A Survey of Various Chinese Ideographic Character Sets for Information Exchange

Hengsheng Cai
Dept.of Systems and Computer Engineering, Carleton University
Ottawa, Canada, K1S 5B6.

Abstract

This paper explores various Chinese ideographic character sets in use in different languages, countries and regions for information exchange. The paper analyzes their respective charcteristics and limitations, compares their commonness and difference. Finally the author presents his opinion on how to treat various national/regional Chinese characters standards for videotex/teletext internationalinteractive information service.

1.Introduction

Chinese characters are now in use in several languages, such as Chinese, Japanese and Korean..., in many countries and regions of the world, including China, Taiwan, Hongkong, Japan, Korean, Singapore and other areas in Asia, Europe, America, Australia. Chinese characters which are used in different languages are called different names, for example, "Chinese character" in Chinese is known as "Hanzi", "Chinese character" in Japanese is known as "Japanese Kanji", "Chinese character" in Korean is known as "Korean Hanja". They are similar in terms of fonts, but not the same. Although the Japanese Kanji and the Korean Hanja are both originally borrowed from Chinese characters, they have undergone considerable migration both in terms of fonts and in their meanings. Some of Chinese characters (or phrases) may look the same, but they mean different things in different areas, some characters (or phrases) in one country do not have the corresponding meanings in the other country. Chinese Hanzi has a long history. With the evolution of society, a large number of ancient characters have been abandoned or simplified, and a great quantity of new characters have been created. In total the accumulated amount is about 60,000 characters, but the frequently used Hanzi number about 6,000. On historical reasons, Mainland China mainly uses simplified Chinese characters except for bibliographies and ancient Chinese literature, while Taiwan uses traditional complicated Chinese characters. Hongkong now uses traditional Chinese characters, but will probably change to simplified Chinese characters after Hongkong returns to China in 1997. Most Chinese outside China still use traditional Chinese characters, but Singapore has changed to simplified Chinese characters same in Mainland China. Because of the differences of social and cultural background as well as different requirements for Chinese character sets, national/regional

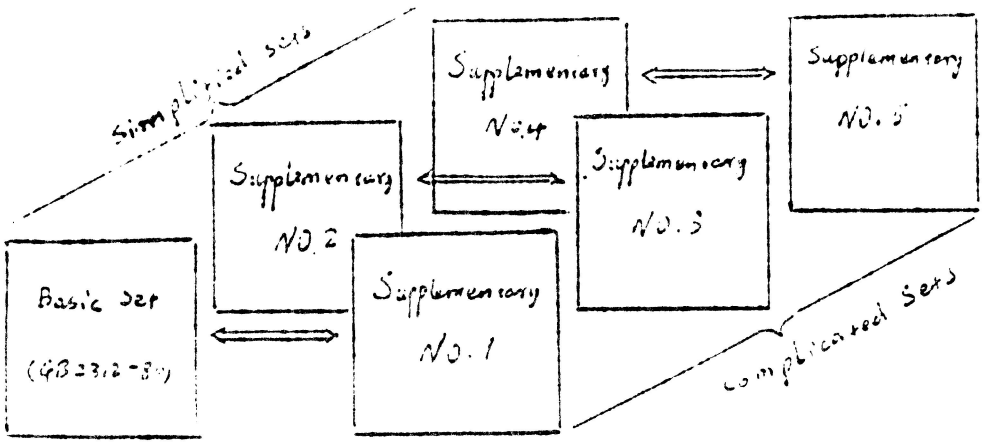
standard codes for individual Chinese character sets have been developed in Japan, China, Taiwan and Korea. Different computer manufacturers support individual standards to the corresponding Chinese character sets, for example, IBM uses a shifted version of the GB 2312 set in its personal computers 5550 series for the People's republic of China, and uses a shifted version of Big 5 and of JIS for Taiwan, Hongkong and Japan respectively.

