TO: UTC **L2/15-045**
FROM: Deborah Anderson, Ken Whistler, Rick McGowan, Roozbeh Pournader, and Andrew Glass
SUBJECT: Recommendations to UTC #142 February 2015 on Script Proposals
DATE: 30 January 2015

The recommendations below are based on documents available to the members of this group at the time they met, and do not include documents submitted later to the document registry.

SOUTH ASIA
Indic
1. *Tamil*
Documents:
L2/15-003 Naming Tamil Symbols in SMP - Vallinam Characters – Ganesan

Older docs:
L2/14-210 Letter on Tamil Fraction Naming - Tamil Virtual Academy
L2/14-212 Tamil names and annotations - Vasu Renganathan / INFITT
L2/14-215 A Proposal as a Standardised Romanisation Scheme for Full Tamilwords Used Inside Code Pages as in Naming of Various Characters Etc & In CLDR – Logasundaram
L2/14-216 Current status of Tamil symbols naming issue (W2 N4622) - Sharma

Discussion: We reviewed the documents, which discuss how Tamil character names and annotations should appear in Roman transliteration. The set of names and annotations under discussion are included in the Tamil Supplement block currently in PDAM 2.2, though two aliases in the current Tamil U+0B80 block are also affected.

Recommendations: We recommend the UTC take no action until expected feedback is received, and then make a decision. (Because Amendment 2 will go out for another PDAM ballot, there will be an opportunity to make ballot comments on the names later, if needed.)

2. *Grantha*
Documents:
L2/14-291 Handling variation in vowelless consonant forms in Grantha – Sharma

Other docs:
L2/14-020 Plain-text ligating virama representation for Grantha script – Ganesan L2/14-097 ZWJ Joiner for Chillu Consonants of Grantha Script – Ganesan
L2/14-110 Comments on L2/14-097 re using ZWJ for Grantha "chillus" - Sharma
L2/14-162 Control Characters (Joiners ZWNJ and ZWJ) in the Grantha Visible Virama and Chillu Consonants - Ganesan
L2/14-164 Chillu examples – Ganesan
L2/14-279 ZWJ for Grantha pre-pausal half-consonants (chillus) - Ganesan
L2/14-268 Recommendations to UTC #141 October 2014 on Script Proposals - Anderson et al.

Discussion: We reviewed these documents, which basically impacts the wording in the 8.0 Core Specification for Grantha.

At the October 2014 UTC, an AI was created to add text for 8.0 saying that Consonant + Virama + ZWJ was an acceptable mechanism to show visual distinctions of the fused virama forms in Grantha. In §2.1 of L2/14-291, Sharma points out that Consonant + Virama + ZWJ is used for C1-conjoining forms in Indic. In Grantha, this sequence is only used for requesting the *reph* in isolation, since all other consonants have C2-conjoining forms. The other South Indian scripts (like Telugu and Kannada) follow this same pattern, so Grantha should also follow the same model, Sharma argues. As mentioned by Sharma, the fused virama forms are not conjoining forms, since they freely alternate with other vowelless forms. It was noted during the script ad hoc's discussion that the use of Consonant + Virama + ZWJ for C1-conjoining forms was the old Malayalam model--Unicode 5.0 and prior, before encoding of the chillus--and that model is now deprecated.

Based on the argument in §2.1 of L2/14-291, we now suggest no change be made to the current text. (Note that Sharma's option §3 is not possible, since Variation Selectors cannot be used with viramas [*TUS* §23.4 "The base character in a variation sequence is never a combining character…"].)

Recommendations: Given the new analysis provided in L2/14-291, we recommend the UTC discuss the topic briefly and drop the action from the last meeting (AI 141-A20 for Roozbeh Pournader: add text to the Specification reflecting the recommendation in L2/14-268 for the use of chillus and ZWJ for version 8).

