Supporting Chinese Character Variants in Hong Kong through Ideographic Variation Sequence

Qin Lu[1], Kwan Hin Cheung[2], Dan Xiong[1], Shing Yu[1], Jian Xu[1]

[1]Department of Computing, The Hong Kong Polytechnic University

{csluqin, csdxiong, cssyu, csjxu}@comp.polyu.edu.hk

[2]Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University

kwan.hin.cheung@polyu.edu.hk

Abstract

This paper will introduce an ongoing project in Hong Kong that makes use of the Ideographic Variation Sequence (IVS) and the associated Ideographic Variation Database (IVD) developed by the Unicode Consortium for character glyph registration. Hong Kong uses the traditional Chinese writing system similar to that of Taiwan and thus used the Big5 encoding for many years. But, Chinese characters used in Hong Kong do have different variant glyphs. The current CJK unification process for the ISO10646/Unicode standard can cause confusion and inconvenience in certain applications. There are applications where different written styles of the same logical character may need to be included in the same document requiring these variants to be separately encoded and specified at the character level. The IVS and IVD developed by the Unicode Consortium for character glyph registration are quite suitable for Hong Kong's Chinese variant specification. This paper will explain how the IVS technology is used to encode these Hong Kong specific Chinese variants, the process of the review, the production of variant, and the production of the Hong Kong Character variant Specification, and the registration in the IVD of Unicode.

Keywords: Chinese variant, Ideographic Variation Sequence, Hong Kong Character glyph

1. Introduction

Current computer coding for Chinese characters is done at the character level. However, a single Chinese character may still have different glyph shapes, namely, variants. Generally speaking, Chinese character variants (異體字 or 多形字), refer to the set of different glyph shapes of the same character. Broadly speaking, the variants can substitute each other without change in the meaning of a given text [1].

Variants normally do not change the meaning of a character. Yet, if coded separately, they will cause problem in searching and indexing. Thus, in the ISO/IEC 10646 (Unicode) character standard, a so-called unification procedure [2-3] is introduced to identify the different glyph shapes of the same character systematically so that different variants are unified to single character for coding with one selected concrete glyph. This unification, however beneficial, can cause confusion and inconvenience in certain applications.

Hong Kong uses the traditional Chinese writing system similar to that of Taiwan and thus used the Big5 encoding for many years. However, Chinese characters used in Hong Kong do have different variant glyphs. However, there are also applications where the different written styles of the same logical character need to be included in the same document requiring these variants to be separately encoded and specified at the character level. The Ideographic Variation Sequence (IVS) and the associated Ideographic Variation Database (IVD) developed by the Unicode Consortium for character glyph registration are quite suitable for Hong Kong's Chinese variant specification.

This paper introduces an ongoing project in Hong Kong that makes use of the IVS and IVD for character glyph registration. The main objective of this project is to specify Chinese character variants at the character level and encode all Chinese character variants under the Big5 coding framework used in Hong Kong so that computer systems can include supports of these variants in different applications. Section 2 gives background on character encoding, variants, and IVS. Section 3 introduces the scope of the project and its workflow. Section 4 presents the principles for the variant specifications. Section 5 gives examples of the glyphs for IVD registration. Section 6 is the conclusion.

2. Background of the Project

The Unicode Standard avoids duplicate encodings of different variants of the same character by unifying them. According to the unification principles in the ISO10646/Unicode standard [2-3], typeface distinctions or local preferences in glyph shapes are considered as sufficient grounds for unification of characters. In Chinese, the character "bone" is a typical example, which takes three different forms due to local preferences, as shown in Table 1. All of the three forms are considered as the same character and encoded at U+9AA8 in the Unicode standard.

Table 1 Variation of the character "bone" due to local preferences

	Hong Kong Glyph	Mainland China Glyph	Taiwan Glyph
Kai	骨	骨	骨
Song	骨	骨	骨

Since the traditional Chinese writing system used in Hong Kong is similar to that of Taiwan, all computer systems have adopted the Big5 encoding scheme of Taiwan and has never attempted to independently develop its own coding standard. However, Big5 is designed without considering Hong Kong's specific needs. Although users have been coping with the limitations of systems based on Big5, the dependence of business activities on computers has highlighted the need to develop a character set for Hong Kong's specific needs, and to seek a better encoding system. The Government of the Hong Kong Special Administrative Region (HKSARG) published the Hong Kong Supplementary Character Set (HKSCS) in 1999 [4] as extension to Big5. Realizing the fundamental limitations of Big5 systems, the government has also tried to promote the use of ISO/IEC 10646/Unicode platforms [5].

