh, and C.C. Yang, the Chinese Character Analysis Group

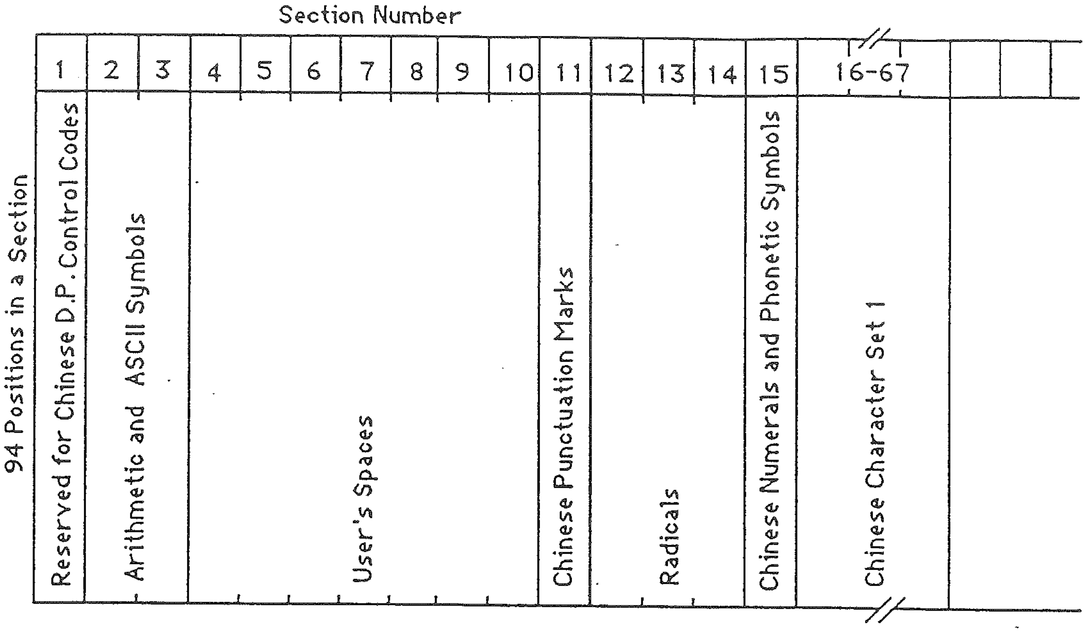
Since certain characters have VARIANT and SIMPLIFIED forms, the coding spaces have been subdivided into layers so that the character codes of the VARIANT and/or SIMPLIFIED form of a given character could be used to trace and to link back to the original characters and vise versa. Such relationships between the orthographic characters and their VARIANT forms could in many cases, have important implications. The 94 PLANES are grouped into 16 LAYERS. Each LAYER, from 1 through 15, is made of 6 consecutive PLANES, while the last LAYER (16th) has only 4 PLANES, figure 3.

Figure 3 Sixteen LAYERS of the CCCII Coding Structure



Each LAYER, thus, can have 53,016 (=94x94x6) coding spaces. However, the first SECTION of each first PLANE of each layer is reserved for the CONTROL CODES that are required in handling Chinese character strings. This reduces the total coding space for Chinese characters in each LAYER to 52,452 (=14x94) POSITIONS per LAYER, and it has a total of 19,740 (=15x1,316) POSITIONS in the whole coding space for CONTROL CODES. These areas are also used for the user's collection of a special set of characters in the private field and/or privately designed new symbols or characters. Therefore, each LAYER actually only has 51,136 (= 52,452 - 1,316) usable coding spaces, figure 4.