3. *Malayalam*

a) Malayalam Anusvara Above

Documents:

L2/14-292 Clarifications re the proposed 0D00 MALAYALAM SIGN COMBINING ANUSVARA ABOVE – Sharma

L2/15-011 Comment on L2/14-003: Any need for Combining Anusvara Sign Above in Malayalam Block? – Ganesan

Older docs:

L2/14-003 Proposal to encode 0D00 MALAYALAM SIGN COMBINING ANUSVARA ABOVE – Sharma

L2/14-029 Feedback on Malayalam Anusvara Above Proposal – Cibu

L2/14-069 Evidence for considerable usage of the Malayalam anusvara above – Sharma

Discussion: We reviewed the documents. The examples on page 3 and 5 of proposal (L2/14-003) show clear examples of the character, and L2/14-292 sufficiently answered questions posed in the earlier script ad hoc's recommendations (L2/14-268).

Recommendations: We recommend the UTC approve U+0D00 MALAYALAM SIGN COMBINING ANUSVARA.

b) Malayalam Ordinal Indicator

Document:

L2/14-303 Malayalam Ordinal Indicator – Davidsson

Discussion: We reviewed this document, which requests a new character *Malayalam ordinal indicator*. According to the author, use of two existing combining characters to represent the *Malayalam ordinal indicator*, U+0D3E MALAYALAM SIGN AA and U+0D02 MALAYALAM SIGN ANUSVARA, does not display

correctly on some browsers or word processing software. The example cited in L2/14-303 was:
135-ↄo vs. 135-◌Ↄ◌o.
It appears to the script ad hoc that there are rendering bugs in displaying the sequence of two combining character after a dash, so the problem lies with the rendering engines, browsers, and/or fonts.

Recommendations: We recommend the UTC respond to the author, recommending the *Malayalam ordinal indicator* be represented by the sequence <U+0D3E U+0D02>, and refer him to the "Anusvara" section of the Malayalam block intro on p. 489 of 7.0, which states:

> ***Anusvara***. *The anusvara can be seen multiple times after vowels, whether independent letters or dependent vowel signs, as in ….<0D08, 0D02, 0D02, 0D02, 0D02>. Vowel signs can also be seen after digits, as in … <0033, 0035, 0035, 0D3E, 0D02>. More generally, rendering engines should be prepared to handle Malayalam letters (including vowel letters), digits (both European and Malayalam), dashes, U+00A0 NO-BREAK SPACE and U+25CC DOTTED CIRCLE as base characters for the Malayalam vowel signs, U+0D4D MALAYALAM SIGN VIRAMA, U+0D02 MALAYALAM SIGN ANUSVARA, and U+0D03 MALAYALAM SIGN VISARGA. They should also be prepared to handle multiple combining marks on those bases.*

The UTC may also suggest the proposal author file bugs with those companies with browsers or software products that are not displaying the sequence correctly.

c) Vertical Bar Virama
Documents:
L2/15-021 Comment on L2/14-015R – Ganesan
L2/14-015R Proposal to encode MALAYALAM SIGN VERTICAL BAR VIRAMA – Cibu (updated 15 January 2015)

Discussion: We reviewed the documents. The updated proposal, revised in January 2015, now includes character properties. On the top of page 6, there is a clear contrast between CANDRAKKALA (U+0D4D MALAYALAM SIGN VIRAMA) and the VERTICAL BAR VIRAMA. The proposed properties appear correct. The discussion in the proposal about VERTICAL BAR VIRAMA and chillus seems valid, i.e., that a separate character for the VERTICAL BAR VIRAMA is warranted, given that there are instances where the VERTICAL BAR VIRAMA does not ligate or only sometimes ligates.