In addition to character extension, Chinese characters used in Hong Kong have different glyphs. To officially make the Hong Kong style characters known for font vendors, the Hong Kong government has developed the reference guides for Chinese computer systems in Hong Kong for both the Song style and Kai style in 2002 [6-7]. Because Big5 is not a Hong Kong coding standard, it is inappropriate for Hong Kong to directly make the reference guide using Big5. Thus, the reference guides are specified on component basis without giving exhaustive list of HK style glyphs. In principle, these reference guides can be used in selected locales to support HK glyphs for character display.

Currently, in ISO/IEC 10646/Unicode, the glyph of a character can be different for each different locale such as the character bone (骨 for mainland China, 骨 for Taiwan) as computer systems can install different fonts for different locales. However, those character glyphs that are unified to the locale chosen shape cannot be separately coded, and thus cannot be used conveniently in computer systems except ad hoc solutions. This issue in principle is not a coding issue, but an implementation issue. Consequently, its solution falls into the scope of the Unicode Consortium (Unicode for short) which in recent years has developed the IVS and the associated IVD.

The IVS is based on ISO/IEC 10646/Unicode the variation selectors coded in U+E0100 to U+E01EF which has a total of 240 codepoints. Variation selectors are categorized as non-printable characters and they cannot be used as single independent characters. To represent an ideographic character variant, a variation selector must be used in combination with a normal printable CJK unified ideograph character to form the IVS. For example, the code sequence <U+4E00, U+E0100>[1] as a tuple is considered an IVS referring to a variant of the character "一"[2]. For any given ideographic character, the use of IVS can define up to 240 different variants (thus in all 241 in total including the normal character defined by a single Unicode codepoint).

The provision of IVS can mainly serve for two purposes. First, it provides a way to describe an ideographic character variant in text form and the variation selector (the second component of the IVS) is by default ignorable by Unicode compliant systems which means that the variation selector should normally not be displayed. In other words, it should not be displayed by default. If an implementation supports variation sequences, and the selected font does as well, then the correct glyph should be displayed. When an ideograph variant is defined, its text behaviour is exactly as that of the normal Unicode character (the left component). Therefore searching and indexing of a variant can be done just like its corresponding character. Second, IVS provides a method to define ideograph variants which are otherwise unified and cannot be displayed. For rendering purpose, however, if the system supports the interpretation of the IVS must be defined in the Unicode IVD, the IVS can be rendered as a single variant character according to its glyph definition.

For example, the two Japanese ideographs 辻 and 辻 are variants of each other. In fact, the Unicode codepoint U+8FBB is assigned to the second glyph 辻 (this one is printed using system

[1] >The symbols "<", "," and ">" are for notations only and are not part of the coded text.

[2] http://www.unicode.org/ivd/

font) and the first one 辻 is unified to it. Historically, the first glyph was published earlier in the Japanese standard (JIS X 0208-1990 36-52). However, as the result of a revision/normalization of Japanese glyphs, the second glyph is defined in a later Japanese standard (JIS X 0213:2004 1-36-52). To support the first glyph, a viable solution is to officially include the IVS sequence <U+8FBB, U+E0100> = 辻 into Unicode's IVD.

Because of the default behaviour of Unicode system, a variant of a character must be close in shape with the ideograph character. In fact, only unifiable characters under the ISO/IEC 10646 unifications in its Annex S as well as it extensions in the IRG working group under the IRG Working Document Series can be defined for a given ideograph character. Each IVS sequence must be officially submitted to Unicode and accepted and then included in the IVD. Unicode compliant systems will consider any IVS not found in the IVD illegal and thus will only take the default interpretation (ignoring the variation selector) without taking the IVS as a variant.

3. Project Scope and Workflow

It is under the IVD framework that the Hong Kong government decided to pursue the Hong Kong variant specification which is chartered to this project. In principle, our work covers all Chinese characters used in Hong Kong, thus including all characters in Big5 and those in HKSCS [4]. Since HKSCS characters already adhere to reference guide of the Hong Kong character glyphs and are not unifiable to any character in Big5, the registration of IVS in Hong Kong should only have characters in the Big5 repertoire.

