# Open Issues on Phags-pa Encoding

**For consideration of Phags-pa ad hoc at Meeting 46, Xiamen**

**Peter Constable**

**Expert Contribution**

**2004-1-22**

This document summarizes issues to be resolved in encoding Phags-pa script, with particular attention to differences between proposals submitted by Andrew West and China/Mongolia.

Relevant documents submitted by each party are as follows:

Submitted by Andrew West:

| Doc # | Title | Note |
|---|---|---|
| N2622 | *Proposal to Encode the Phags-pa Script* | initial proposal, accepted at WG2 M44 and basis for current text in PDAM2 |
| N2719 | *Response to Comments on Phags-pa Proposal in N2706* | response to China's PDAM1 ballot comments |
| N2771 | *Comments on the Chinese-Mongolian joint proposal to encode the HPhags-pa script* | comments on N2745 |

Submitted by China/Mongolia:

| Doc # | Title | Note |
|---|---|---|
| N2666 | *Principles on Encoding Phags-pa Script* | initial comments on N2622, prior to WG2 M44 |
| SC2/N3730 | *Summary of Voting* (PDAM1) | China comments on Phags-pa accepted for PDAM1 (≡ N2622), prior to WG2 M45 |
| N2706 | *Summary of Voting* (Amd. 1 subdivision) | China comments on N2622 (same comments provided in SC2/N3730) |
| N2745 | *HPhags-pa script encoding* | alternate proposal from China and Mongolia |
| N2869 | *Proposal to Encode the Phags-pa Script* | revised alternate proposal |
| N2870 | *Summary of the Revised User's Agreement Related to Phags-pa Script* | |
| N2871 | *Some Problems on the Encoding of Phags-pa Script* | |

It must be borne in mind that N2622 is the basis for what is currently in PDAM2, and that if any changes are to be made in PDAM2 they must be expressed as changes to the content already contained therein.

Where there is disagreement about character names, this document uses names currently in PDAM2 except when explicitly discussing alternate names.

Examples in this document use Andrew West's font solely because it was readily available. No preference of font for use in ISO/IEC 10646 is implied by this.

# 1 Name of script

*Status: resolved*

A consensus was reached in the ad hoc meeting during M45 to use “Phags-pa”.

All references to the script in this document reflect this decision.

# 2 Distinctness of Phags-pa in relation to other scripts

*Status: resolved*

In ballot comments on PDAM1, China raised a concern that Phags-pa not be considered a variant of Tibetan. (Comment T14 in China’s ballot comments on PDAM1.)

It was unanimously held by those in the ad hoc meeting during M45 that Phags-pa script is related to Tibetan script but is considered distinct and is not regarded as a variant of the Tibetan script.

# 3 Choice of font

*Status: open*

N2622 uses a modern design derived from printed texts of the Yuan dynasty. China requests the use of a different font in the “Khubilai” style, as is used in N2869.

(This corresponds to T10 in China’s ballot comments on PDAM1.)

A consensus was reached in the ad hoc meeting during M45 that a different font could be used. A critical issue is availability of a font to the editors of ISO/IEC 10646 and to the Unicode Consortium.

# 4 Specific glyph shapes

*Status: resolved*

In comment T11 of China’s ballot comments on PDAM1, concern was expressed over glyph shapes for PHA and E used in N2622 and that appeared in PDAM1 (and since, in PDAM2). In N2719, Andrew West suggested alternative glyph shapes.

A consensus was reached in the ad hoc meeting during M45 that the alternative glyph shapes suggested in N2719 were better and resolve the problem.

# 5 Names of uncontested characters

*Status: open*

## *5.1 Description*

Among the characters common to both proposals, there were many differences between names used in N2745 and those in N2622. (This corresponds in part to T12 in China's ballot comments on PDAM1.)

