SUPPORTING CHINESE CHARACTER VARIANTS IN HONG KONG THROUGH IDEOGRAPHIC VARIATION SEQUENCE

Qin Lu¹, Kwan Hin Cheung², Dan Xiong¹, Shing Yu¹, Jian Xu¹

¹Department of Computing, ²Department of CBS
The Hong Kong Polytechnic University
OUTLINE

- Introduction
- Background of the Project
- Project Scope and Workflow
- Principles for Specifying Hong Kong Specific Chinese Variants
- Encoding Scheme and Registration
- Progress of the Project
- Conclusion
INTRODUCTION

- A Chinese character may take different forms due to local preferences: U+9AA8

<table>
<thead>
<tr>
<th></th>
<th>Hong Kong Glyph</th>
<th>Mainland China Glyph</th>
<th>Taiwan Glyph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kai</td>
<td>骨</td>
<td>骨</td>
<td>骨</td>
</tr>
<tr>
<td>Song</td>
<td>骨</td>
<td>骨</td>
<td>骨</td>
</tr>
</tbody>
</table>

- Advantages and Limitations of CJK Unification
  - Exchanged text will not change meaning when exchanged
  - Searching and indexing will be easier
  - Different glyphs are supported under different locales
  - Difficult to display different glyphs in the same locale
**STATUS QUO IN HONG KONG**

- Hong Kong never had its own independently developed character standard
  - Use of Big5 as the Defacto standard
  - Extended with HKSCS

- HKSARG’s endeavour to meet it needs :
  - Published HKSCS in 1999 as extension to Big5
  - Promoted the use of ISO/IEC 10646/Unicode platforms
  - Developed the reference guides on component basis for Chinese computer systems in Hong Kong for both the Song and Kai styles in 2002

- Component based as there is no other appropriate
Background: More on Variants and Unification

Definition of Chinese character variants:
- The set of different glyph shapes of the same character
- Do not change the meaning of a character
- Example: (雞 雞 鶴 鴨) (谿 溪) (骨 骨)

Need for ISO/IEC 10646 unification procedure:
- Structure analysis, semantics, and glyph variation at different levels
- Separately coded: (雞 雞 鶴 鴨) (谿 溪)
- Unified: (骨 骨)
NEW ENABLING TECHNOLOGY

- The Ideographic Variation Sequence (IVS)
  - Define a variant with reference to a defined character
    - Ideograph followed by a variation selector

- Variation selectors coded in ISO 10646: U+E0100 to U+E01EF
  - Total of 240
  - Non-printable characters and cannot be used alone

- IVS: <CJK ideograph, VS>
  - U+8FBB: 这
  - <U+8FBB, U+E0100>: 这
HOW IS IVS USED AND STANDARDIZED

- **How to use IVS:**
  - The variant selector will be ignored for display if there is no alternative font
  - They can be searched and sorted in the same way as that of the base character

- **How to standardize**
  - Registration in IVD of Unicode
    - Collection name
    - Permanent site from submitter for public information
    - 3 months public review

- **Advantages**
  - Suitable for unified characters that cannot be coded otherwise
  - Increased sharing and interoperability cross different platforms than using PUA

- **Quite suitable for Hong Kong’s Chinese variants**
  - Officially define our variant collection against Big5
**PROJECT OBJECTIVE AND SCOPE**

**Objectives**
- To specify Hong Kong specific Chinese character variants at the character level
- Under the IVD framework to 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

**Project Scope:**
- The registration of IVS contains characters in the Big5 repertoire
  - HKSCS characters already adhere to the reference guide of the Hong Kong character glyphs and are not unifiable to any character in Big5
WORKFLOW: DIAGRAM

Tasks of the Project

Part 1
- Variants analysis
  - Review of 2 types of fonts (Kai & MingLiu)
    - Development of glyph specification
      - IVS registration
  - Review by CLIAC