2.The four main standards

There are four main national/regional standards for Chinese characters in existence. The national standard chinese character set for Japan is JIS X0208. The part of the character set for Chinese ideographic characters(Kanji) is divided into two roughly equal parts. The first part consists of 693 special characters and 2,965 Japanese Kanji characters which are sorted phonetically according to the predominant reading of the character. The order of the phonetic sort is that of the Japanese syllabary(aiueo, kakikukeko, etc). The second part consisting of 3,388 Kanji characters is sorted by radicals(214 Kang xi radicals -- from "Kang xi dictionary", a famous Chinese dictionary) and sub-sorted by stroke count. JIS X0208 has been registered as the updated Japanese Kanji set by CCITT. Besides the set, the original 1978 Japanese Kanji set, JISC 6226, was also registered. According to ISO 2022, the sequence ESC 2/4 4/2 and ESC 2/4 4/0 is used to designate the two Japanese Kanji sets respectively.

The Korean Standard Information Interchange Code, KSC-5601, is promulgated by the Korea Standards Bureau of the Industrial Promotion Agency. It includes 4,888 frequently used Chinese ideographic characters(hanja) which are arranged in standard phonetic order according to the order of the Korean syllabary(hangul).

The standard hanzi sets including simplified Chinese characters and traditional complicated Chinese characters should be unified, but they have not been unified until now because of political problems. Taiwan established its own standard CNS-11643 in traditional complicated Chinese characters in 1986, while the People's Republic of China announced that GB2312-80 is the national standard for Chinese graphic characters in 1981. For the latter, only the first part of the code, the basic set of simplified Chinese characters or the Primary Set, has been published, it also has been registered with CCITT. This Primary Set is designated and invoked in the 7-bit coded character set by means of 3-charactere escape sequence ESC 2/4 4/1 and shift-in character(SI) respectively. Whthin this Primary Set, there are 7,445 graphic characters including 682 non-Chinese graphic characters and 6,763 Chinese ideographic characters which are arranged in two sequences. The first sequence which is of the 3,755 most used characters is arranged in the phonetic order of their standard Chinese pronunciation(Pinyin). The second sequence of 3,008 characters is arranged in radical order according to a list of 186 radicals, then subarranged by stroke count of the part of the character exclusive of the radical and then further subarranged by stroke type. The order of stroke type is according to the five basic radical order: horizontal stroke(一), vertical stroke(丨), stroke down to the left(丿), dot(丶) and twist(乙,𠂇,乚). Besides the basic set of Chinese characters, the compilation of the supplementary sets of Chinese characters has been accomplished, which will be officially announced in the near future. These supplementary sets of Chinese characters are divided into two series. Series of simplified



characters include the basic set(GB2312-80), supplementary set No.2 and supplementary set No.4. Series of traditional complicated characters include supplementary set No.1, supplememtary set No.3 and No.5. The two series are image reflections of each other. The respective simplified and complicated characters having the same meaning are provided with the same coding number. The following figure shows the relationship between respective sets.

CNS-11643 has been promulgated by Taiwan since March 1986. This code contains 13,051 characters arranged in two sets. The first set contains the 5,401 most frequently used characters together with 651 special symbols, while the second set contains the 7,650 next most frequently used characters. Both sequences are in the order of the total strokes in the character and subsequenced by the traditional 214 Kang xi radicals. Before CNS-11643 was published, BIG-5 had been the most common code set for traditional complicated Chinese characters used by computer manufacturers in Taiwan. It was published by the Institute of Information Industry in 1984. It has three parts -- the graphic symbols(441), the most common characters first (5,401) and then the next most common characters (7,652), intotal 13,494 characters. Within these sections the characters are arranged by total stroke count first and then by radical (214 radical Kang xi system). Since the BIG-5 code is a de facto manufacturers standard not a government controlled exchange code, anomalies can be found among its various implementations. For example, most manufacturers provide some additional characters beyond standard Chinese characters. Although CNS-11643 is a national standard for information interchange, it has not been universally adopted as a standard internal coding system by various manufacturers in Taiwan. This is because : (1) the differences between CNS-11643 and BIG-5 in the size of the character sets and the Hanzi sequences; (2) no simple corresponding relationship between the two coding systems; (3) the purpose of CNS-11643 is only for information interchange, its coding system can not identify whether Chinese characters or ASCII characters; (4) not enough room for users to create characters. In December of 1987, the commercial professional computer union proposed a new internal code system, called the union code, the sequence and the character sets of the union coding system are according to the CNS-11643, it, therefore, is easy to switch between these two coding systems. The first set of the union code contains