Figure 4 Structure of Each LAYER

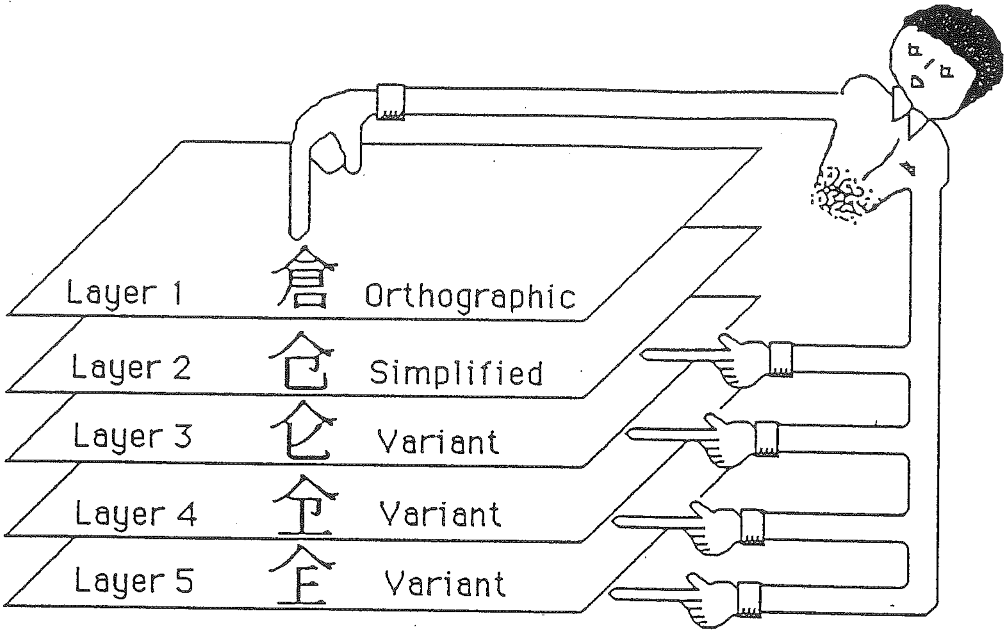

Section Number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16-67
94 Positions in a Section
Reserved for Chinese D.P. Control Codes
Arithmetic and ASCII Symbols
User's Spaces
Chinese Punctuation Marks
Radicals
Chinese Numerals and Phonetic Symbols
Chinese Character Set 1

The usage of these 16 LAYERS is described as follows:

(1) LAYER 1, i.e., (PLANE 1 through 6) is used to designate the GRAPHIC CHARACTERS of the arithmetic and ASCII symbols (PLANE 1, SECTIONS 2-3), 35 Chinese punctuation marks (PLANE 1, SECTION 11), 214 RADICALS (PLANE 1 SECTION 12-14), 41 Chinese numerical characters, 37 Chinese PHONETC SYMBOLS and TONE MARKS (PLANE 1, SECTION 15) [special note: up to this point, it's in the user area of the first LAYER], the 4,808 MOST FREQUENTLY USED CHINESE CHARACTERS (PLANE 1, SECTION 16-67), the 17,032 NEXT FREQUENTLY USED CHINESE CHARACTERS (PLANE 1, SECTION 65 to PLANE 3, SECTION 64), and all other orthographic forms of Chinese characters (PLANE 3, SECTION 65 to PLANE 6, SECTION 5).

(2) LAYER 2, i.e., PLANES 7 through 12, is used for designating the SIMPLIFIED forms of Chinese characters which are used in the mainland China.

Figure 5 Relationship between Orthographic and Variant characters



(3) LAYER 3 through 14 are used to designate the VARIANT forms of the Chinese characters that appeared in LAYER 1. The handling of these VARIANT forms will be discussed in the following sections. Currently, only up to the 12th LAYERR are used for VARIANTS. The JIS C 6226 characters are treated as VARIANT forms and have been placed in LAYER 13. LAYER 14 is reserved for the Chinese characters used in Korean.

(4) LAYER 15th is reserved for other usages.

(5) The 16th LAYER is the coding area for other languages which are most closely correlated to the Chinese language. Also, the SHIFTED JIS C 6226 KANJI characters are placed in PLANE 91 of this LAYER. PLANE 92 is reserved for the Korean KIPS. The 93rd PLANE is reserved for the supplementary Chinese characters. The last PLANE (94) will be used to contain the non-Chinese characters of CB2.

4. Current Status

(1) The Most Frequently Used Character Set:

April 15,1980, the first edition of the Chinese Character Code For Information Interchange was published and the complete coding structure of CCCII was announced by the CCAG.

(2) Next Frequently used Character Set:

In October of 1982, Second edition of the CCAG's works were published [2]. It contains 33,544 characters, in which 21,885 were orthographic characters and

11,660 VARIANT chaaracters, which in turn contained 3,752 SIMPLIFIED characters. A separate book on the VARIANT characters [3], and a printed form of the Chinese Character Data Base were also published at the same time [4].

(3) Rarely Used Character Set:

In February of 1987, The CCAG released the latest updated revision in a similar format of the earlier books [5]. However, in this set, it contains 20,583 Rarely Used Characters. This brings the total published number of characters to 53,940.

The Chinese Character Analysis Group has planned to release the the most comprehensive set of publications, which will contain information of more than 74,000 characters CCAG collected, and analyzed throughtout these many years. No one can be absolutely sure this will contain "all" Chinese characters, but it will be the most complete job ever accomplished in this field.

Based on preliminary survey and statistics, the collection made by Census Automation Project of Taiwan there are at least 12,000 to 18,000 Chinese characters that were estimated by the Project, are outside the collection of 48,174 characters collected by the Ministry of Education previously. Now these characters are under analysis by the CCAG, and will be coded and included in the future Variant Characters and the Rarely Used Character Set of CCCII.

