JTC 1/SC2/WG2/IRG N 1090

Date: Nov. 29, 2004

ISO/IEC JTC 1/SC2/WG2/IRG

Ideographic Rapporteur Group

Source: Yasuhiro, ANAN

Meeting: IRG#23 @ Jeju

Title: Thoughts on CJK Basic Strokes

Keywords: (none)

Status: Individual Contribution

Short Description:

This paper presents bits and bytes of ideas and concerns to define CJK Basic Strokes repertoire.

Proposed Conclusion / Requested Action: IRG to discuss

Thoughts on CJK Basic Strokes

0. Preface

At the last WG2 meeting 45 in Markham, 16 stroke characters from HKSCS were accepted to be encoded to a new block called CJK BASIC STROKES and the scope of IRG tasks has been expanded to include CJK strokes with SC2 approval. The intention of the expansion of the scope would not be at all to catalogue every single instance of ideographic strokes to ISO/IEC 10646, but to consult the IRG with the development of the appropriate encoding model for CJK strokes and the basic repertoire based on the model.

It is expected that the following items are fully discussed within the IRG:

1. The purpose and scope of CJK Strokes,
2. Encoding model and/or taxonomy of strokes,
3. Issues identifying a stroke – difference of typeface, regional conventions
4. Basic repertoire – criteria for basic strokes.

1. The purpose and scope of CJK Strokes

No one in the IRG would argue about the importance of ideographic stroke concept. It's the smallest unit of an ideograph and it gives detail features to ideographs as well as numeric characteristics (stroke count). However, encoding ideographic strokes to UCS requires its own justification because they are generally not ideographs in their own right.

In the document IRG N927 (Two more ideographic strokes for CJK_C1), China argued about the importance of encoding two ideographic strokes in terms of "ideograph decomposition, analysis and for making ideographic strokes subset", but it does not answer to the question why they have to be encoded to UCS. In fact, HKSAR provides the list of all basic components including strokes, to each of which the code and the name based on other scheme than UCS is given for the similar purpose and made it available on their Web site (http://glyph.iso10646hk.net/english/hkcharacters_2.html). In order to answer to the question why ideographic strokes in UCS, the purpose and scope have to be specified in terms of how UCS applications are involved.

2. Encoding model and/or taxonomy of strokes

CJK C1 project adopts 5 basic stroke types (横, 竖, 撇, 点, 折) as part of ideograph categorization scheme. This is one well-established encoding model of ideographic strokes and we can call them as CJK Basic Strokes. Yet another well-known encoding model of ideographic strokes is 8 basic strokes (侧, 勒, 努, 钩, 策, 掠, 啄, 磔) used to compose 永. One implementation example of this encoding model is a commercial Input Method which takes the eight stroke types as input sequence from a user and converts each stroke combination to ideographs.

The IRG N927 indicates that there're already good enough number of ideographic strokes as CJK Ideographs except two strokes in the proposal. It is unknown which strokes are already encoded and which encoding model sits behind the proposal, but

China is expected to clarify them further. The other proposal, IRG N987 (CDL strokes) provides 39 basic stroke types. Stroke categorization in this document seems well explained and comprehensive, but the question remains what makes the 39 (or 64) set as basic without evaluation of the actual CDL data.

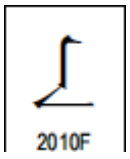
3. Issues identifying a stroke – difference of typeface, regional traditions

Because ideographic strokes are the most primitive elements of ideographs and the unification of ideographs is generally evaluated on the higher nodes of component analysis (S.1.3.1), ideographic strokes wouldn't follow the same unification rule under which CJK Unified Ideographs are classified. The first 4 strokes of 言 (U+8A00) in Ming style are 横 strokes in 5 basic strokes model (or 勒 strokes in 8 basic strokes model). In Song style, the glyph for U+8A00 is represented by 言. The 1st stroke is not 横 but 点, but obviously the 2 types of strokes in 5 basic stroke model are not unified however the 2 glyphs 言 in Ming style and 言 in Song style are unified because it is considered that the different appearance of the 1st stroke does not change the abstract shape of the ideograph. In other words, ideograph decompositions to strokes would not be unique depending on typefaces and/or regional traditions.

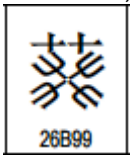
According to CDL stroke model, 撇 (丿丿丿) has 3 different subtypes -撇 (p), 弯撇 (wp) and 竖撇 (sp). Because the difference between those subtypes is very subtle, even when typeface is given, sometimes it is difficult to identify the subtype of 撇 (丿). If they are to be distinguished, people might want to ask why 横 (h) stroke unifies subtypes of 勒 (horizontal) and 策 (slanted to the upper-right) as in 8 basic strokes model.

4. Basic Repertoire – criteria for basic strokes

ISO/IEC 10646 encodes more than 70K Unified Ideographs and it's not a surprise there're handful of ideographs with unusual and specialized strokes as shown in the following examples. It would be difficult to call those strokes as basic, so the IRG is expected to define criteria to judge which stroke is basic and which is extended or specialized. One way would be to specify the repertoire, say URO. Any ideograph in URO has a decomposition only with CJK Basic Strokes.

U+2010F  and U+20114  (reversed upside-down),

U+20137  , U+21100  , U+211A2  , U+22013  ,

U+26B99 

----end----