L2/14-060

Title: **Proposed properties for incorrectly or insufficiently documented characters in the Unicode 7.0 repertoire**
Source: **Laurențiu Iancu (Microsoft Corporation) and Ken Whistler (SAP AG)**
Status: **Individual contribution**
Action: **For consideration by the Unicode Technical Committee**
Date: **2014-02-05**

1. Abstract

Unicode 7.0 adds over 2,800 new characters, spread across 23 new scripts and 32 new blocks, besides adding to several existing blocks. The new character repertoire is diverse enough to present challenges and carry the risk of classification errors or omissions when assigning properties.

Occasionally, the documents proposing the new characters may contain errors or may lack the full documentation needed for assigning properties in a correct and complete manner. This document samples several characters which were analyzed during the property definition work for Unicode 7.0 and determined to require property changes or additions compared to their proposal documents. The respective modifications were applied in the alpha versions of the UCD 7.0 files. They are being brought to the attention of the UTC for awareness and validation or further tuning before the publication of Unicode 7.0.

2. Notable property assignments in alpha UCD 7.0

The set of property assignments discussed in this document is partitioned into three groups:

1. Egregious errors that were fixed in the alpha UCD 7.0 files, shown here for awareness rather than recommended for discussion.
2. Safe assignments, when proposal documents did not include the respective properties, or adjustments applied to property values from proposal documents, which were determined to be inadequate during the property definition work. An example of adjustment is for consistency across sets of similar characters. These items are included for transparency and grouped separately in case they raise any concerns.
3. Debatable cases, either insufficiently documented or not documented at all in the original proposals. Those are brought to the UTC to analyze, discuss, and make resolutions for.

In the following subsections, the properties marked 'Proposed' are those from the respective proposal documents, and those marked 'Assigned' are those from the alpha UCD 7.0 files as of February 1st, 2014 (LineBreak-7.0.0d1.txt, PropList-7.0.0d23.txt, Scripts-7.0.0d25.txt, UnicodeData-7.0.0d17.txt, etc.).

2.1. Egregious errors

Khojki -aa

U+1122C KHOJKI VOWEL SIGN AA	Proposed gc=Mn, bc=NSM	Assigned gc=Mc, bc=L

Post-base dependent vowel sign similar to the Devanagari counterpart. Classified in error in the proposal [11-021].

Tirhuta gvang

U+ 114C5 TIRHUTA GVANG	Proposed gc=Mc (implying lb=CM)	Assigned gc=Lo, lb=AL

Described as a *gomukha* in Section 4.11 of [11-175R], i.e., a form of anusvara (modifier indicating nasalization) similar to the Vedic *gomukha*s U+1CE9–U+1CEC, which are treated structurally as letters.

Modi abbreviation sign

U+11643 MODI ABBREVIATION SIGN	Proposed lb=BA	Assigned lb=AL

Proposed as lb=BA in [11-212R2], corrected to lb=AL for consistency with existing Indic abbreviation signs.

2.2. Safe assignments or adjustments

Vertical lines used in Lithuanian dialectology

U+2E3D VERTICAL SIX DOTS	Proposed lb=AL	Assigned lb=BA
U+2E3E WIGGLY VERTICAL LINE	Proposed lb=AL	Assigned lb=BA

Indicate phrasal or breathing pauses. Proposed as lb=AL in [11-223] by analogy with U+2016 DOUBLE VERTICAL LINE; assigned lb=BA for consistency with U+205E FOUR VERTICAL DOTS and the prototypical vertical bar U+007C, which are lb=BA. The value lb=BA provides more line-breaking opportunities after, and fewer before, than lb=AL (in combination with other line-breaking classes). Also assigned Term=N for consistency with U+205E.