5. Chinese Character Data Base (CCDB)

The Chinese Character Data Base (CCDB) provides more than 20 attribute indexed files of the CCCII with cross-references. Any character can be searched by the following methods, such as:

- Pronunciation
- Radicals
- Stroke-account
- Four Corner Code
- Three Corner Code
- Telegraph Code
- Big-5 Code
- Dragon Code
- General Chinese Character Standard Interchange Code

The publications of CCDB were released in:

- Second Edition, October, 1982
- Revised Edition, May, 1985
- Edition of CCDB for Rarely Used Characters, Feburary, 1987

6. The usage of the First Layer of CCCII

Layer 1, i.e., Plane 1 through 6, is used to designate the GRAPHIC CHARACTERS of the:

(1) Arithmetic and ASCII Symbols (Plane 1, Section 2-3)

(2) 35 Chinese punctuation marks (Plane 1, Section 11).
(3) 214 Radicals (Plane 1, Section 12-14).
(4) 41 Chinese numerical characters, 37 Chinese Phonetic Symbols and Tone marks (Plane 1, Section 15).

[special note: up to this point, these are in the user area of the First Layer].

(5) 4,808 Most Frequently Used Chinese Characters (Plane 1, Sections 16-67, Hexadecimal values starts from 213031 to 216330).
(6) 17,032 Next Frequently Used Chinese Characters (Plane 1, Section 68 to Plane 3, Section 64, Hexadecimal values starts from 216421 to 236072).
(7) 20,583 Rarely Used Characterse (Plane 3, Section 65 to Plane 6, Section 5, Hexadecimal values starts from 236121 to 262543).

Above, in total, there are 42,423 orthographic Chinese characters.

In Plane 6, there still 8,366 coding spaces (i.e., 89 Sections) remaind unused, these are reserved for extention of the orthographic characters.

7. Up to this point, 42,423 orthographics and 11,517 variants, 53,940 Chinese characters in total were coded in the CCCII by the CCAG.

8. Looking into the Future of the CCCII

(1) Comprehensive organization of the Variants: We are re-working on the alreday published 11,517 variants and their corresponding Rarely Used Characters (about 20,000 characters) and 12,000 characters used in people's name but without clearly known meaning or pronunciation. During this process, we found that the variants of many Rarely Used Characters already exist as orthographics in the volume I and II. Many of these variants are borrwoed or synonym, particulary the simplified characters from mainland China. The shape symbol of many characters were deleted with only the sounding symbol left. In order to be consistent with one-character-one-code pricinple, these variants need some further works. And that may influence the orthographics. This difficult task is expected to be finished within 9 months.

(2) The collections of Japanese JIS and Korean KIPS codes and character cards are completed. We are working on these characters following the REACC CJK project of RLIN (Research Library Information Network), and placed them into the CCCII coding space. The third examination is undergoing. There are about 500 characters in Layer 7 and 8 of the REACC are under re-examination.

(3) We are collecting Chinese minority characters, such as Manchurian, Mongolian, Tibetain and Moslem. It is expected to be included in the CCCII at the end of 1990.

(4) We are also collecting all ISO registered languages. They will be included in the CCCII to ensure that the CCCII can be used in a multi-lingual and library environments.

(5) Developing the necessary software tools, including all available input methods and internal codes. These will be included in a cross-index table.

(6) Producing a graphic based, resolution independent Chinese character generator to produce different fonts with different size without distortion. This can be used in the publishing industry.

9. Conclusions

The design of the CCCII already included many considerations. We can not say that this is the best, but we also did not see anything better than this yet. After a long period of endurance, due to the limitation of manpower and equipment,we finally obtained gracious assistances, encouragements and supports from the Council of Cultural Planning and Development, the elders and authorities of the country, and many intellegent people from abraod and domestic. During these long 8 years, no matter under what kind of criticism or even abusement, we still keep going with our original vision. Now, we start to see some of our results and deserved recognitions which are the best encouragement to us. Recently, we have learned that the CCCII/EACC (East Asian Character Code) is in the process of becoming an ANSI standard. Now, may we present all what we have to the public to express our most sincere gratitute.

References:

1. Chinese Character Analysis Group, Chinese Character Code for Information Interchange, Volume I, April, 1980
2. Chinese Character Analysis Group, Chinese Character Code for Information Interchange, Second Edition, October, 1982
3. Chinese Character Analysis Group, Variant Forms of Chinese Character Code for Information Interchange, Volume II, Second, December 1982
4. Chinese Character Analysis Group, Chinese Character Data Base, Second Edition, October, 1982
5. Chinese Character Analysis Group, Chinese Character Code for Information Interchange, Volume III, February, 1987