

Universal Multiple-Octet Coded Character Set

International Organization for Standardization

ISO/IEC JTC1/SC2/WG2/IRG

REVISED PROPOSAL TO DISUNIFY U+5F50

Individual Contribution

Alexander Zapryagaev

For consideration by the IRG

Date: 2021-08-08

rudetection@gmail.com

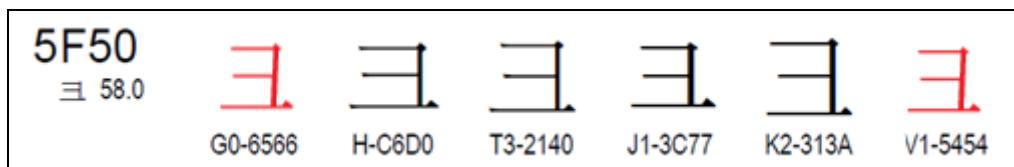
TABLE OF CONTENTS

| | |
|--------------------------------------|----|
| The Proposal..... | 2 |
| The Characters | 3 |
| Local Data..... | 5 |
| Can't They Stay Unified?..... | 9 |
| Possible Solution | 10 |
| Appendix: Various Attestations | 11 |

THE PROPOSAL

This document proposes a series of changes that could serve as a solution to the current situation when three highly graphically similar characters, mostly used as components of other characters were, I believe, mistakenly unified in Unicode, leading to unwanted consequences. The following changes are proposed.

1. Alter the glyphic representation of the character G0-6566 in the GB 2312-1980 standard so it conforms to the 𠄎 style; do the same with V1-5454.



2. Add a character corresponding to 𠄎 standard to U-Source. Change the status of UTC-01005 𠄎 from U (“Encoded in the URO or as a unified ideograph in the CJK Compatibility Ideographs block”) to N (“Planned to be submitted for a future extension”).
3. Encode, appending to the URO, two new characters with the representative glyphs following the 𠄎 and 𠄎 standards. Use the U-Source of UTC-01005 and the character encoded in Item 2, respectively.

The 𠄎 character may, beside UTC-01005, use at later moment such evidence as GZH-0657.02, GZ-45901, JMJ-011225.

The 𠄎 character may, beside UTC pending, use at later moment such evidence as JMJ-011226.

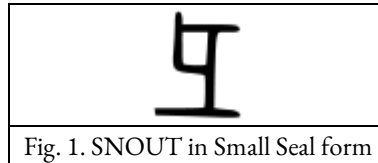
4. Attach the following comment to U+2E95 in the table (also, replace the approximately equal sign with an arrow and correct the glyph in the commentary to U+2F39):

| | | |
|--------------------------|--|-----------------------|
| 2F39 | 𠄎 | KANGXI RADICAL SNOOT |
| | → 5F50 | 𠄎 |
| 2E94 | 𠄎 | CJK RADICAL SNOOT ONE |
| | → 5F51 | 𠄎 |
| 2E95 | 𠄎 | CJK RADICAL SNOOT TWO |
| <input type="checkbox"/> | Despite the name, is actually a version of 2F1C KANGXI RADICAL AGAIN | |
| | → XXXXX | 𠄎 |

This document is the revision of [L2/20-112](#), which is in its course a response to the feedback ([Feedback 1](#), [Feedback 2](#)) to [IRGN 2414 Proposal to De-Unify One Obsolete Simplified Chinese Character](#), where a different solution was offered which, upon discussion, cannot be maintained anymore.