Recommendations: We recommend the UTC encode MALAYALAM SIGN VERTICAL BAR VIRAMA, with properties as given in updated proposal. The proposed code point is U+0D3B.

d) Circular Virama Sign
Documents:
L2/14-014R Proposal to encode MALAYALAM SIGN CIRCULAR VIRAMA – Cibu, Siju, Sunil (revised 15 January 2015)
L2/15-024 Comment on Malayalam circular Virama sign (L2/14-014r) – Ganesan

Discussion: We reviewed the documents. The proposal shows contrastive use of the circular virama and the CANDRAKKALA (U+0D4D MALAYALAM SIGN VIRAMA) in images 4, 6, 7 and 8.

The script ad hoc discussion focused on the properties. The CANDRAKKALA,  , is Mn with Indic_Positional_Category=Top. The CIRCULAR VIRAMA holds the same positional slot, so its general category should, we believe, also be Mn.

Recommendations: We recommend the UTC encode MALAYALAM SIGN CIRCULAR VIRAMA, but the general category should be Mn and bidi NSM. The proposed code point is U+0D3C.

The recommendations regarding the properties for this Malayalam sign suggest that the UTC may wish to reconsider the gc property for the Takri virama, which has similar rendering issues.

 U+116B6 TAKRI SIGN is currently gc=Mc, but has Indic_Positional_Category=Top. It should, perhaps, be changed to gc=Mn.

4. Bengali

Documents:

L2/14-304 Bengali Vowel Letter Aw (U+0985 U+09D7) Used in Kokborok – Sanghmitra Sahu

L2/15-010 Feedback on proposal L2/14-304 to encode BENGALI LETTER AW – Sharma

Discussion: We reviewed these documents. The document L2/14-304 requests the encoding of a vowel letter in the Bengali script for AW that is used in the Kokborok language of India. Users are currently representing AW by the sequence U+0985 BENGALI LETTER A followed by U+09D7 BENGALI AU LENGTH MARK. However, the proposal author reports of difficulties in getting this combination to work properly, and mentions it is not "standard practice" nor universally supported. (Examples of AW are found on lines 2-3 of the sample text.)

Discussion about the proposal in the script ad hoc raised the following points:

- Although the sequence <0985, 09D7> may not work on older operating systems, this combination should work on modern OSes.
- For most Indic scripts, the Core Specification often includes a table that recommends use of the atomically encoded vowel letter, and not the vowel letter followed by dependent vowel sign. However, in this case, encoding a new vowel letter could introduce a new ambiguity that the language does not currently have.

Recommendations: We recommend the UTC not encode this character, but instead have the Editorial Committee add text to Bengali section of 8.0 that recognizes this form exists, recommend the form be represented by the sequence <U+0985, U+09D7> and note in the text that it is an exception to the rule that independent vowels are separately encoded.

5. Gondi

Document:

L2/15-005 Proposal to encode the Gondi script – Pandey

Discussion: We reviewed this proposal, which is a revision of the 2012 proposal. The discussion touched on a several points:

- Reword text in section 3.7 "The Gondi VIRAMA is identical in shape and function to the corresponding character … U+094D DEVANAGARI SIGN VIRAMA. It is rendered by default as a

visible sign. “
(The VIRAMA is not visible in the sequence KA + VIRAMA.)

- Explain figure 22, which shows the visible virama. What is this document about and what is it saying?
- Discuss the encoding sequence options for *repha* and *ra-kāra* (which have implications for the user’s anticipated typing sequence):
  (a) for *ra-kāra* (2 options): Consonant + *ra-kāra* + Vowel or Consonant + Vowel + *ra-kāra*
  (b) for *repha* (3 options): *repha*+ Consonant + Vowel or Consonant + *repha* + Vowel or Consonant + Vowel + *repha*
- The present model of using ZWJs and ZWNJs for presenting various forms of virama and *ra* is confusing. We suggest an alternative model be considered, with two different virama characters, an always invisible character, and an always visible one, similar to Myanmar and Khmer. For *repha* and *ra-kāra*, encoding separate characters may be useful, similar to the Malayalam *dot reph* and and the *medial ra* in Myanmar and Tai Tham.