The revised China/Mongolia proposal in N2869 resolves most of these differences, though two remain:

| Letter | Name in PDAM2 | Name in N2869 |
|---|---|---|
| 𡨴 | PHAGS-PA LETTER AA | PHAGS-PA LETTER MINISCULE A |
| 𡩡 | PHAGS-PA LETTER GGA | PHAGS-PA LETTER QA |

## *5.2 Discussion*

There is a naming conflict in N2869, as PHAGS-PA LETTER QA is used for both 𡩡 and 𡨸.

# 6 Punctuation of Mongolian or Chinese origin

*Status: open*

## *6.1 Description*

N2622 documented the use in Phags-pa text of certain punctuation characters of Mongolian or Chinese origin, and concluded that these can be unified with existing characters in the UCS. In N2745, China and Mongolia proposed distinct Phags-pa punctuation characters, but have revised their proposal in N2869 to unify punctuation with existing characters in the UCS.

There is not complete agreement on what existing UCS characters are to be used, however, specifically in regard to the small-circle punctuation mark:

| Mark | UCS character assumed in N2622 | UCS character assumed in N2869 |
|---|---|---|
| ᠂ | U+1802 MONGOLIAN COMMA | U+1802 MONGOLIAN COMMA |
| ᠃ | U+1803 MONGOLIAN FULL STOP | U+1803 MONGOLIAN FULL STOP |
| ᠅ | U+1805 MONGOLIAN FOUR DOTS | U+1805 MONGOLIAN FOUR DOTS |
| ◌ | *U+3002 IDEOGRAPHIC FULL STOP* | *U+02DA RING ABOVE* |

## *6.2 Discussion*

The small-circle punctuation mark is used in both Chinese and Mongolian text, and for both of these the appropriate UCS character is U+3002 IDEOGRAPHIC FULL STOP. Given the likelihood of Phags-pa text being set together with either Mongolian or Chinese text, it seems appropriate to use the same character for the Phags-pa punctuation mark.

If these marks are likely to be used in mixed Phags-pa and Mongolian text, then it is important to consider whether font implementations can use the same glyphs for both scripts. If the same glyphs would not be suitable for both, there may be benefits to disunification.

# 7 Punctuation of Tibetan origin

*Status: open*

## *7.1 Description*

N2622 documents the use of certain punctuation signs of Tibetan origin:

| Mark | Name in PDAM2 | Similar Tibetan character or character sequence |
|---|---|---|
| 𐡷 | PHAGS-PA SINGLE HEAD MARK | U+0F04 TIBETAN MARK INITIAL YIG MGO MDUN MA |
| 𐡸 | PHAGS-PA DOUBLE HEAD MARK | < U+0F04 TIBETAN MARK INITIAL YIG MGO MDUN MA, U+0F05 TIBETAN MARK CLOSING YIG MGO SGAB MA > |
| 𐡹 | PHAGS-PA MARK SHAD | U+0F0D TIBETAN MARK SHAD |
| 𐡺 | PHAGS-PA MARK DOUBLE SHAD | U+0F0E TIBETAN MARK NYIS SHAD |

N2871 suggests that the shad marks can be unified with the corresponding Tibetan characters, and that the head marks can be represented as U+1800 MONGOLIAN BIRGA "and its variant" (i.e., the double head mark would be represented using a variation-selector sequence).

N2622 considers possible unification of the head marks with Tibetan characters, but argues against this on the grounds that the two marks are used contrastively.

## *7.2 Discussion*

N2622 does not provide rationale for disunification of the SHAD and DOUBLE SHAD from Tibetan counterparts.

N2622 mentions that the DOUBLE HEAD MARK corresponds to a ligature of < 0F04, 0F05 >, but does not discuss the possibility that this mark be represented by precisely this sequence. It also does not discuss a potential relationship to the related Mongolian punctuation marks, which is reasonable to consider given the use of other Mongolian punctuation in Phags-pa text.

If these marks are likely to be used in mixed Phags-pa and Tibetan text, then it is important to consider whether font implementations can use the same glyphs for both scripts. If the same glyphs would not be suitable for both, there may be benefits to disunification.