THE CHARACTERS

U+5F50 彡 (J-Source form shown) opens the segment of the URO dedicated to the Kāngxī radical 58 (U+2F39 彡 KANGXI RADICAL SNOOUT according to the naming system of the code chart). It is followed by U+5F51 彡, providing the alternate form of the radical-character as given by U+2E94 彡 CJK RADICAL SNOOUT ONE in the “CJK Radicals Supplement” block. Both represent the graphical variations of the character with the Early Middle Chinese (EMC) reading k-jeiH (Baxter), corresponding to modern Mandarin *jì* and defined by *Shuōwén* as 豕之頭 “head of a boar”, which is probably what the original form of the character depicts (Figure 1):



U+5F50 is thus a well-known and universally spread alternate graphic representation of the same character. In fact, in modern simplified and shinjitai forms the occurrence of 彡 is much more probable than that of the unchanged elements; see, for example, U+9332 録, which was simplified in Japan from U+9304 録, and its Chinese version 录 (in 录, which originally denoted “to trim/carve wood,” the whole grapheme probably depicts a decapitated tree, but the top was early reinterpreted as the “decapitated” head of a “pig”¹).

The characteristic lower line of 彡, frequently extended in length, is retained in 彡 as well, according to its usage. Currently, the word 彡 is not used to describe a pig’s snout anymore, with general words like 口鼻部 *kǒubíbù* taking the job, but there is at least one situation where the character is still relevant: that is, speaking about the radical itself (彡部 *jìbù* “Radical 58”). We will further refer to 彡 as SNOOUT.

However, this character has an extreme graphical similarity to another one, also popular as a component: I refer to 彡. As the graphical form shows, the difference lies in the manner of the contact between the right vertical and lower horizontal strokes (here the vertical slightly protrudes² and, more importantly, the horizontal stops at the vertical). This character is one, the most moderate, of the descendants of the glyphs for “right hand” (Figure 2), and we will apply the name HAND to it:



¹ <https://wenlin.co/wow/Zi:%E5%BD%95>.

² Or exactly touches the end of the horizontal, as might be rendered in several fonts and is, at least according to the representative glyphs, the normal Vietnamese rendering.

This character directly represents a right hand, now normatively 又, which is the descendant of the same glyph.

Later, the basic glyph suffered a phonetic loan, now standing for the identically read adverb: 又 *yòu* “again;” this required to modify the glyph meaning “right,” now 右 *yòu*, with a 口 “mouth” appended.

Still, many characters containing “right hand” as a component show an explicitly 𠃉-like form, such as (Figure 3):



Fig. 3. U+5F57, G-form

This character denotes a broom, with the top side being a drawing of the twigs of the broom, while the bottom is the hand carrying it.

Perhaps, the more recognizable form of this HAND is, however, one with the horizontal line in the middle protruding beyond the vertical, like in the Kāngxī Dictionary (Fig. 4):

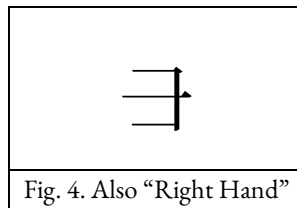


Fig. 4. Also “Right Hand”

We will denote it as LONG HAND and, once again, demonstrate its usage on the broom (Fig. 5):

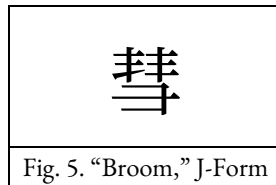


Fig. 5. “Broom,” J-Form

Thus, we consider three elements:

| | |
|-------------------------|---|
| SNOUT | 𠃉 |
| HAND | 𠃉 |
| LONG HAND | 𠃉 |
| Table 1. The characters | |

They are currently unified as U+5F50. This document proposes their disunification, with retention of SNOOUT for the current codepoint and encoding HAND and LONG HAND anew.

LOCAL DATA

PRC

The source of the complexity is probably the decision of the PRC to unify the graphical form of these components and choose 𠄎 as the joint representation of them. To quote the PRC feedback of 2019.10.23 to IRGN2414:

‘The only reason why 𠄎𠄎𠄎 is are always mix used mistakenly in ancient times, and Ministry of Culture and Reform and arrangement Committee of Chinese characters of China released a document called 《印刷通用汉字字形表》 in 1965 to solve this chaotic situation in general use. This document use 𠄎 instead of 𠄎𠄎𠄎<...>, so that the dictionaries and GB2312-1980 followed that decision.’