Recommendations: We recommend the UTC review the proposal and forward feedback from the discussion to the author, including the questions raised above.

CENTRAL ASIA
6. Khotanese
Document:
L2/15-022 Preliminary Proposal to Encode the Khotanese Script – Wilson

Older document:
L2/14-192 Preliminary Proposal to Encode the Turkestani Script – Wilson

Discussion: We reviewed this preliminary proposal, which appears to be a solid start. The proposal was based on the earlier Turkestani proposal (L2/14-192), which laid out solid reasons for disunifying “Turkestani” in §5 (of L2/14-192) into two scripts, Khotanese and Tocharian.

A few comments:

- It might make sense to move the code points into a different alignment, thus making the chart easier to use. The suggested change is to put:
  - independent vowels in the first column, starting with LETTER A at U+11E60
  - consonants in the second and third columns
  - dependent vowels in the fourth column in the same relative offsets as the independent vowels, for easier comparison in the chart
  - other characters in the last two columns
- §4.9 Numbers
  - The proposal recommends use of a virama for representing combinations of numbers. While the original Brahmi proposal did propose a virama for the old Brahmi additive-multiplicative number system, the model was changed in Unicode 7.0 to one that uses a U+ 1107F BRAHMI NUMBER JOINER. For Khotanese, should a script-specific number joiner be encoded or the BRAHMI NUMBER JOINER? While a script-specific character may be the cleaner option, the next revision of the proposal should discuss the different options.

    - Include discussion on how one would write "102" and "200" in Khotanese. (If the 2 is on the right side of 100 in one and the left side of 100 in the other, we don't need any special mechanism, we can just suggest mandatory ligatures.)
- Indic properties (i.e., Indic_Positional_Category and Indic_Syllabic_Category) should be added.
- The proposed location of the script, U+11E60...11EBF, should be discussed with the Roadmap Committee, since Chola is currently in that location.
- The author should seek input from experts in Khotanese.

Recommendations: We recommend the UTC members review this preliminary proposal and send feedback to the author, including the comments above.

7. Tocharian
Document:
L2/15-023 Preliminary Proposal to Encode the Tocharian Script – Wilson

Older document:
L2/14-192 Preliminary Proposal to Encode the Turkestani Script – Wilson

Discussion: Since it is preliminary, we recommend Tocharian experts and seasoned Unicode proposal authors read it carefully. The Roadmap Committee should be consulted, as the proposed location, U+11E00..11E6F is taking the spot allocated for Chalukya and spills over into the proposed location for Khotanese (U+11E60-). (See also comments above, under "Khotanese", for background.)

Recommendations: We recommend the UTC members review this preliminary proposal and send feedback to the author, including the comments above.

8. Soyombo
L2/15-004 Proposal to Encode the Soyombo Script – Pandey
L2/15-009 N4653 Comments on Proposals of Zanabazar Square and Soyombo Script from Mongolian Experts – Toshiya Suzuki, et al

Discussion: We reviewed the documents. This revised proposal has adopted the virama model, unlike the earlier version of the proposal (L2/13-069), which had a model with subjoined forms for each consonant. The latest proposal encodes a virama character (called "SOYOMBO SUBJOINER") used to stack consonants, four cluster-initial letters that behave like Malayalam *dot reph* (and always cluster with a following letter), and additional head marks and terminal marks. A new category was added for the cluster-initial letters ("Consonant_Prefixed").

It was noted that the script has a font and working implementation, which was demonstrated by John Hudson and Andrew Glass at the 2014 Unicode Conference.

This proposal was deemed very mature.

Comments from the script ad hoc:
- How are the three head marks different from one another?
- Based on feedback from the Mongolian experts (L2/15-009) and the Japan NB, engagement with the experts in Mongolia is needed.