8,836 characters, including graphic symbols(651), control codes(33) and the most used Chinese characters(5,401), the second set includes the next most used Chinese characters(7,650). The internal code system will probably be adopted by various organizations in Taiwan.

3. Summary

Given that many countries/regions have established their own Chinese character code sets for information exchange, it is impossible to choose one character set from these countries/regions or try to establish a new Chinese character set as commonly-accepted international standard for Chinese characters. From the cultural point of view, Chinese, Japanese and Korean scripts should be treated as separate and distinct scripts and respective national agencies are the principle control and standard-setting organizations. Because of a more political than technical problem, GB2312-80 and CNS-11643 for Chinese Hanzi should be respectively applied for Mainland China and Taiwan at present. For videotex/teletext international information service, we should choose a corresponding Chinese character set as a Gset suited to the requirement for different countries/regions, and it is listed as follows:

(1) JIS X0208 set for Japanese kanji, for use in Japan.

(2) KSC-5601 set for Korean hanja, for use in Korea.

(3) GB 2312-80 set for Chinese simplifiedHanzi, for use in People's Repubicof China, Singapore, Hongkong, the United Nations, other areas and international organizations.

(4) CNS-11643, BIG-5, or Union Code for traditional complicated Hanzi, for use in Taiwan and Hongkong.

According to the principles and concepts of the ISO 2022 Code Extension Standard, the NAPLPS (North American Presentation Level Protocol Syntax) videotex/teletext standard has included an extension to all other languages. The extension of NAPLPS, in two sets termed C-NAPLPS and J-NAPLPS for Chinese Hanzi and Japanese Kanzi respectively has been proposed. Using these principles, we can further mixed NAPLPS with Chinese traditional complicated Hanzi and Korean Hanga sets. These national/regional standards for Chinese characters are similar in structure and are 2 byte coding systems, most characters are the same or image reflections in terms of fonts and meaning. For international information exchange, the rational and practical way is to establish a commonly-used conversion table, which is for online or offline exchange while admitting the character sets used by various countries and regions. An ideal internal code model runs under the same OS environment and allows various Chinese character sets to be called, which can coexist through the use of various Rom chips in megabits or Chinese characters boards containing the complete Chinese character sets. Some computer manufacturers in China, Taiwan and Japan have provided these types of products respectively.

References:

1. Jack Cain, "A Servey of Character Sets That Include Chinese Ideographic Characters With Special Reference to Sort Algorithms and the Retrieval of Data",The International Symposium on Standardization for Chinese Information Processing (SCIP-89), Beijing, March 21-23, 1989.

2. Zhu Yan, Chen Yaoxing, "Chinese MARC and its Processing", First International Conference on Scholarly Information Network, Tokyo, December 8-11, 1987.

3. Jin-Tuu Wang, "Comment Paper - from the Cultural Aspect", First International Conference on Scholarly

Information Network, Tokyo, December 8-11, 1987.
4. Hengsheng Cai, C.Douglas O'Brien, and J.SpruceRiordon, "A Proposed Scheme for Chinese Videotex Standard", Computer Processingof Chinese & Oriental Languages (ISSN 0715-9048), Vol.2, No.4, 181-197, October 1986.
5. Hengsheng Cai, J.Spruce Riordon, and C.Douglas O'Brien, "A Proposed Chinese Language Videotex and Teletext Standard", International Conference on Communication Technology, Nanjing, China, November 1987.
6. Y.F.Lum, Hengsheng Cai and J.S.Riordon, "Extension of NAPLPS Videotex/Teletext Standard to the Chinese Language", ICCC Symposium'89, Beijing, China, September, 1989.