Thus, the 1980 establishment of GB 2312-1980 produced the following aberrant representative glyph G0-6566:

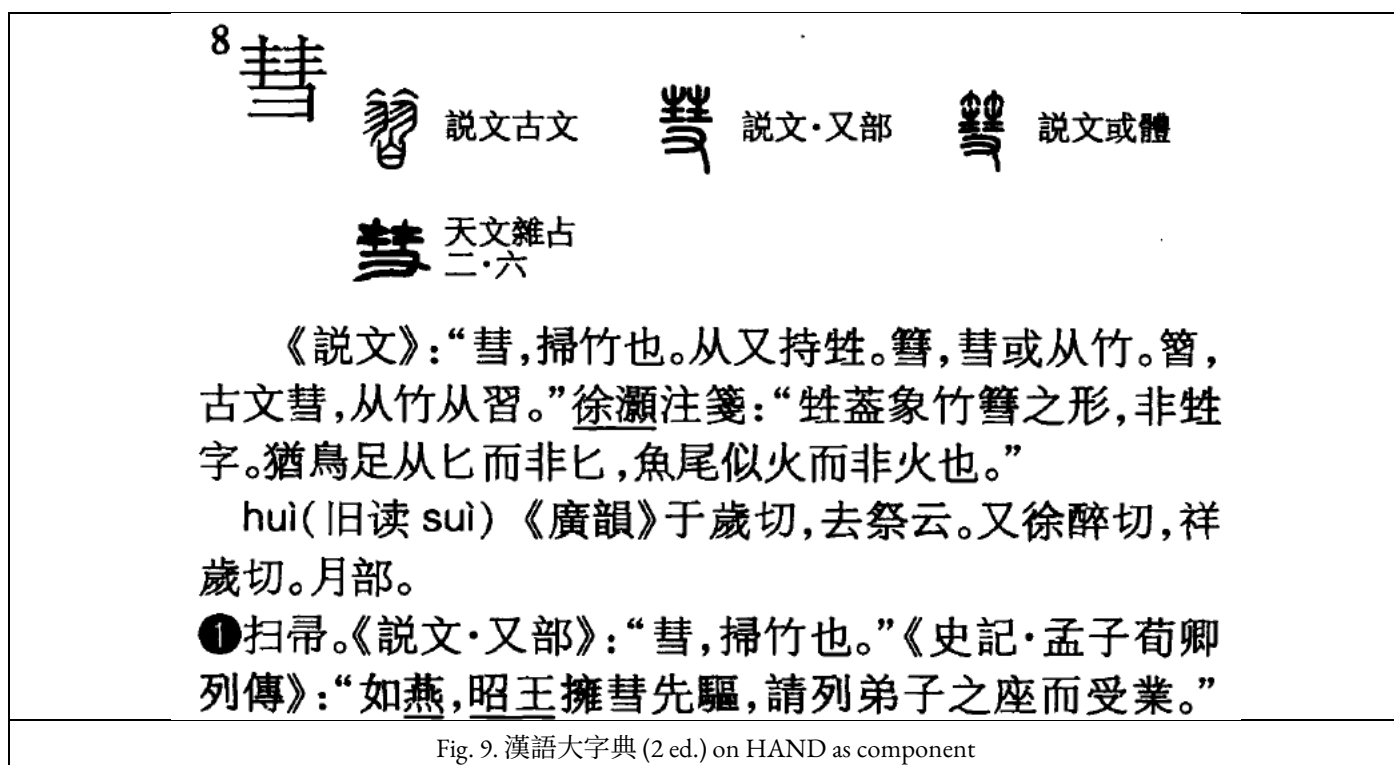
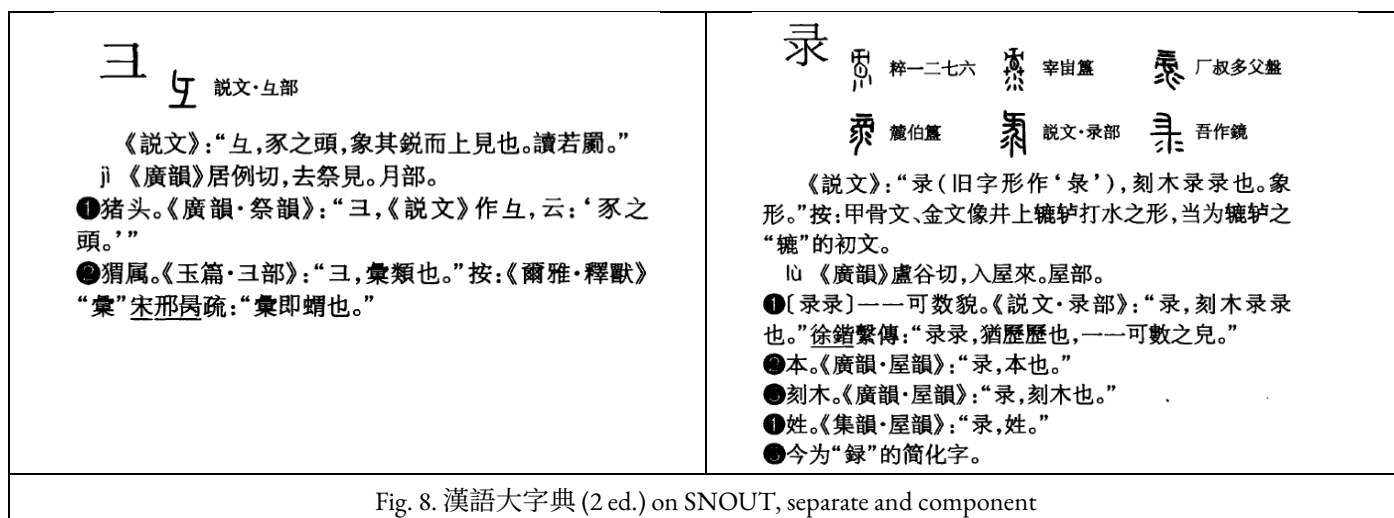
| | | | | | | |
|----------------|---------|--------|---------|---------|---------|---------|
| 5F50 𠄎 58.0 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 |
| | G0-6566 | H-C6D0 | T3-2140 | J1-3C77 | K2-313A | V1-5454 |