# 8 Inter-syllable delimitation

*Status: open*

## *8.1 Description*

Phags-pa is a cursively-connecting script, like Mongolian and Arabic. At the boundaries between syllable clusters, the cursive connection between characters is broken; typically, a small

amount of white space also appears. For these reasons, it is necessary that syllable clusters be separated by some non-connecting, white-space character.

(This issue constitutes part of comment T13 in China's ballot comments on PDAM1.)

N2622 proposes that U+0020 SPACE be used between each syllable cluster (as also between words).

N2869, however, states that the advance between syllables must be 1/3 of a space, and on that basis proposes that U+202F NARROW NO-BREAK SPACE be used between syllables, but U+0020 SPACE between words.

### 8.2 Discussion

It is noted that, in Mongolian text, SPACE is used between words, while NARROW NO-BREAK SPACE is used for smaller gaps that break cursive connection at certain word-internal boundaries.

While NNBSP would provide the appropriate behaviour in relation to cursive connection (breaking the connection between syllables), it is unclear whether it would provide the desired default line-breaking behaviour. If NNBSP were used between every syllable, the default behaviour would be that lines would break at word boundaries but would not break at syllable boundaries. (Tailored line-breaking implementations could provide different behaviour, however.)

## 9 Encoding of conjoining consonants in consonant clusters

*Status: resolved in relation to WA and YA; open in relation to RA*

### 9.1 Description

Consonants WA, YA and RA have alternate presentation forms when they occur as the second consonant in a cluster. Thus, there is a contrast between  /hay/ and  /hya/. Also, RA has an alternate presentation form when it occurs as the first consonant in a cluster. Thus, there is a contrast between  /rang/ and  /rnga/.

(Note: these Phags-pa consonants, like all Phags-pa letters in general, have cursively-connecting forms – isolate, initial, medial and final. Both proposals assume that cursive-connecting forms have no linguistic significance and are determined by context. These variant forms for WA, YA and RA, however, are linguistically significant and, thus, are a distinct issue from cursive-connecting forms.)

N2622 encodes the alternate forms of these consonants as distinct characters, following the model used for Tibetan script in the UCS.

N2869 does the same for WA and YA, though not for RA. Instead, it proposes that the variant RA forms be encoded using variation-selector sequences.

(This corresponds to T8 in China's ballot comments on PDAM1.)

### 9.2 Discussion

The basis for the encoding model for RA proposed by China and Mongolia initially in N2745 is that the UCS follows a character-glyph model in which various presentation forms of a character are all encoded in terms of a single character.

In the case of the forms in question, they have semantic significance; they are not predictable on the basis of context alone, nor are they aesthetic variants selected at the discretion of a user. The character-glyph model does not stipulate what encoding mechanisms must be used in such cases, and different options exist:

1. Encode the presentation forms as distinct characters.
2. Encode a character that reflects the linguistic context within a consonant cluster (lack of inherent vowel) and that controls the corresponding presentation.
3. Use variation-selector sequences.

The first approach has a precedent in the case of Tibetan script, from which Phags-pa was historically derived. The second approach has a precedent in the virama used for most other Indic scripts. (The scripts using a virama model include Khmer, which is like Phags-pa in that there is never an overt "halant" marking a dead consonant.) In both cases, the presentation forms have the same linguistic significance in Phags-pa as in the other scripts.

The precedent for use of variation-selector sequences is Mongolian script, which also was an historical influence in the development of Phags-pa. Mongolian variation selectors are not used for presentation forms that have linguistic significance as in the case of these presentation forms in Phags-pa, or their counterparts in Tibetan and other Indic scripts, however.

It should be noted that some software processes will want to capture the linguistic relationship between the various forms for each of the three consonants, RA, WA and YA, and that existing implementations that support Tibetan will already have the means to do this using the first model. More importantly, the means to do this will already be incorporated into software that supports Phags-pa to deal with the WA and YA cases, and would require only minor changes to do this for RA as well.