As Figure 1 shows, our workflow for the standardization falls into four major parts. In the first part, Hong Kong character glyphs with all the characters in the Big5 standard [8] will be reviewed and analysed to find their differences by the project team in accordance with the existing specification and guides [6, 7, 9] and some other references [10-12]. In the second part two types of Chinese character fonts used in Hong Kong, that is, Kai and Song, will be reviewed and developed based on the analysis results. Kai font is useful for teaching of Chinese characters as it maintains the traditional Chinese brush style of writing. The Song font (also called MingLiu font in Hong Kong and Taiwan) is used in ISO/IEC 10646's code chart and generally considered printed form of Chinese. The fonts from collaborating vendors will be reviewed and produced according to analysis results. The project's industrial collaborators will then produce the revised fonts after these variants are identified. This deliverable will be reviewed and finalized in phases to allow overlap of development and review. It is planned that in addition to the review by this project team, review by the Chinese Language Interface Advisory Committee (CLIAC) under Office of the Government Chief Information Officer (OGCIO) of HKSAR government will also be carried out in different stages. In the third step, the Hong Kong character glyph specification and glyph list will be prepared, in which the differences between the Chinese character variants in Hong Kong and the characters in the Big5 standard will be illustrated at the character level. The documents will be passed to CLIAC as an official file for review and approval. In the last step, the list of variant characters will be submitted to the Unicode standard for inclusion in its IVD through Unicode registration, all required information and relevant data will be posted to the working website of the project and submitted to Unicode for registration.

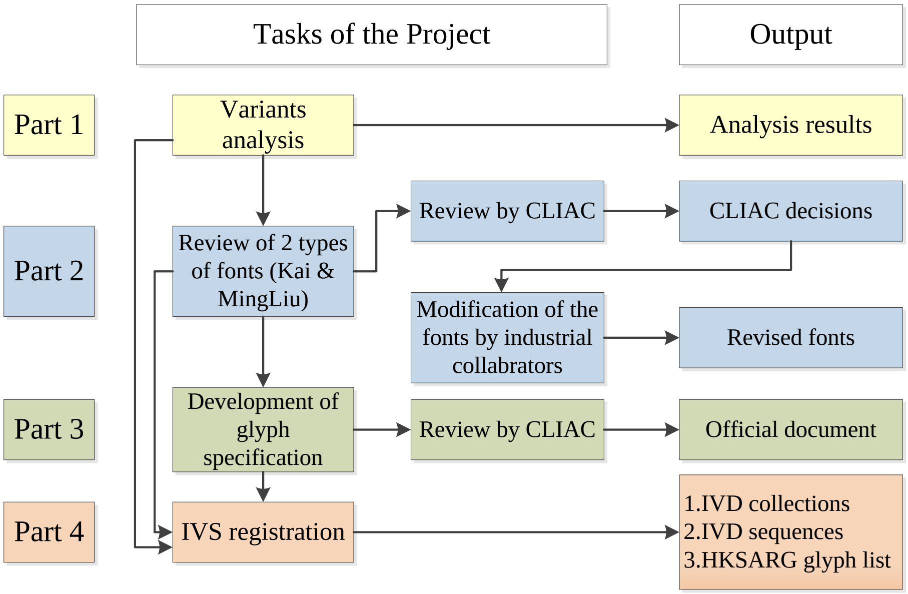


Figure 1 Workflow of this project

4. Principles for Specifying Hong Kong Specific Chinese Variants

Under the unification principles of the ISO10646/Unicode standard [2-3], we review all the characters in the Big5 standard of Taiwan [8] so as to identify their differences to the Hong Kong specific Chinese variants at the character level. In our review, we distinguish variants using a set of principles. The most important principle is to consider four types of differences: (1) stroke types, (2) stroke count, (3) relative positions of strokes/components, and (4) any other difference that may cause confusion of components or word meaning. For Hong Kong glyphs that are different from Big5 characters in any of the 4 types of differences, they will be listed as the Hong Kong specific Chinese variants. Examples of these differences are listed below.