Fig. 6. The Unicode chart for U+5F50

As we observe, even the Vietnamese form was influenced by the PRC decision. As Lee Collins (p. c.) reports, the Vietnamese usage of the character is a unification of two characters, one with the reading *kẻ*, which corresponds to Mandarin *jì* and is thus connected etymologically to SNOOUT, while the other (used in Tày language with the meaning *mự* “a time or turn”), *might* be derived from a LONG HAND-using 歸, but has an unambiguously SNOUTy form in the reference (Figure 7).

| |
|---|
| <p>𠄎 MỰ²</p> <p>Lân. ĐN. Vữa.</p> <p>許免嗜双𠄎三𠄎</p> <p>Hầu mìn khăn soong mự sam mự (Kh.T.Q.S)</p> <p>(Để cho nó gáy hai lần ba lần).</p> |
| Fig. 7. Từ điển chữ Nôm Tày |

Note, however, that the dictionaries of the PRC are not consistent with implementing the decision (acknowledged by the PRC response as well), and the most authoritative editions in the matter of graphic correctness separate them (see Figures 8–9).



The LONG HAND is not used in the PRC standard, as HAND 彑 is treated as the graphic norm; thus, any character containing a HAND will be rendered with 彑; for example, ‘snow’ 雪, possessing a simplified 彗 in the bottom part, is not rendered with 彑.