Part 2
- Review of 2 types of fonts (Kai & MingLiu)
  - Modification of the fonts by industrial collaborators
    - Revised fonts

Part 3
- Development of glyph specification
  - Review by CLIAC

Part 4
- IVS registration
  - Review by CLIAC

Output

- Analysis results
- CLIAC decisions
- Revised fonts
- Official document
- 1.IVD collections
- 2.IVD sequences
- 3.HKSARG glyph list
PROJECT COMPONENTS

- **Part1**: Review and analyze Hong Kong character glyphs with all the characters in the Big5 standard
  - Use different references (by the project team)
- **Part2**: Produce two types of Chinese character fonts Kai and Song, based on the analysis results (by the project team)
  - Produce the revised fonts after the variants are identified (by the industrial collaborators)
  - Review and finalize the deliverables (by CLIAC)
- **Part3**: Prepare the glyph specification and exhaustive list (by the project team)
  - Review and approve the documents (by CLIAC)
- **Part4**: Submit the list of the variant characters to the Unicode standard for inclusion in its IVD through Unicode registration
  - Post all required information to the working website of the project (http://www.iso10646hk.net/ivd/1/)
PRINCIPLES TO IDENTIFY VARIANTS

The most important principle is to consider four types of differences:
1) Stroke types
2) Stroke count
3) Relative positions of strokes/components
4) Any other difference that may cause confusion of components or word meaning
DIFFERENCE IN STROKE TYPES

- Basic stroke types: 一 (Héng, horizontal), | (Shù, vertical), \ (Piě, left slanting), \ (Diǎn, dot), ↘ (Zhé, turn)

- Difference in stroke types:

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5 Glyph</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>83AB</td>
<td>B2F6</td>
<td>莫</td>
<td>莫</td>
<td>The last stroke is different.</td>
</tr>
<tr>
<td>4EA2</td>
<td>A4AE</td>
<td>亢</td>
<td>亢</td>
<td>The first stroke is different.</td>
</tr>
<tr>
<td>5316</td>
<td>A4C6</td>
<td>化</td>
<td>化</td>
<td>The 3rd stroke is different.</td>
</tr>
</tbody>
</table>
# Difference in Stroke Count

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>6334</td>
<td>D1C0</td>
<td>梅</td>
<td>梅</td>
<td>The stroke counts of “梅” and “梅” are different.</td>
</tr>
<tr>
<td>6C0F</td>
<td>A4F3</td>
<td>氏</td>
<td>氏</td>
<td>The stroke counts of “氏” and “氏” are different.</td>
</tr>
<tr>
<td>82E0</td>
<td>D0A2</td>
<td>蒡</td>
<td>蒡</td>
<td>The stroke counts of “萆” and “萆” are different.</td>
</tr>
<tr>
<td>7B64</td>
<td>DF47</td>
<td>籬</td>
<td>籬</td>
<td>The stroke counts of “籬” and “籬” are different.</td>
</tr>
</tbody>
</table>
## Difference in Relative Positions of Strokes/Components

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5 Glyph</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>5317</td>
<td>A55F</td>
<td>北</td>
<td>北</td>
<td>The relative positions of “北” are different.</td>
</tr>
<tr>
<td>5E78</td>
<td>A9AF</td>
<td>幸</td>
<td>幸</td>
<td>The relative positions of the horizontal strokes of “幸” are different.</td>
</tr>
<tr>
<td>5BFA</td>
<td>A678</td>
<td>寺</td>
<td>寺</td>
<td>The relative positions of the horizontal strokes of “寺” are different.</td>
</tr>
</tbody>
</table>
Any Other Difference

Any other difference that may cause confusion of components or word meaning: considered as Hong Kong specific Chinese variant

1) Difference in protrusion

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>5468</td>
<td>A950</td>
<td>周</td>
<td>周</td>
<td>The vertical stroke “</td>
</tr>
<tr>
<td>6025</td>
<td>ABE6</td>
<td>急</td>
<td>急</td>
<td>The horizontal stroke “一” in the middle of the Hong Kong glyph protrudes rightwards</td>
</tr>
</tbody>
</table>
ANY OTHER DIFFERENCE