4.1 Difference in stroke types

Chinese characters are formed by strokes. According to [13], the strokes of Kai style Chinese characters can be classified into dozens of types based on their writing orientation and shape. The basic types include "一" (*Héng, horizontal*), "丨" (*Shù, vertical*), "丿" (*Piě, left slanting*), "丶" (*Diǎn, dot*), and "㇇" (*Zhé, turn*).

For the two characters listed in Table 2, since the glyphs used in Hong Kong are different from those in the Big5 standard in stroke types, they will be considered as Hong Kong specific Chinese variants.

Table 2 Examples of difference in stroke types

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
83AB	B2F6	莫	莫	The last stroke is different.
4EA2	A4AE	亢	亢	The first stroke is different.

4.2 Difference in stroke count

If the stroke count of a Hong Kong glyph is different from that in the Big5 standard, it will be treated as a Hong Kong specific Chinese variant, like the characters listed in Table 3.

Table 3 Examples of difference in stroke count

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
6334	D1C0	挴	挴	The stroke counts of "母" and "毋" are different.
6C0F	A4F3	氏	氏	The stroke counts of "㇙" and "㇄" are different.

4.3 Difference in relative positions of strokes

As Table 4 illustrates, for the two glyphs, the relative positions of the stroke "㇀" are different. Therefore, this glyph will be treated as a Chinese variant in Hong Kong.

Table 4 Examples of difference in relative positions of strokes

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
5317	A55F	北	北	The relative positions of "㇀" are different.
5BFA	A678	寺	寺	The relative positions of the horizontal strokes of "土" are different.

4.4 Any other difference that may cause confusion of components or word meaning

4.4.1 Difference in protrusion

Whether stokes protrude at the stroke initiation/termination or at the folded corner of strokes can make the structural shapes look quite different. As Table 5 illustrates, this kind of glyphs are considered variants and will be listed as Hong Kong specific Chinese variants.

Table 5 Examples of difference in protrusion

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
5468	A950	周	周	The vertical stroke "丨" of the Hong Kong glyph protrudes downwards.
6025	ABE6	急	急	The horizontal stroke "一" in the middle of the Hong Kong glyph protrudes rightwards

4.4.2 Difference in rotation of strokes

Since such differences make the shapes of the characters quite different, they are considered as variants. Table 6 gives two examples.

Table 6 Examples of difference in rotation of strokes

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
68B2	D5BF	梲	梲	The strokes of "丷" are rotated.
706B	A4F5	火	火	The first stroke " 丶 " is rotated.

4.4.3 Difference in contact of strokes that cause different word meanings

The examples listed in Table 7 indicate that whether the stroke "一" touches the left or right stroke will cause confusion of different components and affect the word meaning of the characters. In consideration of this, we specify two different glyphs and will list all the characters containing the two components as Chinese variants.

Table 7 Examples of difference in word meaning

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
66F0	A4EA	曰	曰	The "一" of the Hong Kong glyph does not touch the right "丨".
5192	AB5F	冒	冒	The "一" of the upper component of the Hong Kong glyph touches neither the left nor the right "丨".

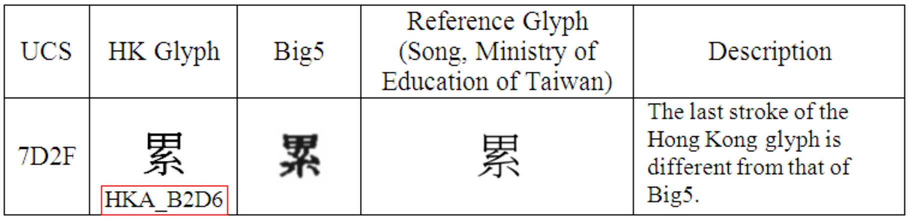
Other differences which we consider as font design differences that may vary from vendor to vendor are classified as trivial differences and will not be treated as glyph variants. For example, the two components "阝" and "阝" are considered as trivial differences. In the first case, the first stroke and the second stroke touch only at the starting point. Whereas, in the second case, the first stroke and the second one touch each other at all points. The components "艹" (the vertical strokes slightly slant inward) and "艹" (completely vertical) are also considered as trivial differences. Table 8 shows two examples using these components.