## 10 Cursive-connecting forms

Phags-pa is a cursively-connecting script, like Mongolian and Arabic. Within a syllable cluster, letters are generally cursively connected, and letters generally have distinct isolate, initial, medial and final forms. Some letters do not require a change in form when connecting on their leading edge, and so have the same form in isolate and final contexts. Similarly, some letters do not require a change in form when connecting on their trailing edge, and so have the same form in initial and medial contexts.

Cursive connection in Phags-pa has an additional factor not found in Mongolian or other cursive-connecting scripts: some letters connect toward the baseline (on the right) while other letters connect away from the baseline (on the left).

Issue related to cursive connection are divided into two parts: basic selection of connecting forms, and selecting of left- versus right-connecting forms.

## *10.1 Basic selection of connecting forms*

*Status: resolved*

In general, selection of connecting glyph forms – initial, medial, final or isolated forms – can be determined by context alone. It is necessary to be able to select isolate, initial, medial or final forms contrary to contextual expectations, however; e.g., to display a medial form in isolated context.

Both N2622 and N2745 assume that connecting forms are normally determined by context, and that ZWJ and ZWNJ can be used to select isolate, initial, medial or final forms contrary to contextual expectations.

## *10.2 Selection of left- versus right-connecting forms*

*Status: open*

### 10.2.1 Description

In general, selection of left- versus right-connecting glyph forms can be determined by context alone. It is necessary to be able to select right-connecting and left-connecting forms contrary to contextual expectations, however; e.g., to display the right-connecting form of a letter following a left-connecting letter.

N2622 proposes that explicit selection of left- versus right-connecting forms be done using some control character inserted after the letter in question. It suggests the possibility of a new control character for this purpose, but leaves open the issue of what control character should be used.

N2745 and N2869 are unclear in their explanation of how selection for the side on which connection occurs is controlled, thought they do mention a proposed JOINER character in this regard. (All of the examples in N2869 include a sequence of < JOINER, VS1 >.) N2871 provides a marginally clearer description (though considerable uncertainty remains due to the lack of examples and possible errors of “left” for “right”) and appears to employ a revised model from that in the earlier documents: the JOINER character is placed between letters to select the side for connection of the following letter as being *opposite* of that of the preceding letter.

Assuming this interpretation of N2871 is correct, both proposals use a single character for explicit selection of left- versus right-connecting forms. In other respects, though, the mechanisms differ, as shown in the following examples. (To reduce non-essential differences, the behaviour-controlling character for both proposals will be referred to as ALTERNATE JOINING CONNECTION OVERRIDE – AJCO.)

| Text element | Sequence proposed by N2622 | Sequence proposed in N2871[1] |
|---|---|---|
| 𑙐 | < TTHA, I > | < TTHA, I > |
| 𑙖 | < TTHA, I, AJCO > | < TTHA, AJCO, I > |
| 𑙖 | < THA, I > | < THA, I > |
| 𑙐 | < THA, I, AJCO > | < THA, AJCO, I > |

It is important to note that it is unclear whether the correct interpretation of N2871 has been made here. In particular, it should be noted that N2870, which was submitted concurrently with N2871, lists variation-selector sequences for selecting left- or right-connecting forms.

### 10.2.2 Discussion

Both of the proposals illustrated by the examples above assume that the alternate connecting behaviour can be specified at the juncture between any pair of letters and is not limited to certain letters only. Also, both would *reverse* the connection from the side that was contextually expected as opposed to selecting an absolute side for the connection (e.g., presence of a control character forces the second letter to connect specifically on the left).

There is no precedent for such a control function in the UCS, though this is certainly the case because no script already encoded in the UCS has similar behaviour.

Since it is unclear whether China and Mongolia intend to propose variation-selector sequences to select left- versus right-connecting forms, this possibility will be considered here.

