SAT Feedback to "Preliminary proposal to add a new provisional kIDS property (Unihan)" (IRGN2492) and "Proposal to encode five new Ideographic Description Characters" (IRGN2572)

Date: 2022-08-29

1. Usage of IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBSTRACTION

We generally acknowledge the usefulness of the newly proposed binary operator IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBTRACTION (hereinafter SS) for human users, but at the same time, we are concerned, as Dr. Qin LU in a comment about the document at IRG #57, that it would introduce to the IDS system a new dimension of ambiguity which is hostile to machine checking algorithms (such as search and generating canonical forms/decompositions). We hence suggest that **reasonable constraints should be imposed on its usage in the IDS data of new characters to be submitted to future IRG working sets**.

In the system using traditional IDCs (including SURROUND FROM RIGHT and SURROUND FROM LOWER RIGHT in this discussion), a character is described with a combination of one or more two-dimensionally separable components in the way designated by each IDC. Provided every existing CJK ideograph is associated with an IDS, we can recursively decompose a character until it ultimately reduces to a combination of a limited number of graphemes that are, for most practical purposes, atomic. Although in reality an ideograph does not always resolve to only one canonical sequence, there are only a finite number of paths that can in principle be collated[1]. This is the basic principle that (we assume) most IDS machine checkers presuppose, as well as the rationale why we have been freely choosing a short, intuitive IDS from among a vast number of options to describe an ideograph. (That is, the precise choice of IDS ought not matter because IDSes are assumed to be confluent in a system of rewrite rules.) As regards the subsequent discussion, however, we note a limitation of the traditional IDS system: namely that it does not have a notation or other mechanism to indicate a relationship between two components that are not connected in the IDS forest. One purpose of UCVs is to bridge this gap, by instructing machines on the perceived equivalence of glyphs that are not technically linked in the network, as well as kStrange, to help human users look up hard-to-reach orphaned (often atomic) components for better machine checking coverage.

SS, by its nature, can be regarded as an inverse operator for any suitable traditional IDC (ignoring the difference between CJKUI and CJK Strokes for the purpose of this discussion):

氏 = ⊖氏丶 ← 氏 = ⊟氏丶

[1] Of course, a number of reservations should be made, such as possible incompatible/incommensurable decompositions allowed by the operator OVERLAID.

自 = ⊖百一 ← 百 = ⿱一自

其 = ⊖其ノ ← 其 = ⿴其ノ

With the support of additive (compositional) sequences defined elsewhere, it can theoretically be incorporated into the interdependent network of IDSes. However, as stated in the proposal, the motivation for introducing the new operator is to describe an ideograph otherwise difficult to compose with additive operations, such as the third example (from the original document IRGN2492), U+2CEBB (豕 without the last two strokes). Suppose that we do not have a character 大 in the Han repertoire, but only 太 and 犬. Now two submitters try to encode this new character, one using ⊖太丶 and the other using ⊖犬丶. These two IDSes are not conflatable by an algorithm if 太 and 犬 are atomic (not decomposable), or share no already describable component (which would link them using "traditional IDC" semantics). Thus, if we assume the current IDS data, it is impossible to detect whether two different SS sequences represent an effectively identical shape, or even whether such a sequence is identical to an existing character. This potential for false negatives makes SS qualitatively more dangerous than OVERLAID, which is often termed “ambiguous” and for which a fuzzy search is likely to return false positives if a decomposition strategy is sufficiently reasonable. (False positives are relatively harmless; false negatives are not.)

We therefore believe that, although using SS has a clear advantage for human recognition, the descriptiveness of a subsequence led by it is basically equivalent to ？ in automatic duplication checking. We suggest that some safety measures should be taken for the use of SS in IRG WS submissions, such as (but not restricted to) one or more of the following:

- The submitter must also provide an additive IDS of a character for which SS has been used.
- The submitter must also declare the shape which the SS sequence represents in their supplementary components list (or elsewhere).
- The submitter (or another authority in the pipeline) must confirm that the intended component described with SS has not been encoded, as a part of quality assurance.
- Restrict the choice of subtrahend to a small set of minor strokes to avoid arbitrary variation for a shape associable with multiple characters.[2]
- Do not use sequence as the subtrahend component of SS sequence, or find a mechanism to reduce ambiguity if multiple strokes are subtracted.[3]
- Do not use SS inside an SS sequence.

