L2/02-052

Proposal to add the Ideographic Taboo Variation Indicator to ISO/IEC 10646-1 and Unicode™

John H. Jenkins
Apple Computer, Inc.

8 February 2002

A. Administrative

1. Title:

Request to add Ideographic Taboo Variation Indicator to ISO/IEC 10646-1 and Unicode

2. Requester's name:

John H. Jenkins

3. Requester type (Member body/Liaison/Individual contribution):

Individual contribution

4. Submission date:

8 February 2002

5. Requester's reference (if applicable):

N/A

6. (Choose one of the following:)
This is a complete proposal: ; or, More information will be provided later:

This is a complete proposal

B. Technical - General

1. (Choose one of the following:)

b. The proposal is for addition of character(s) to an existing block:
Name of the existing block:

CJK Symbols and Punctuation (U+3000—U+303F) would be ideal, but there's no room left there. Somewhere like U+2FFF might therefore be a candidate to keep it as close as possible to U+303E IDEOGRAPHIC VARIATION INDICATOR. (U+3040 is close and unoccupied, but that would make the character abut Hiragana, which seems weird.)

2. Number of characters in proposal:

1

3. Proposed category (see section II, Character Categories):

B.1 Specialized (Small Collections of Characters)

4. Proposed Level of Implementation (see clause 15, ISO/IEC 10646-1):

Level 1

Is a rationale provided for the choice?
If Yes, reference:

This is not a combining character.

5. Is a repertoire including character names provided?:

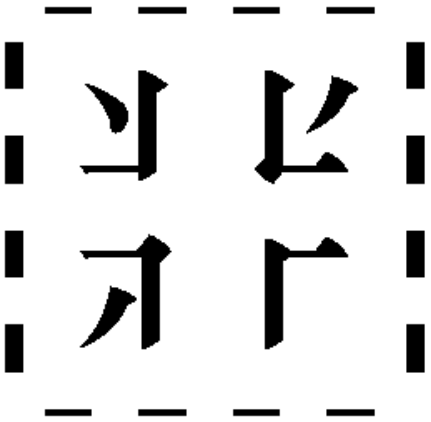
Yes, IDEOGRAPHIC TABOO VARIATION INDICATOR

a. If YES, are the names in accordance with the 'character naming guidelines' in Annex K of ISO/IEC 10646-1?

Yes

b. Are the character shapes attached in a reviewable form?

Yes



6. Who will provide the appropriate computerized font (ordered preference: True Type, PostScript or 96x96 bit-mapped format) for publishing the standard?

John H. Jenkins

If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

N/A

7. References:
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

Yes, see below

b. Are published examples (such as samples from newspapers, magazines, or
other sources) of use of proposed characters attached?

Yes, see below

8. Special encoding issues:

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information):

Yes

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?

No

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?

Yes, other experts

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?

The character would be used by sinologists transcribing historical Chinese texts

4. The context of use for the proposed characters (type of use; common or rare)

Rare

5. Are the proposed characters in current use by the user community?

Not in computerized text

6. After giving due considerations to the principles in N 1352 must the proposed characters be entirely in the BMP?

Yes. This is a single character similar in function to U+303E IDEOGRAPHIC VARIATION INDICATOR and is most naturally encoded near it.

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?

N/A

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

No

9. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

No

10. Does the proposal include use of combining characters and/or use of composite sequences (see clause 4.11 and 4.13 in ISO/IEC 10646-1)?

No

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

N/A

11. Does the proposal contain characters with any special properties such as control function or similar semantics?

Yes.

There is no distinction within Chinese between names and other words written with ideographs, it is entirely possible that the ideograph used for any individual's name will occur in a normal text. This can be awkward when the individual is someone of high rank or otherwise deserving of special respect; using the ideographs from their name in ordinary writing was traditionally seen as disrespectful and unacceptable. To get around this, the practice developed of using "taboo" forms (避諱 or 避讳) for these ideographs in normal texts. The ideographs would be deliberately distorted (usually, but not always by omitting the final stroke) to avoid writing the personal name of the individual in question. The precise form of the taboo form is not specified and not necessarily predictable.

The purpose of the IDEOGRAPHIC TABOO VARIANT INDICATOR is to mark places in a text where this has been done. An IDEOGRAPHIC TABOO VARIANT INDICATOR followed by an ideograph indicates that the ideograph was written with a taboo form in the original text. The precise nature of the taboo form is not specified (and is, indeed, irrelevant for purposes of transcribing the text).

If the font provides for this, the combination of the IDEOGRAPHIC TABOO VARIANT INDICATOR plus ideograph may be treated as a ligature and directly rendered by a taboo form glyph for the given ideograph. Otherwise, it should have a visual appearance such as specified in this proposal.

The following examples regarding the use of the taboo forms are provided by Richard Cook of UC Berkeley:

Here're 3 examples of [U+907f][U+8af1] (bihui) from p. 202 of <Song Ben Guang Yun> (SBGY) the Song Dynasty rhyming dictionary (YU Nae-wing, Chinese U. of Hong Kong, 1993):

http://linguistics.berkeley.edu/~rscook/images/SBGY/SBGY-YNY-202.jpg

The 3 green arrows indicate taboo-deformed head entries, specifically, deformations of [U+6046], [U+63ef] and [U+7dea]. You can see the 1st 2 of these same 3 hanzi deformed also in the definitions.

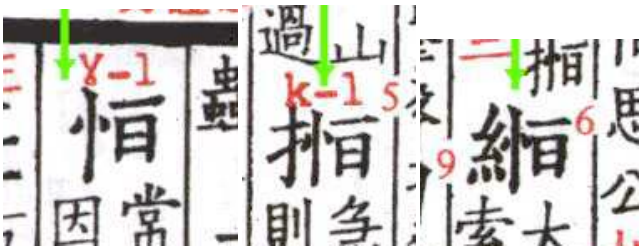
The tabbo deformation in all 3 is omission of the last heng stroke.

The "correct" (undeformed) writings of [U+63ef] and [U+7dea] can be seen in the (modern) footnotes 5 and 6 at the bottom of the page.

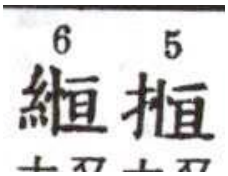
Note that [U+63ef] and [U+7dea] both have [U+6046] as phonetic, and so Taboo Deformation of [U+6046] spreads to all hanzi in which [U+6046] is a component. This is a general truth of TD, which we might call the "Taboo Deformation Spreading Principle" (TDSP).

Note that the TD forms of [U+6046], [U+63ef] and [U+7dea] are all non-hanzi, which is to say that omission of the last stroke does not result in confusion with preexisting non-TD hanzi. I have yet to see a case in which regular omission of the final stroke would result in confusion, and I have also not seen a case in which a TD (or TDSP) hanzi omits a stroke other than the final stroke. But I may just not have looked hard enough yet.

Here are the three taboo forms from Richard's scan:



These are the two non-taboo forms in the footnotes:



==========

D. SC 2/WG 2 Administrative (To be completed by SC 2/WG 2)

1. Relevant SC 2/WG 2 document numbers:

_____

2. Status (list of meeting number and corresponding action or disposition):

_____

3. Additional contact to user communities, liaison organizations etc:

_____

4. Assigned category and assigned priority/time frame:

_____