If VS sequences are used to select right- versus left-connecting forms, there would be certain implications for this connecting behaviour. First, since VS sequences select particular glyph forms, a given sequence does not *reverse* the side on which connection is made from what would be contextually expected; rather, an absolute selection of left or right connecting forms would be made. Also, each VS sequence must be explicitly defined. Therefore, a contrary-to-context left- or right-connecting form could be used only in certain pre-specified cases. Indeed, N2870 proposes exactly eight such VS sequences.

The use of VS sequences for this purpose in N2870 results in certain inconsistencies and complexities:

- In most cases, the VS sequence results in a left-connecting form, though in one case the VS sequence results in a right-connecting form.
- Because N2870 proposes VS sequences for other purposes as well, different VS characters are used to control the side of connection in different cases: VS1 is used in six cases, and VS2 and VS3 are each used in one case.
- N2870 uses VS sequences for another independent function – selection of free-variation glyph variants – that intersects with this function. Thus, < E, VS1 > selects the triple-toothed form of E, < E, VS2 > selects left-connecting E (*in final context only* – in initial context, < E, VS2 > selects nominal-E with preposed A), and < E, VS3 > selects both the left-connecting and triple-toothed form of E.

[1] This assumes that the interpretation of the intent of N2871 is correct, which is by no means certain.

The use of sequences that include a variation selector or other control character introduces an issue of the interaction between this control mechanisms and ZWJ or ZWNJ. In the case of VS sequences, a particular connecting form would be selected, so it is unclear what the result should be of such a sequence preceded or followed by ZWNJ.

For a control other than VS sequences, there is less of a problem: presumably, the presence of a JOINER or NON-JOINER controls *whether* a connection happens, while the alternate-connection mechanism controls *how* a connection happens. Thus, it would not be essential to have a control character that has a complementary function to ZWJ and ZWNJ. Such an option is a possibility, however: a third joiner that functions similar to ZWJ but that selects alternate connecting forms:

*Joiners:*

- *ZERO WIDTH NON-JOINER: inhibits cursive connection*
- *ZERO WIDTH JOINER: selects normal cursive-connecting forms*
- *ZERO WIDTH ALTERNATE JOINER: selects alternate cursive-connecting forms*

An important note with regard to the control required for Phags-pa is that an alternate connecting form is selected only for the character *after* the juncture.

# 11 Status of the letter A

*Status: open*

## *11.1 Description*

Section II.8 of N2871 raises concern regarding the status of the letter A, stating that some consider it to represent a vowel while others consider it a vowel, and that it has another role as *titum 'top of character'*. N2871 states on page 9,

> "Therefore, we think that Phags-pa script encoding… should adapt itself to all these views without being partial to one while restricting another. So it is inevitable to [provide encoded representations] in different ways, though if properly handled, confusion can be avoided."

## *11.2 Discussion*

It must be noted that, in general, no specific linguistic interpretation is imposed on any character in the UCS, and that encoding distinctions are not made in order to support different views among experts on the appropriate linguistic interpretation of characters. In accordance with the principles outlined in N2652R *Principles and Procedures*, characters are to be encoded in the UCS once, and that multiple encoded representations for the same text element are to be avoided.

Thus, any difference in view held by the authors of competing proposals on the status of A as a vowel or consonant have no bearing on encoding decisions to be made, with the possible exception of decisions regarding the binary ordering of encoded characters.

# 12 Encoding of A-vowel forms

*Status: open*

## 12.1 Description

N2869 refers to variant forms for the vowels in which the nominal glyph (or a connecting variant) of the vowel is preceded by the glyph for A. For example,  is considered a variant of  /u/.

This document will refer to these (for lack of an obvious alternative) as *A-vowel forms*. This terminology is not used in N2869. It is adopted here purely as a way to refer to these forms for sake of discussion. No particular analysis is implied thereby.

Related to the A-vowel forms is the appearance in section IV of N2870 of  as a variant of SUBJOINED WA. It is assumed that the same encoding mechanisms will be used in this case as for the vowels.