In addition, shapes previously described with SS should be recorded in the proposed components block or some IRG documents, so that they can be found. Here, the `kStrange` property comes in handy.

---

[2] E.g.: 宀 = ⊖家豕 = ⊖安女 = ⊖完元 = ⊖客各 = ⊖容谷 = ⊖守寸

[3] In the original document's example of U+2CEBB (豕 without the last two strokes) as ⊖豕⿱ノ乀, people might also consider the IDSes ⊖⊖豕ノ乀 and ⊖⊖豕乀ノ.

Finally, **one idea for handling subtractive sequences from an algorithmic perspective would be to rewrite them as additive ones** (possibly using OVERLAID as a position-agnostic addition operator) in a database-internal preprocessing step.

2. Usage of IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION and IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN ROTATION (IDEOGRAPHIC DESCRIPTION CHARACTER ONE HUNDRED EIGHTY DEGREE ROTATION)

The potential problems of unrestricted usage of those two IDCs in the IRG work have been already covered in L2/18-012 (by Taichi KAWABATA) and Kushim JIANG's Feedback to IRGN2273R, which can be summarized as:

- The possibility of multiple interpretations regarding which component is transformed in a complex component (e.g., 壯 = ↔□↷干片 (?))
- The ability to create idempotent and/or redundant notations, which can cause infinite loops in the algorithm (e.g., 字 = ↔↔字 = ↔↔↔↔字 = …)
- (We would like to add: the ambiguity that ↔↷ = ↷↔)

Thus, we suggest the following for the safe handling of those unary operators in IRG WS submissions:

- Prohibit sequences as the argument of these new unary operators; that is, allow only single characters and strokes.
- The algorithm should strip away all of the unary operators from IDSes before matching.
- The submitter (or another authority in the pipeline) must confirm that the component intended to be described by the unary operators is not encoded, as a part of quality assurance.

Also note that Kushim JIANG questions whether those unary operators should be named with *IDEOGRAPHIC DESCRIPTION CHARACTER* where their function is closer to that of 〾 U+303E IDEOGRAPHIC VARIATION INDICATOR.

3. Name and mapping of the components block

We agree with Eiso CHAN's Feedback to IRGN2492, that:

(a) The new ideograph components block seems better placed somewhere near the end of the SIP, where a considerable number of code points remain unassigned. The range U+2EBF0 through U+2F7FF has 3,088 code points, and the range U+2FA20 through U+2FFFD has 1,502 components available, which will be more than sufficient for prospective maximal number of components. We especially note that the range after the CJK Compatibility Ideographs Supplement is the least expected place to take in any further extension of CJKUI and might be good to accommodate a smaller block, unless other factors are considered.

(b) The block name (with the proposed name *CJK Unified Ideographs Components*) could be more explicit about being for technically plain CJKUIs, to minimize misunderstandings of potential

users. The best name will depend on the intended purpose of the block, but could be e.g., *CJK Unified Auxiliary Ideographs* or *CJK Unified Accessory Ideographs* (besides Eiso's suggestion).

4. Other

- As for the name IDEOGRAPHIC DESCRIPTION CHARACTER HALF-TURN ROTATION, we note that the phrase "half-turn rotation" is unusual. Rotations are rarely full-turn (360°) rotations. If numerals are permitted, explicitly calling it a 180°-rotation might be preferable. Otherwise, we would prefer IDEOGRAPHIC DESCRIPTION CHARACTER ONE HUNDRED EIGHTY DEGREE ROTATION (as in IRGN2492) or simply IDEOGRAPHIC DESCRIPTION CHARACTER HUNDRED EIGHTY DEGREE ROTATION (*i.e.,* without the "ONE").
- On the side, we note that in the IDS syntax, the term `IDS_Trinary_Operator` does not match standard usage in programming language semantics (e.g., C and C++ have a *ternary* operator with the following intended/abstract semantics:

  `λ(b:bool, x:τ, y:τ).(b ? x : y):τ)`

  We propose to consider using `IDS_Ternary_Operator` if possible.

Acknowledgments

We thank Stephan Hyeonjun STILLER for comprehensive suggestions and proofreading.

(End of document)