Manichaean punctuation

U+10AF0 MANICHAEAN PUNCTUATION STAR	Proposed lb=QU	Assigned lb=BA
U+10AF1 MANICHAEAN PUNCTUATION FLEURON	Proposed lb=QU	Assigned lb=BA
U+10AF2 MANICHAEAN PUNCTUATION DOUBLE DOT WITHIN DOT	Proposed lb=EX	Assigned lb=BA
U+10AF3 MANICHAEAN PUNCTUATION DOT WITHIN DOT	Proposed lb=EX	Assigned lb=BA
U+10AF4 MANICHAEAN PUNCTUATION DOT	Proposed lb=EX	Assigned lb=BA
U+10AF5 MANICHAEAN PUNCTUATION TWO DOTS	Proposed lb=QU	Assigned lb=BA

Based on a single citation in the proposal [11-123R] which describes these characters generically as *Interpunktion*, an assignment of lb=BA for undifferentiated archaic separator punctuation marks seems appropriate, in the absence of more precise information.

Psalter Pahlavi punctuation

U+10B99 PSALTER PAHLAVI SECTION MARK	Proposed lb=B2	Assigned lb=AL
U+10B9A PSALTER PAHLAVI TURNED SECTION MARK	Proposed lb=B2	Assigned lb=AL
U+10B9B PSALTER PAHLAVI FOUR DOTS WITH CROSS	Proposed lb=B2	Assigned lb=AL

U+10B9C Psalter Pahlavi Four Dots with Dot Proposed lb=EX Assigned lb=AL

The property value lb=B2 was designed specifically for em dashes in Western typography and is not applicable to archaic punctuation and behavior inferred from fragmentary manuscript material. Assigned lb=AL following the model of U+0700 Syriac End of Paragraph from the historically related Syriac punctuation mentioned in the proposal [11-147].

Mahajani section mark

U+11175 Mahajani Section Mark lb N/A Assigned lb=BB

Assigned lb=BB similarly to Tibetan and 'Phags-pa head marks, given the explanation and examples in [11-274].

Sharada sutra mark

U+111CD Sharada Sutra Mark Proposed lb=BB Assigned lb=AL, Term=Y, STerm=Y

Proposed as lb=BB but described as being "used for indicating the end of a *sūtra*, or *rule*" in [12-171R]. Assigned lb=AL because it can appear flanked by double dandas and by spaces, to avoid forcing any particular line breaking. Also, per UAX #29 rule SB8a, STerm × STerm, so <double-danda space* sutra space* double-danda> has a single sentence boundary at the end.

Khojki word separator

U+1123A Khojki Word Separator lb N/A Assigned lb=AL, Term=Y, STerm=N

Shown as a trailing separator in the examples in [11-021], with no real apparent function in breaking lines, and commonly occurring next to double (and, occasionally, single) dandas. Defaulted to lb=AL to avoid a redundant line-breaking opportunity.

Grantha sign pluta

U+1135D Grantha Sign Pluta lb N/A Assigned lb=AL

Essentially a letter, and described as "break not allowed before pluta" in [10-331].

Pau Cin Hau sentence-final tones

U+11AE8 Pau Cin Hau Rising Tone Long Final STerm N/A Assigned Term=N (hence STerm=N)
… … …
U+11AF8 Pau Cin Hau Glottal Stop Final STerm N/A Assigned Term=N (hence STerm=N)

Ten of the Pau Cin Hau tone letters have a conflated function as sentence-final punctuation [11-104R]. However, tailored sentence segmentation for Pau Cin Hau is out of scope for the default UAX #29 algorithm.

Bassa Vah Full Stop

U+16AF5 Bassa Vah Full Stop N/A Assigned lb=BA, Term=Y, STerm=Y

Properties of U+16AF5 not discussed in [10-382R]. Of two candidate models, U+A60E Vai Full Stop which is lb=EX (no indirect break before) and U+A6F3 Bamum Full Stop which is lb=BA (indirect break before), followed the lb=BA model in the absence of strong evidence for preventing indirect breaks before.