It appears that China and Mongolia intend that these be encoded as sequences using the vowel character with some control character, but it is not entirely clear what representations are permitted. Section II.4(c) of N2869 seems to indicate that ZWNJ is used in at least some cases. In section IV of N2870, however, such forms are shown to be encoded as VS sequences, or as sequences using ZWJ, or possibly without any control character, the presentation form being determined by word position. Comparison of the two documents appears to indicate that multiple spellings for a given form would be available in at least some contexts.

N2622 does not make any particular reference to these A-vowel forms. N2771 makes clear that, under that proposal, these forms would be considered simply sequences of < A, *vowel* >.

## 12.2 Discussion

The text elements in question are visually indistinct from sequences of the form < A, *vowel* >; e.g.,  would be visually indistinct from the sequence <  PHAGS-PA LETTER A,  PHAGS-PA LETTER U >. Thus, encoding these as *vowel + control* sequences would result in multiple spellings (i.e., multiple encoded representations). One of the design principles for the UCS is that such multiple spellings are to be avoided. (See Clause 2.3 and also Annex G of N2652R *Principles and Procedures* for related discussion of UCS encoding principles.) The concern of multiple spellings is augmented if a given A-vowel form can be represented by multiple control-character sequences in addition to a sequence < A, *vowel* >.

N2869 and supporting documents do not provide a clear rationale indicating why it would not be adequate to encode these text elements as sequences of the form < A, *vowel* >. Such a rationale would need to be provided before an encoding decision contrary to established UCS principles can be considered.

It is noted that the proposed representation of these text elements requires the use of variation-selector sequences, and that this complication does not arise if the text elements are encoded as < A, *vowel* > sequences.

# 13 Encoding of YA, SHA, HA, FA variants

*Status: open*

## *13.1 Description*

Both N2622 and N2869 refer to variant forms of YA, SHA, HA and FA that are attested in *Menggu Ziyun*. Both proposals agree on the need for a distinct encoded representation for these variant forms but differ with regard to the analysis of how these forms were used in *Menggu Ziyun* and with regard to the means of encoded representation to be used.

(This corresponds to a portion of comment T13 in China's ballot comments on PDAM1.)

N2622 claims that the variant forms are used in cases of phonemic contrasts from an earlier period that had been neutralised in Yuan Chinese, and notes that these forms are not used in known period texts apart from *Menggu Ziyun*. N2622 further notes the minor visual differences between these variant forms and the nominal forms, and warns of user confusion if the *Menggu Ziyun* variant forms are encoded as distinct characters. For these reasons, N2622 proposes that these be represented as VS sequences involving the characters corresponding to the normal forms of YA, SHA, HA and FA (e.g., < YA, VS1 >).

N2869 claims that the variant forms are used to represent contrastive phonemes in Chinese (contrasts not found in Mongolian) and on that basis proposes that these be represented as distinct characters, HAN YA, HAN SHA, HAN HA and HAN FHA.

## *13.2 Discussion*

It is noted that N2622 chooses to unify these variants with their corresponding nominal forms on the basis that there is no phonemic contrast. Yet there is no disagreement that there is a contrast of letterforms, and that these contrasting forms stand in a corresponding relationship to contrasting Han characters.

# 14 Encoding of vowels /ö/ and /ü/

*Status: open*

## *14.1 Description*

These vowels are written as compound (multi-graph) forms: ᠤᠺ (or ᠰᠤᠺ) for /ö/, and ᠤᠥ (or ᠰᠤᠥ) for /ü/.

N2622 does not make any particular reference to these vowels, though it is clear from N2719 that, under that proposal, these would be considered multi-graphs, encoding as character sequences, < EE, O > (or < A, EE, O >) and < EE, U > (or < A, EE, U >).

N2869 treats these as atomic, encoding them as distinct characters, PHAGS-PA LETTER OE and PHAGS-PA LETTER UE. The rational provided is that these represent distinct phonemes and are based on the Mongolian characters ᢎ OE and ᢏ UE.