2) Difference in rotation of strokes: the differences make the shapes of the characters quite different

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>68B2</td>
<td>D5BF</td>
<td>稅</td>
<td>稅</td>
<td>The strokes of “\” are rotated.</td>
</tr>
<tr>
<td>67B0</td>
<td>AC69</td>
<td>杼</td>
<td>杼</td>
<td>The strokes of “\” are rotated.</td>
</tr>
<tr>
<td>706B</td>
<td>A4F5</td>
<td>火</td>
<td>火</td>
<td>The first stroke “\” is rotated.</td>
</tr>
</tbody>
</table>
3) Difference in contact of strokes that cause different character semantics

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>66F0</td>
<td>A4EA</td>
<td>日</td>
<td>日</td>
<td>The “一” of the Hong Kong glyph does not touch the right “一”.</td>
</tr>
<tr>
<td>5192</td>
<td>AB5F</td>
<td>冒</td>
<td>冒</td>
<td>The “一” of the upper component of the Hong Kong glyph touches neither the left nor the right “一”.</td>
</tr>
</tbody>
</table>
**TRIVIAL DIFFERENCES**

Font design differences that may vary from vendor to vendor: classified as trivial differences and will not be considered as variants, eg, “β” and “β”, “++” and “++”

<table>
<thead>
<tr>
<th>U+</th>
<th>Big5</th>
<th>HK Glyph</th>
<th>Big5 Glyph</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>90A3</td>
<td>A8BA</td>
<td>那</td>
<td>那</td>
<td>The first and second strokes of “β” of the Hong Kong glyph touch each other only at the starting point.</td>
</tr>
<tr>
<td>827E</td>
<td>A6E3</td>
<td>艾</td>
<td>艾</td>
<td>The vertical strokes of “++” of the Hong Kong glyph slightly slant inward.</td>
</tr>
</tbody>
</table>
SCOPE OF THE VARIANT LIST

- The variation list includes all Chinese character glyphs used in Hong Kong which are different from Big5.

- The number of the characters in the list is around 9,000.
OTHER CONSIDRATIONS

- When Big5 is inconsistent:
  1. When most HK characters containing the component are different from Big5: all related characters are listed in order to specify Hong Kong character glyphs in a consistent way, eg, for 言, Big5 has three glyphs: 謁 計 訤 (the 3rd one is the same as Hong Kong glyph) In this case, all characters containing 言 are included in the variant list.
  2. Only a very few characters containing the component are different: eg, 果: In all Big5 characters containing this component, only two have a hook: 墭 𨦵 In this case, they are not included in the variant list but listed in another document for reference.
ENCODING AND REGISTRATION

- **Encoding format:**
  - HKA_big5-code, eg, 累：HKA_B2D6
  - “A” refers to Big5 characters used in Hong Kong

<table>
<thead>
<tr>
<th>UCS</th>
<th>HK Glyph</th>
<th>Big5</th>
<th>Reference Glyph (Song, Ministry of Education of Taiwan)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7D2F</td>
<td>累</td>
<td>累</td>
<td>累</td>
<td>The last stroke of the Hong Kong glyph is different from that of Big5.</td>
</tr>
</tbody>
</table>

- **Registration**
  - All required information will be submitted to Unicode for registration
PROGRESS OF THE PROJECT

- The variant glyph collection is currently being carried out by the CLIAC.
- The target completion time of the project is **May 2015**.
- It is expected that the Chinese character variant collection for Hong Kong should be included in Unicode before **the end of 2015**.
CONCLUSION

An ongoing project that applies the IVS and IVD for Hong Kong specific Chinese variant registration:

- To specify Hong Kong specific Chinese character variants at the character level
- To encode all variants under the Big5 coding framework used in Hong Kong so that computer systems can support these variants in different applications