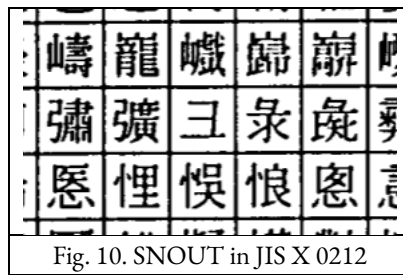
Additionally, I should mention that not even all fonts aimed at PRC usage follow the “U+5F50 = HAND” principle; while this seems to be generally true for Song (sans-serif) fonts as well as heiti (sans-serif), fonts derived from the regular script consistently show SNOOT, though perfectly able to produce HAND as component (see FangSong in Table 2):

| Font name | 5F50 | 'Snow' |
|-----------|------|--------|
| FangSong | ㄣ | 雪 |
| KaiTi | ㄣ | 雪 |

Table 2. Kaiti fonts shipped with OS Windows

Japan

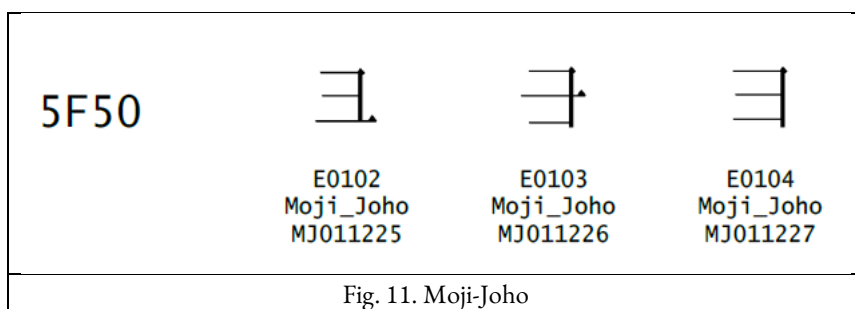
Japanese standard JIS X 0212-1990 (J1), to which the code chart refers, includes a version of the characters under consideration. Under 28-87, it incorporates the character U+5F50 (Figure 10):



The characters surrounding it make it clear that it is SNOUT that is meant. The combination U+5F50 U+E0100 as well as just U+5F50 in both the 1990 and 2004 versions of the mappings (as well as U+2F39) are mapped to this glyph.

But this is far from the end of the story, as both of the remaining forms occur in Japanese names, and thus an *ad hoc* solution was required. Adobe-Japan1 incorporated U+5F50, U+E0101 as the denotation of U+2E95, depicted there as ㄣ, LONG HAND. But this is a conservative solution.

Moji-Joho, on the other hand, uses all three (Figure 11):



Obviously, as the characters are non-cognate, such a solution is suboptimal, but it allows to encoded all the necessary forms.

Finally, we must observe that, unlike PRC, in Japan the difference between the forms of HAND was considered not a question of style but a method of separation between the frequent and rare characters (approximately Jōyō vs. non- Jōyō), and thus both occur: for comparison, see the ‘broom’ 箒 [Jinmeiyō] vs. ‘snow’ 雪 [Jōyō].

Hong Kong, Taiwan, Macau

All the sources from these locations provided in the code chart do refer to SNOUT.

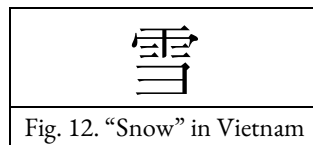
Taiwan, however, maintains HAND as well, which is stored at Plane 8, 7359 of CNS-11643 (<https://www.cns11643.gov.tw/wordView.jsp?ID=553817>); the Unicode mapping it is forced to use is U+2E95 in the “CJK Radicals Supplement” block.

Korea

Korea maintains only the 冫 glyph, as K2 [KS X 1027-1:2011 (formerly PKS C 5700-1 1994)], character 313A.

Vietnam

As alluded before, HAND as component takes in Vietnamese glyphs a unique form with the horizontal and the vertical joining at the same location; the Nom Na Tong form of “Snow” will be sufficient as an example (Figure 12):



Yet, their representative glyph of U+5F50 is a copy of the Chinese one (Figure 13).