Pahawh Hmong punctuation

U+16B37 PAHAWH HMONG SIGN VOS THOM	Proposed lb=EX	Assigned lb=BA, Term=Y, STerm=Y
U+16B38 PAHAWH HMONG SIGN TSHAB CEEB	Proposed lb=EX	Assigned lb=BA, Term=Y, STerm=Y
U+16B39 PAHAWH HMONG SIGN CIM CHEEM	Proposed lb=IS	Assigned lb=BA, Term=Y, STerm=N

Described in [12-013] as behaving like punctuation marks '?', '!', and ','. However, that functional relationship does not require identical line-breaking behavior, and given the terse evidence, assigned lb=BA (and corresponding terminal-punctuation properties).

Pahawh Hmong arithmetic symbols

U+16B3C PAHAWH HMONG SIGN XYEEM NTXIV	OMath N/A	Assigned OMath=N
…	…	…
U+16B3F PAHAWH HMONG SIGN XYEEM FAIB	OMath N/A	Assigned OMath=N

Described in [12-013] as arithmetic symbols behaving like '+', '–', '×', and '÷'. Assigned Math=N (as a result of OMath=N) because Math=Y is primarily associated with international mathematical symbols.

Duployan thick letter selector

U+1BC9D DUPLOYAN THICK LETTER SELECTOR	Proposed gc=Cf	Assigned gc=Mn

From the description in [11-303], it is functionally equivalent with a variation selector. Assigned properties as for all other variation selectors.

Wingdings and webdings

Multiple code points	Proposed lb=AL	Assigned lb=AI, AL, ID, NS, or QU

Examples include:

U+1F10B DINGBAT CIRCLED SANS-SERIF DIGIT ZERO	Assigned lb=AI, ea=N (as the sets at U+2780 etc.)
U+1F322 BLACK DROPLET	Assigned lb=ID (as most weather symbols)
U+1F336 HOT PEPPER	Assigned lb=ID (as other plant symbols)
U+1F679 HEAVY INTERROBANG ORNAMENT	Assigned lb=NS (consistently with U+203D INTERROBANG)
U+1F676 SANS-SERIF HEAVY DOUBLE TURNED COMMA QUOTATION MARK ORNAMENT	Assigned lb=QU

Document [11-344] proposed that all symbols in this set be given lb=AL. However, a blanket lb=AL would be inconsistent with many existing pictographic symbols. Appropriate lb values were assigned consistently with similar characters. A few examples are shown above.

Sample Script property assignments

Notable characters in terms of sc property include the following:

U+0605 ARABIC NUMBER MARK ABOVE	Assigned sc=Zyyy (decided by the UTC)
U+AB5B MODIFIER BREVE WITH INVERTED BREVE	Assigned sc=Zyyy (as similar phonetic modifiers)
U+AB65 GREEK LETTER SMALL CAPITAL OMEGA	Assigned sc=Grek (as similar phonetic letters)
U+1BCA0 SHORTHAND FORMAT LETTER OVERLAP	Assigned sc=Zinh (like ZWJ, ZWNJ)
…	…
U+1BCA3 SHORTHAND FORMAT UP STEP	Assigned sc=Zinh (like ZWJ, ZWNJ)
U+102E0 COPTIC EPACT THOUSANDS MARK	Assigned sc=Zyyy (decided by the UTC)
…	…
U+102FB COPTIC EPACT NUMBER NINE HUNDRED	Assigned sc=Zyyy (decided by the UTC)

2.3. Debatable cases

Siddham separators

U+115C4 SIDDHAM SEPARATOR DOT	Proposed lb=BA	Assigned lb=EX, Term=Y, STerm=N
U+115C5 SIDDHAM SEPARATOR BAR	Proposed lb=BA	Assigned lb=EX, Term=Y, STerm=N

Described as marking "boundaries between syllables, words, and phrases" [12-234R]. From the samples in proposal, the separators do not seem to have a predisposition to occur at the ends of lines like dandas. Assigned lb=EX by analogy with Tibetan *shad*s to prevent indirect line breaks and distinguish them from the dandas also occurring in Siddham.