Recommendations: We recommend the UTC encode the script. We also suggest the UTC convey to the Japanese and Mongolian National Bodies that once the script is on a ballot, there will be ample opportunity to make comments. (The earliest opportunity for Soyombo to be put on a ballot would be after the WG2 and SC2 meetings in October 2015. According to this schedule, Soyombo could conceivably be in line to be published as part of Unicode 10 in 2017.)

EUROPE

9. North Eastern Iberian

Document:

L2/15-012 Preliminary proposal to encode the north-eastern Iberian script – University of Barcelona

Discussion: We reviewed this preliminary proposal, which is for an ancient script attested in over 2000 inscriptions, dating from 5 c BC– 1 c AD. The proposal is part of the ongoing Hesperia project, which is making available a database with critical editions of all the Palaeohispanic inscriptions.

Recommendations: We recommend the UTC members review this proposal and send comments to the authors.

PUNCTUATION

10. Punctuation

Document:

L2/14-302 Short hyphen proposal – Moore

Discussion: We discussed this proposal, which asks for a new hyphen to be encoded, but one with left-to-right directionality. The author, creator of the artificial language Peoplese, currently uses the character U+06D4 ARABIC FULL STOP, which has strong right-to-left directionality (Bidi_Class=AL). The character is used as a "hyphenette" after prefixes, which contrasts with a standard hyphen (presumably U+002D HYPHEN-MINUS).

Recommendations: We recommend the UTC not accept this proposal, but send the author feedback saying that there are other eligible encoded characters available to fulfill his needs, such as U+2010 HYPHEN (Bidi_Class=ON) or U+2043 HYPHEN BULLET (Bidi_Class=ON).

NOTATIONAL SYSTEMS

11. Pitman Shorthand

Document:

L2/14-254 Encoding Pitman Shorthand scripts into Unicode Character Set – Rajaram

Discussion: We reviewed this proposal, a revised version of L2/14-167 which incorporates changes made based on feedback from the August 2014 UTC. The revised proposal now includes two punctuation characters and a chart of the 43 proposed characters (pp. 9-11). It proposes removing 47 characters from the earlier proposal (identified on pages 3-7 by red type). A list of 34 points for "clarification or confirmation" are listed on pages 12-14.

A few questions were raised:

- Does the proposed model follow that of Duployan?

- Can the author provide examples of running text alongside the representation using the proposed code points? (If such information could be provided, it might be possible to see if Pitman Shorthand would work in current rendering engines. The earlier version of the proposal contained single word examples on pp. 9-13, but the examples were not included in the later revision and were not full sentences.)
- Additional feedback on this proposal from those familiar with encoding a shorthand script (such as Van Anderson) and experts would be useful, particularly to provide input on the list of 34 points for clarification/confirmation on pages 12-14.

Recommendations: We recommend the UTC carefully review this proposal and discuss it, and relay to the author comments and questions that arise from the discussion, including the feedback above.

Previous recommendations (carried over for script and character proposals not yet discussed in the UTC)

EAST AND CENTRAL ASIA

12. Small Seal Script

L2/14-242 Proposal to encode Small Seal Script – TCA and China

Discussion: We reviewed this proposal, which proposes 799 characters out of a projected 10,516. In our opinion, the proposal is still far from mature, and would benefit from coordinating work with experts in the U.S. and Japan in order to formalize mapping data, which is needed to evaluate a final proposal. The proposal should also provide demonstrated need for including the script in the international standard.

Recommendations: We recommend the UTC members review this proposal and consider sending the authors the comments above.

13. Naxi Dongba

L2/14-241 Supplement on Proposal for Encoding Naxi Dongba Pictograph Script (L2/11-178) - China
L2/14-245 Feedback on Naxi Dongba Supplement document - Anderson

Discussion: We reviewed the "Supplement" document, which answered questions posed at the June 2011 WG2 meeting in Helsinki, Finland (see Naxi Dongba Ad Hoc report, L2/11-244). Specifically, the authors in the "Supplement" confirmed that the encoding is for modern use, not traditional use of the characters, and that alphabetical ordering is preferred.