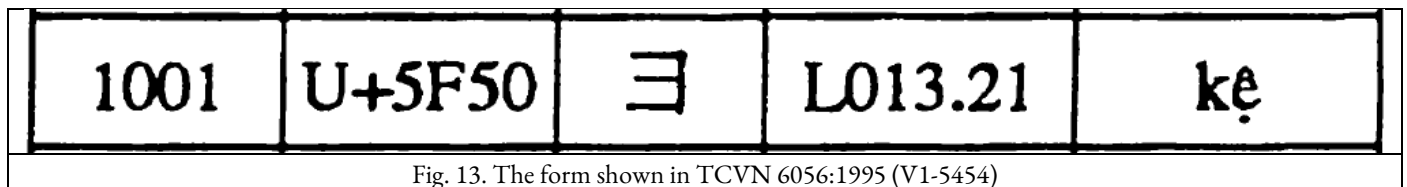


Fig. 13. The form shown in TCVN 6056:1995 (V1-5454)

Yet, the fact that it appears in the substring

彈彊冫形彩

is very much suggestive of SNOUT.

As the definitions and readings in Nom Foundation dictionary show, at least some of the intended usages of U+5F50 in Vietnamese refer to SNOUT, not HAND (Figure 14).

| Quốc Ngữ | Hán-Nôm | Codepoint | Context | Ref. | English |
|----------|---------|------------------------|-----------------|----------------------|--------------------------|
| kê | 𠄎 | U+5f50 | mặc kê, thầy kê | vhn | not to pay attention to |
| kẹ | 𠄎 | U+5f50 | lúa kẹ | btcn | snout; KangXi radical 58 |
| kí | 𠄎 | U+5f50 | bộ kí | gdhn | snout; KangXi radical 58 |

Fig. 14. U+5F50 in Vietnamese

CAN'T THEY STAY UNIFIED?

No, they cannot, I argue, and this is why. Though it worked out before, consistency required the disunifications, for at least two reasons (beside the fact that they are obviously non-cognate): one conceptual and one practical.

Correspondence with Radicals

U+5F50 is considered the Unified version of the radical 2F39 (Fig. 15):

| | | |
|------|---|-----------------------------------|
| 2F39 | 𠄎 | KANGXI RADICAL SNOOT ≈ 5F50 𠄎 |
| 2E94 | 𠄎 | CJK RADICAL SNOOT ONE → 5F51 𠄎 |
| 2E95 | 𠄎 | CJK RADICAL SNOOT TWO → 5F50 𠄎 |

Fig. 15. Code charts for radicals

However, the unfortunately named SNOOT TWO (which should have better been CJK RADICAL AGAIN, referring to U+2F1C 𠄎, the main allomorph of HAND in modern Sinitic writing) is also matched to U+5F50, without a better place to be sent (not to U+53C8 𠄎, surely?). This is the only exception to the general principle that the Unicode Kangxi Radicals and CJK Radicals each have a corresponding Unified analogue, with different characters sent to different Unified Ideographs.

Second Stage

The addition of Extension G introduced the Second Stage Simplifications. These Second Stage Simplifications (第二次 汉字简化方案—草案, SSS) were an abortive project of PRC government in late 1970s. They were supposed to become a continuation of the highly successful campaign of simplifications implemented throughout the 1950s-60s. Part One was released on 20th December 1977 and was consistently used in all the publications in 人

民日報 *Rénmín Ribào* until July 1978. It was widespread during that period and gained mass currency, but, announced a failure, soon practically dropped out of usage, though an official withdrawal (though not declaration of abandoning any simplification plans) was postponed until 24th June 1986.

The characters of Part One (except those produced by analogy, that have been separately offered to WS-2020) are now encoded in Unicode. They include the following (Figure 16):

| | | | | |
|-------------------------|-----------------|---|---|--------------|
| 1.1.136 | ㄣ ¹¹ | ㄣ | ㄣ | 96EA [TGH 1] |
| Fig. 16. The new “Snow” | | | | |

Basically, for “Snow” the top is cut, retaining just HAND. And, just like that, HAND gets a new semantic meaning. Now U+5F50 is overloaded, playing the roles of two different morphemes of Mandarin: *jì* ‘pig’s head, name of a radical’ – and *xuě* ‘snow.’ And while their graphical identity was probably desirable by the compilers of the actual simplification scheme in the 70s – after all, they tried to make ㄣ and ㄣ graphically identical all this time – but not to the current use, when any font maker deciding to support Second Stage Simplifications but render SNOOT and any kind of HAND differently is forced to the complex decision of keeping them the same, though they might co-occur, even contrastively, in the same sentence.