Table 8 Examples of trivial differences

Unicode	Big5	HK Glyph	Big5 Glyph	Differences
90A3	A8BA	那	那	The first and second strokes of "阝" of the Hong Kong glyph touch each other only at the starting point.
827E	A6E3	艾	艾	The vertical strokes of "艹" of the Hong Kong glyph slightly slant inward.

5. Encoding Scheme and Registration

Once the Chinese character variants are identified, the IVS technology is used to encode them. The name of the new glyph variant collection will be labeled as HKA, in which the letter "A" refers to the whole class of Big5 characters used in Hong Kong. The encoding format is "HKA_big5-code". For example, the code of "累" is "HKA_B2D6", as shown in Figure 2. Once completed, all the required information will be submitted to Unicode for registration and Unicode IVD registration required documentation and character list will be posted at the working website of the project (http://www.iso10646hk.net/ivd/1/) and later transferred to the Hong Kong Government website for public access.



UCS	HK Glyph	Big5	Reference Glyph (Song, Ministry of Education of Taiwan)	Description
7D2F	累 HKA_B2D6	累	累	The last stroke of the Hong Kong glyph is different from that of Big5.

Figure 2 Example of an encoded Hong Kong specific Chinese variant for registration

6. Conclusion

This paper introduces an ongoing project that applies the IVS and IVD for Hong Kong specific Chinese variant registration. The main objective is to specify Chinese character variants at the character level and encode all Chinese character variants under the Big5 coding framework used in Hong Kong so that computer systems can support these variants in different applications. This paper presents our project motivation, the workflow, the principles for identifying Hong Kong specific Chinese variants with examples, and also describes our encoding and registration process. The review of the variant glyph collection is currently being carried out by the CLIAC. The number of the characters in the list is around 7,000. The target completion time of the project is May 2015. With the Unicode review cycle of 3 months, we expect the Chinese character variant collection for Hong Kong should be included in Unicode before the end of 2015.

References:

[1] Qin Lu, Ideograph Variants-What They Are and How To Handle Them, the 21st International Unicode Conference, Dublin, Ireland, 14-17 May 2002

[2] The Unicode Standard Version 6.2 – Core Specification, Chapter 12 East Asian Scripts, Principles of Han Unification, pp.414-418, The Unicode Consortium, 2012

[3] Annex S (Informative) Procedure for the Unification and Arrangement of CJK Ideographs, ISO/IEC 10646-1:2000(E)

[4] Hong Kong Supplementary Character Set, HKSARG, September 28, 1999

[5] Information Technology Strategy, HKSARG, February 2002, http://www.digital21.gov.hk/eng/index.htm

[6] Reference Guide on Kai Style Character Glyphs for Chinese Computer Systems in Hong Kong, HKSARG, 2002

[7] Reference Guide on Song Style (Print Style) Character Glyphs for Chinese Computer Systems in Hong Kong, HKSARG, 2002

[8] 《電腦用中文字型與字碼對照表》（台灣工業標準「大五碼」之正式文獻）， 技術通報 C-26，財團法人資訊工業策進會，1984 年 (Computer Chinese Glyph and Character Code Mapping Table, the Industrial Standard of Big5 in Taiwan, Technical Report C-26, Institute for Information Industry of Taiwan, 1984)

[9] 李學銘（主編）, 《常用字字形表》, (二零零零年修訂本), 香港教育學院出版, 2000 年 (LEE Hok-ming as Chief Editor, Common Character Glyph Table, Hong Kong Institute of Education, 2000)

[10] 《國字標準字體宋體母稿-教育部字序》， 臺灣教育部，1998 年 (Master Copy of Standard Song Typeface for Chinese Characters, Ministry of Education of Taiwan,1998)

[11] 《康熙字典》，中華書局影印本，1997 年 (KangXi Dictionary, Chung Hwa Book Co. Ltd., 1997)

[12] 《漢語大字典》第一至八卷，湖北辭書出版社，四川辭書出版社，1986 年 (the Great Compendium of Chinese Characters, Vols 1-8, Hubei Dictionaries Press and Sichuan Dictionaries Press, 1986)

[13] 《信息處理用 GB13000.1 字符集 漢字部件規範》，國家語言文字工作委員會，1997 年 (Chinese Character Component Standard of GB13000.1 Character Set for Information Processing, The State Language Commission CLC of China, 1997)