The "Feedback" document posed additional questions and made suggestions. During WG2 discussion, the Naxi Dongba proposal authors stated the script is both a logography and syllabary, and the variation shown in some glyphs is due to regional differences, but only one glyph per character is warranted in the encoding. They agreed to revise the proposal and provide information on the proposed characters, with glyphs, Romanized transcription, Chinese gloss (and English translation) and references.

Recommendations: We recommend the UTC members review this proposal and send comments to the authors.

14. Shuishu

L2/14-243 Proposal for encoding Shuishu – China

Discussion: We reviewed this proposal, which is still at an early stage. In our view, it is not yet clear that Shuishu is an encodable writing system. In order to move forward, we recommend the authors prepare and publish a standard sign list for Shuishu, which can then be circulated for review by other scholars and gain scholarly support. The next version of the proposal should also provide a rationale for the digital representation of their sign list, answering the question why these shapes should be put into an international character encoding standard.

Recommendations: We recommend the UTC members review this proposal and send comments to the authors. The UTC may want to relay the suggestions to the authors above, regarding recommended next steps.

15. Khitan Large Script
L2/14-234 Proposal on Encoding Khitan Large Script – China
L2/14-233 Preliminary Review of Proposal on Encoding Khitan Large Script – West
L2/14-246 Ad hoc reports for Tangut and Khitan Large Script – Anderson

Discussion: We reviewed these documents. As noted in L2/14-233, the Khitan Large Script is largely undeciphered without any character list or recent dictionaries, vocabulary lists, or secondary linguistic materials, so the current proposal should be viewed as preliminary.

Also as mentioned in L2/14-233, the script appears to have a significant percentage of characters (18%) that are either Han borrowings or identical in shape to already encoded CJK ideographs. A revised proposal should discuss the pros and cons of unifying those Khitan Large Script characters with CJK characters already encoded: what are the costs/benefits to unification? Because Khitan Large Script is an historical script, the security risk would not arise if Khitan Large Script used CJK characters, only if it encoded a large set of identical CJK characters.

Additionally, we suggest the proposal also create a “Uni-Khitan” database (or spreadsheet) to document sources.

Recommendations: We recommend the UTC members discuss these documents.

16. Ranjana
L2/09-192 Preliminary proposal for encoding the Rañjana script in the SMP (WG2 N3649)
L2/14-221 Comparison between Ranjana Proposals - Anderson
L2/13-243 Proposal to Encode Ranjana Script - Manandhar
L2/14-253 Recommendations to UTC from Script Meeting in Nepal - Anderson

Discussion: We discussed these documents. Since decisions on the repertoire and encoding model for Ranjana depend upon those for “Nepaalalipi”, discussion on Ranjana was limited. It was noted that a future Ranjana proposal should also discuss the unification with Wartu and Lanydza, and should provide details on any specific characters and behaviors of the script in Tibet and other locations outside Nepal.

Recommendations: We recommend the UTC review the document, but postpone discussion until after the “Nepaalalipi” encoding is resolved.

17. Bhujinmola
L2/14-253 Recommendations to UTC from Script Meeting in Nepal

L2/14-283 Introducing the Bhujinmol Script - Pandey

Discussion: We briefly discussed the section in the "Recommendations" on Bhujinmola. Bhujinmola has a characteristic wavy headline (see examples in "Roadmapping the Scripts of Nepal" L2/09-325). The question on whether Bhujinmola represents a stylistic variation of "Nepaalalipi" or should be separately encoded needs to be discussed in a separate document, with examples of how vowels and consonants join differently from "Nepaalalipi" and other rendering issues.

Recommendations: We recommend the UTC review the document, but wait for further research to support separately encoding Bhujinmola. (Note: The script ad hoc did not yet review L2/14-283 Introducing the Bhujinmol Script by Pandey.)