

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2/IRG N1371

Date: 2007-11-13

Source:	John Knightley
Title:	Further Annex S issues
Status :	For discussion
Distribution:	IRG members and Ideographic Experts
Medium :	Electronic

Below are list several Annex S related issues apparent in documents submitted to the IRG.

1) Dealing with New components

One of the challenges with regard to applying the principles of Annex S is when new, or seldom seen, components of an usual nature. A new component is a character, or part of a character that can not be decided on solely by considering breaking it down into parts and applying known examples. In this sense the word component is used to mean can not be meaningfully divided a component itself may in fact have several parts.

Annex S is a convention whose strength comes from being kept, for completely new components a convention needs to be established, in the cases shown below it is suggested that there is a choice as to whether or not to consider them to have the same abstract shape.

There are some glyphs with several parts that are can considered as base components rather than subdivided, two such examples are 𠂇 vs 𠂈 and 𠂉 vs 𠂊 (IRG N1367 glyphs 115 and 116). Whilst it is correct to say that if 𠂉 and 𠂊 are considered to have different abstract shapes, then according to Annex S '—' and '...' have different abstract shape, it does not follow that all pairs that differ by '—' vs '...' must have different abstract shapes. Since have never been included in UCS before whether a convention is agreed that they have the same abstract shape or that they have a different abstract shape it could be argued the basic integrity of Annex S has been maintained.

Once a decision has been agreed whether or not the components have the same abstract shape processing them is straight forward, as is seen by the 1082 and 144 encoded characters containing the components 𠂉 and 𠂊 respectively.

Following the Rules

Annex S contains both unification and disunification examples

S.1.4.3 Different structure of a corresponding component

The examples below illustrate rule c). The structure of one (or more) corresponding components within the two ideographs in each pair is different.

扌·擴, 策·筭, 𠂇·然, 𠂇·徑, 𠂇·僉, 区·區, 夾·夾, 单·單, 隹·隹, 𠂇·藁, 贊·贊, 襄·襄, 隹·隹, 問·問, 朶·朶, 隹·隹, 恒·恆, 奘·奘, 人·入, 𠂇·宗, 𠂇·又

Shows that if the right hand side of characters differ by 人 vs 入 then they should not be unified however it is suggested 𠂇 be unified with U+201A7 𠂇.

𠂇 00218 G_CYY00075	𠂇 201A7				G: unified H: unifiable with 201A7 provided that they have the same meaning and pronunciation	J: unifiable/duplicate. If any objection show more evidence. (*5)
--------------------------	------------	--	--	--	--	---

(IRGN1353_CJK_D_V22-D_Unification.pdf page 10)

Stating which the unification example a decision is based upon would a way to eliminate such errors.

The next two cases however are not so clear cut, and may benefit from reconsideration:-



H: Unifiable with 34B1 provided that they have the same meaning
 J: unifiable/duplicate. If any objection show more evidence. (*5)

Page 67

(IRGN1353_CJK_D_V22-D_Unification.pdf page 67)

 05461 V04-4F21	 250F1				H: Not unified with 250F1 V: Unifiable	J: unifiable/duplicate. If any objection show more evidence. (*5)
--	--	--	--	--	---	---

(IRGN1353_CJK_D_V22-D_Unification.pdf page 70)

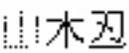
Discussing IRG N1350
Additional Examples of Annex S

Given any pair of CJK characters either have the same or different abstract shape. However at this stage in the review process the examples should be divided in three groups namely:-

- (1) examples definitely having the same abstract shape
- (2) examples definitely having differences of abstract shape
- (3) examples requiring further evidence or discussion

The list of examples is by no means complete. Suggests for additional examples are invited. Some of these may come for IRG discussions.

For instance discussion of

52	JH-JTB2FC		
----	-----------	---	---

(IRGN1367_J-urgent.pdf page 5)

suggests U+4E11 𠃉 vs U+4E12 𠃉 be added to the list of examples.

Some of the examples in IRGN1350_Appendix_A_unifiable_Part01.pdf and IRGN1350_Appendix_A_unifiable_Part02.pdf require discussion. Any examples which only have two or three strokes need to be considered very carefully, as do any that seem to have overlap with disunification examples. It is suggested that the following examples might best be put in the group requiring further evidence or discussion:-

001, 008, 010, 012, 032, 033, 039, 041, 091, 098.

Most of the 151 disunification examples in IRGN1350_AppendixB_non-unifiable.pdf are clear cut, except examples 137 and 138.

Additional Sections

At this stage in the review process most of these do not need to be discussed in detail. However comments made on Extension D unification cases illustrate a difference of opinion regarding (C06) this is either true or false.

敝 03482 G_zjw01074	魁 29C81				G: not unified,different structure&component H: Not unified with 29C81	J: not unifiable because of different structure of a corresponding component (see S.1.4.3). (*3)
--------------------------	------------	--	--	--	---	--

(IRGN1353_CJK_D_V22-D_Unification.pdf page 10)