3. References

[10-331] Shriramana Sharma, *Request to encode 1135D GRANTHA SIGN PLUTA*, L2/10-331, August 2010, http://www.unicode.org/L2/L2010/10331-grantha-pluta-sign.pdf.

[10-382R] Michael Everson and Charles Riley, *Final proposal for encoding the Bassa Vah script in the SMP of the UCS*, L2/10-382R, October 2010, http://www.unicode.org/L2/L2010/10382r-n3941r-bassavah.pdf.

[11-021] Anshuman Pandey, *Final Proposal to Encode the Khojki Script in ISO/IEC 10646*, L2/11-021, January 2011, http://www.unicode.org/L2/L2011/11021-khojki.pdf.

[11-104R] Anshuman Pandey, *Proposal to Encode the Pau Cin Hau Alphabet in ISO/IEC 10646*, L2/11-104R, April 2011, http://www.unicode.org/L2/L2011/11104r-paucinhau-alphabet.pdf.

[11-123R] Michael Everson, Desmond Durkin-Meisterernst, Roozbeh Pournader, and Shervin Afshar, *Second revised proposal for encoding the Manichaean script in the SMP of the UCS*, L2/11-123R, May 2011, http://www.unicode.org/L2/L2011/11123r-n4029r-manichaean.pdf.

[11-147] Michael Everson and Roozbeh Pournader, *Proposal for encoding the Psalter Pahlavi script in the SMP of the UCS*, L2/11-147, May 2011, http://www.unicode.org/L2/L2011/11147-n4040-psalter-pahlavi.pdf.

[11-175R] Anshuman Pandey, *Proposal to Encode the Tirhuta Script in ISO/IEC 10646*, L2/11-175R, May 2011, http://www.unicode.org/L2/L2011/11175r-tirhuta.pdf.

[11-212R2] Anshuman Pandey, *Proposal to Encode the Modi Script in ISO/IEC 10646*, L2/11-212R2, November 2011, http://www.unicode.org/L2/L2011/11212r2-n4034-modi.pdf.

[11-223] Vladas Tumasonis and Karl Pentzlin, *Second revised proposal to add characters used in Lithuanian dialectology to the UCS*, L2/11-223, May 2011, http://www.unicode.org/L2/L2011/11223-n4070.pdf.

[11-274] Anshuman Pandey, *Proposal to Encode the Mahajani Script in ISO/IEC 10646*, L2/11-274, July 2011, http://www.unicode.org/L2/L2011/11274-n4126-mahajani.pdf.

[11-303] Van Anderson, *Proposal to include Duployan Shorthands and Chinook script and Shorthand Format Controls in UCS, as approved by WG2*, L2/11-303, July 2011, http://www.unicode.org/L2/L2011/11303-duployan.pdf.

[11-344] Michel Suignard, *Updated proposal to add Wingdings and Webdings Symbols*, L2/11-344, September 2011, http://www.unicode.org/L2/L2011/11344-wingdings.pdf.

[12-013] Michael Everson, *Final proposal to encode the Pahawh Hmong script in the UCS*, L2/12-013, January 2012, http://www.unicode.org/L2/L2012/12013-n4175-pahawh-hmong.pdf.

[12-171R] Anshuman Pandey, *Proposal to Encode the Sutra Mark for Sharada*, L2/12-171R, May 2012, http://www.unicode.org/L2/L2012/12171r-sharada-sutra-mark.pdf.

[12-234R] Anshuman Pandey, *Proposal to Encode the Siddham Script in ISO/IEC 10646*, L2/12-234R, August 2012, http://www.unicode.org/L2/L2012/12234r-n4294-siddham.pdf.