The character UTC-01005 in the U-Source was introduced by Andrew West in order to represent unambiguously the simplification of “Snow”; however, it was withdrawn (wrongly, I argue) due to the identity with U+5F50.

The arguments proposed above, I hope, make it sufficiently clear that U+5F50 should be split, with the forms of HAND being separate.

POSSIBLE SOLUTION

However, the solution offered by the PRC response is utterly untenable. It offers the following:

‘...China decides to move the G glyph and source reference out of the code point and put it in the next working set, and submit a new one for U+5F50 to match the common view.’

No change should ever alter the correspondences of the standard as basic as GB 2312-1980, frequently used on simpler devices that will never be able to update to the newer versions of Unicode; furthermore, if a new one would be submitted, it will likely be encoded outside the BMP, which is highly undesirable.

In fact, the following is clear: the location in the URO, the matching with the standards (especially Japanese JIS 0212) already implied in the table, even the graphical forms of the radicals as now presented in Unicode 13.0 lead to the conclusion that the codepoint U+5F50 must be the Unified correspondence of the radical SNOOT. And this means only one solution.

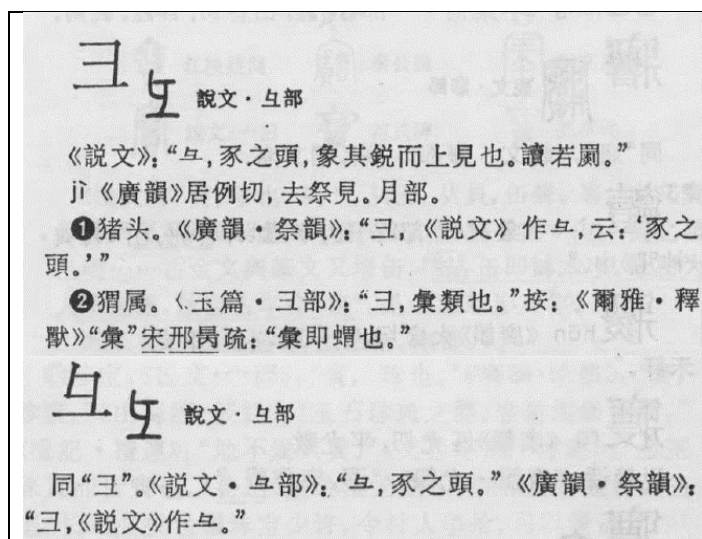

I propose to alter the graphical form of G0-6566 itself. If the representative glyph in the actual standard GB 2312-1980 is made to represent SNOUT and not HAND, the relationships between various standards and intentions become clearer. It is HAND (preferable in the LONG HAND form which is orthodox in the Kāngxī Dictionary) that needs to be re-encoded, preferably appended to the URO as this would (if the corresponding, though optional, rewriting of the IDS is undertaken) fix the contradiction that a TIP character can be used in the IDS of BMP characters. The Second Stage simplified “Snow” can be just unified to this new character.



Furthermore, even if the graphical form of G0-6566 (and V1-5454) is not altered, this is not an impediment towards the disunification of U+5F50. It is well-known that for some distinct codepoints graphical identity in some particular locales does not invalidate separation. Even with older forms of G0-6566 and V1-5454 the Mainland and Vietnam may just choose the present their glyphs identical to the newly encoded characters.