N2871 elaborates on the rationale, stating that in past implementations significant difficulties in processing of Phags-pa text were encountered if these vowels were encoded as sequences. Details regarding the problems encountered are not provided, however.

N2871 clarifies that multiple, distinct representations are to be allowed for, stating that this permits different researchers to assume different interpretations of the letterforms,

(This corresponds to T7 in China's ballot comments on PDAM1.)

## *14.2 Discussion*

The text elements in question are visually indistinct from sequences < EE, O > and < EE, U >. Thus, encoding these as distinct characters would result in multiple spellings (i.e., multiple encoded representations). N2871 (p. 4) makes explicit reference to the availability of multiple spellings for these vowels.

While they may constitute graphemic units within written forms for Mongolian or other languages, such multi-part graphemes are not normally encoded in the UCS in order to avoid multiple encoded representations. (See Clause 2.3 and also Annex G of N2652R *Principles and Procedures* for related discussion of UCS encoding principles.)

A function of multi-graphs to represent distinct phonemes of a language is not considered sufficient grounds on which to encode a multi-graph as a distinct character. It is noted, in particular, that N2869 says,

> "We maintain that the nominal glyphs of the Phags-pa alphabet is a system with pronunciation as its content, and with their graphic symbols to distinguish themselves."

An assumption that character distinctions to be made in the UCS should be based on pronunciation distinctions would contradict a basic design principle of the UCS.

N2871 cites difficulties for text processing that arise if these vowels are encoded as sequences, and gives as an example Latin transliteration ('must never be "aeo",' etc.). Such considerations arise in all cases of multi-graphs proposed for addition to the UCS, and are generally not found to be adequate grounds for encoding. For instance, in the Latin-transliteration example, there is only real difficulty in transliterating occurrences of  if this text element is truly ambiguous – that is, it must be interpreted as /ü/ in some cases but as some other reading in other cases. Further details on the nature of text-processing difficulties experienced in past implementations should be provided.

It is noted that the proposal in N2869 requires several variant forms for OE and UE, including variation-selector sequences. These complexities, including the need for VS sequences, do not arise if characters are encoded based on letterforms, with OE and UE represented as character sequences.

It is also noted in passing that many Brahmi-derived scripts use multi-graph representations for vowel phonemes, but these normally have representations as character sequences in the UCS. (Khmer is exceptional in this regard.)

# 15 Other glyph variants

## *15.1 Description*

N2869 identifies other "free variant" glyphs for WA and E; a variant for SUBJOINED YA is identified in N2871, as is another rare variant for U.

| Nominal glyph | Variant glyph | Note |
|---|---|---|
|  |  | |
|  |  | |
|  |  | |
|  |  | Initial form only |

Proposed variation-selector sequences for each of these (or their connecting variants) are presented in N2870.

N2622 listed some of these glyph variants but did not discuss any explicit encoding for them. N2771 states that these distinctions are no more than aesthetic in nature and should be treated as font considerations.

(This corresponds in part to T13 in China's ballot comments on PDAM1.)

## *15.2 Discussion*

There is no disagreement that the glyphs in question are variants of the nominal glyphs as identified; the only issue is whether VS sequences are required so that these glyph variants can be given explicit encoded representation.

The contributions from China and Mongolia cite frequencies of occurrence and mention the co-occurrence of both WA variants within the same word in one particular document. They do not explain why these variants must be supported by VS sequences in plain text rather than being considered font variants (provided in different fonts, or supported in a single font using a font feature).

N2771 states that these should be treated as font considerations, but does not identify any reasons why VS sequences for these would be problematic.

# 16 Visual versus logical encoding order for CANDRABINDU

*Status: open*

## *16.1 Descriptions*

The character CANDRABINDU appears visually on the leading edge of a syllable cluster. Phonologically, it is used to indicate nasalization of the syllable nucleus, implying that its

*reading* or "*logical"* order is after the vowel component of the syllable. In other scripts of Brahmic origin, the candrabindu is typically encoded at the end of an orthographic-syllable cluster.

Section 6 of N2622 identifies the question of where the PHAGS-PA LETTER CANDRABINDU should occur within a syllable-cluster sequence, and proposes that it be encoded in its visual position, at the start of the syllable.

Consensus must be established on whether a logical- or visual-encoding model is to be used for CANDRABINDU as this will have significant bearing for Phags-pa implementations.

## *16.2 Discussion*

For certain text processes, such as sorting or transliteration, visual ordering for CANDRABINDU may result in slightly greater level of complexity for implementations, as this character would need to be processed as though it were in *reading* order rather than visual order.

For rendering implementations, logical ordering would result in a slightly greater level of complexity, though any implementation that supports a variety of scripts, or even the minority scripts of China and Mongolia (which would include New Tai Lue script), must already be able to deal with the glyph re-ordering that would be required.

N2622 rightly observes that logical ordering would have a usability impact:

> "It would be… inconvenient for the end user to encode the Candrabindu sign as the last character in a syllable unit (its logical position) and yet render it as the first glyph in a syllable unit (its visual position), as cursor movement, text selection and delete/backspace operations would be confusing."

In most Indic scripts, the difference between logical and visual order of candrabindu is minimized due to the visual orientation of a syllable cluster in relation to the line direction. For Phags-pa, however, the logical- versus visual-position difference is greater, comparable to the i-kaar in Devanagari.

N2622 goes on to suggest that input methods may be designed to permit users to enter candrabindu in its logical order, but have it inserted into the data in its visual order. This would require an advanced-capability input method – an *input method editor* – that provides a temporary processing buffer and user-interface to edit text prior to it being inserted into the data. This would create a significant obstacle to implementation, greater perhaps than the need for glyph re-ordering during rendering under a logical-order model. Thus, if it is important to users that logically-ordered entry of candrabindu be supported, a logical order may be preferable. It would seem, though, that visually-ordered keyboard entry and a visually-ordered encoding should be acceptable to users.

# 17 Character names for contested characters

*Status: open, dependent on resolution of prior issues*

## *17.1 Description*

For characters proposed in N2869 that have not already been accepted for encoding by WG2 (i.e. do not appear in PDAM2), in the event that there is consensus to recommend one or more of these for encoding, there will also be a need to reach agreement on the character names.

The following are the potential characters in question:

| Glyph | Name proposed in N2869 |
|---|---|
|  | PHAGS-PA LETTER OE |
|  | PHAGS-PA LETTER UE |
|  | PHAGS-PA LETTER 0A |
|  | PHAGS-PALETTER VOICELESS SHA |
|  | PHAGS-PA LETTER VOICED HA |
|  | PHAGS-PA LETTER ASPIRATED FA |
|  | PHAGS-PA JOINER |

(This corresponds in part to T12 in China's ballot comments on PDAM1.)

## *17.2 Discussion*

It is noted that Rule 1 of the character-naming guidelines in Annex L of N2652R *Principles and Procedures* specifies that only Latin capital letters A to Z, space and hyphen are to be used, with allowance for digits 0 to 9 only under special circumstances. The use of the digit zero in "0A", therefore, is a concern. It is also noted that potential confusion between "O" (capital O) and "0" (zero) could result from the use of "0A".

# 18 Ordering of characters

*Status: open; dependent on resolution of final character repertoire*

Once the final character repertoire is determined, the code-position order of characters must be established.

(This corresponds to T9 in China's ballot comments on PDAM1.)

In N2871, China and Mongolia indicate that they largely agree to the ordering in PDAM2. Apart from the additional character that they propose which are not in PDAM2, their revised proposal in N2869 differs from PDAM 2 in the follow respects:

- Tibetan retroflex consonants (TTA, TTHA, DDA, NNA) follow the core set of consonants and vowels.
- Non-Tibetan consonants QA, XA, FA and GGA and the subjoined consonants come after the core vowels.
- Non-core vowels follow the subjoined consonants.