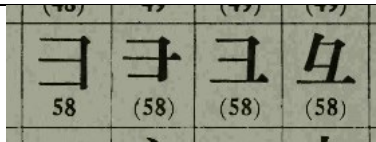
APPENDIX: VARIOUS ATTESTATIONS

This section provides various attestations of characters SNOUT, HAND, and LONG HAND being distinct from the dictionaries and sources including those that happen to be used as source references in Unicode.

漢語大詞典 (GHC)

| | |
|---|---|
|  |  |
| p. 960, only SNOUT definitions | p. 4058, the graphical rendering of a HAND-component |

| | | | | | | | |
|---|---|-------|---------|-------|---|--------|-------|
|  4204 ■ 4-3-1 | ▶KEIGASHIRA <table border="1" data-bbox="683 1668 1013 1720"> <tr> <td>58</td> <td>Radical</td> <td>U2E95</td> </tr> <tr> <td>彡</td> <td>S3-3-0</td> <td>F0000</td> </tr> </table> radical variant → see 彡 at 4205 | 58 | Radical | U2E95 | 彡 | S3-3-0 | F0000 |
| 58 | Radical | U2E95 | | | | | |
| 彡 | S3-3-0 | F0000 | | | | | |
|  4205 ■ 4-3-1 | ▶KEIGASHIRA <table border="1" data-bbox="683 1870 1013 1921"> <tr> <td>58</td> <td>Radical</td> <td>U5F50</td> </tr> <tr> <td>彡</td> <td>S3-3-0</td> <td>F0000</td> </tr> </table> | 58 | Radical | U5F50 | 彡 | S3-3-0 | F0000 |
| 58 | Radical | U5F50 | | | | | |
| 彡 | S3-3-0 | F0000 | | | | | |
| RADICAL <i>keigashira</i> 'pig's head', used for character classification Variants: 彡 彡 <i>keigashira</i> (彡彡彡) | | | | | | | |
| <i>The Kodansha Kanji Dictionary Updated and Revised Edition</i> (Kodansha, 2013): treats them as equivalent, but requires LONG HAND in plain text | | | | | | | |






Andrew N. Nelson's *The Modern Reader's Japanese-English Character Dictionary Second Revised Edition*: requires all three distinguished in plaintext

| | |
|---------|---------|
| 全字庫字型 | 全字庫字型 |
| | |
| T3-2140 | T8-7359 |

| | | |
|------------------------------------|----------|--|
| ヨ | MJ011225 | <p>ヨ jì 音计] ①猪头。见《广韵》。②刺猬一类的动物。见《玉篇》。</p> <p>ヨ 曾作“雪”的简化字,后停用。见《第二次汉字简化方案(草案)》。</p> |
| ヨ | MJ011226 | |
| ヨ | MJ011227 | |
| Moji-Joho: all three distinguished | | GZH-0657.02: Second Stage Simplification for “Snow”, specifically opposed to SNOUT |

| | |
|--|---|
| | <p>ヨ siet [θi:tʰ] 雪。</p> <p>雪 (ヨ)</p> <p>〈方〉siet [θi:tʰ] 前面加“比”，义为假如；假设；如：比～。beijsiet. 比如；假如。</p> |
| GB3001-1997 lists all three, counter the previous intention of unification | 古壮字字典 p. 459 (GZ): possible G-Source |

| | | | | | | | |
|--|-------|---|---|----------|-------|--|--|
| 105 | Z3-03 |  | babelstone-41 GF1997-212 GF2009-415 | top of 寻 | 倥 𠂇 妇 | | katakana 𠂇, there is ongoing discussion about disunification of U+5F50 |
| 106 | Z3-04 |  | babelstone-42 GF1997-213 GF2009-224 | top of 录 | 𠂇 录 彙 | | there is ongoing discussion about disunification of U+5F50 |
| 107 | Z3-05 |  | babelstone-43 irgn2225-Z3-01 GF1997-214 cdp-8bab | top of 𠂇 | 争 𠂇 事 | | there is ongoing discussion about disunification of U+5F50 |
| L2/21-134 “Collections of ideograph components for use in IDSeS” | | | | | | | |