Document Number: L2/01-430R

Title: UTC Response to L2/01-304, “Feedback on Unicode Standard 3.0”, an article published in Vishwabhara@tdil (Newsletter of the TDIL Programme of Ministry of Information Technology, Government of India).

URLs: `http://vishwabharat.tdil.gov.in/newsletter1.htm`
`http://www.unicode.org/L2/L2001/01304-feedback.pdf`

Source: Rick McGowan on behalf of Unicode Technical Committee

Date: November 20, 2001

ABSTRACT

The document “Feedback on Unicode Standard 3.0”, an article published in Vishwabhara@tdil (Newsletter of the TDIL Programme of Ministry of Information Technology, Government of India) asks for some additional characters, provides annotations and information for block introductions, and also requests a number of codepoint changes. The present document is an initial technical response by the Unicode Technical Committee to the points raised in that document.

UTC has recently approved several characters in the attached document L2-01/431R.

INTRODUCTION

UTC would like to thank the authors of “Feedback on Unicode Standard 3.0” for writing this detailed analysis of Indic script encoding within the Unicode standard.

“Feedback on Unicode Standard 3.0” received a UTC document number of “L2/01-304” and will may be referred to by that number in the discussion below. This response document and accompanying documents form a detailed analysis of all the requests and suggestions made in L2/01-304, with suggested actions for UTC and questions for the authors of L2/01-304.

This document is an initial technical response, and the position of the UTC on specific points may change in view of additional information from the Government of India on particular characters. The committee looks forward to discussion of the various points raised by the document, so that understanding and agreement can be reached about specific resolutions.

To expedite the addition of characters and annotations to the standard, UTC would like to engage in direct dialogue and meetings between the experts listed as contributors to L2/01-304 and various UTC members who would be working on technical details of the proposals.

In the following discussion, sections of the document L2/01-304 will be referred to by page number, beginning with "Page 15", as well as sub-heading, such as "Bengali".

Three documents available on the Unicode web site are also relevant to various parts of this discussion:

1. The Unicode FAQ "Indic Scripts and Languages":
`http://www.unicode.org/unicode/faq/indic.html`

2. Unicode Technical Report #15 "Unicode Normalization Forms" available on the Unicode web site:
`http://www.unicode.org/unicode/reports/tr15/`

3. The joint report with W3C, "Unicode in XML and other Markup Languages":
`http://www.unicode.org/unicode/reports/tr20/`

PAGE 15.

In point (iii) requests a name change with regard to the terminology "virama" versus "halant", in various scripts. This unfortunately cannot be accommodated due to UTC and WG2 policy about name changes (as explained in detail below), but explanatory text and/or annotations in the name list will be written to clarify the issue, and to discuss the two terms.

Point (iv) on page 15 appears to request a change in the rendering model and glyph selection so that halant (virama) would render conjuncts horizontally while "ZWJ" would be used for vertical rendering of conjuncts. Such a model change would be a violation of a set of important stability policies guiding the development of the standard (more on the details and rationale of these policies below) and it would invalidate all existing data and implementations—both font implementations and software. If there are issues with the current model and its description that can be addressed within the constraints of the stability guarantee, UTC welcomes future discussion of the various issues raised by this point, then the Devanagari block introduction in chapter 9 of the Unicode Standard may be clarified.

PAGE 16.

Point (vi) of page 16 suggests that the authors of L2/01-304 are offering to engage in writing detailed block descriptions for the remainder of the Indic scripts that are not already detailed in the Unicode Standard. This offer is very welcome. The more detailed block descriptions will help to ensure that implementers of the standard handle the remaining scripts correctly.

Point (viii) expresses the desire that transliteration between scripts be simple and one-to-one. Clearly it would be a desirable state of affairs, but cannot be achieved at this time. The differences between North and South Indian scripts, as well as the script-specific characters already encoded, make this a practical impossibility.

Point (ix) of page 16 is headed “Updating constraints of Unicode Consortium regarding character encoding stability”. UTC would like to obtain clarification from the authors regarding point (ix), page 16. It is hoped that the current policies posted on the Unicode web site are clear. Many of the Unicode Consortium policies are aligned with policies of WG2. The policies for Unicode and WG2 are available online at the following URLs:

```
http://www.unicode.org/unicode/standard/policies.html
http://anubis.dkuug.dk/JTC1/SC2/WG2/docs/principles.html
```

Stability policies with regard to encoded characters and normalization forms are universally applicable. They are considered vital for many reasons: existing data does not always get updated; creation of uncertainty in models leads to wariness on the part of system vendors about supporting the encoding; other standards bodies such as W3C and IETF need to be extended guarantees of stability because Unicode and 10646 form a foundation for their standards as well. Furthermore, because characters are never removed from the standard, changes in the model for a script's encoding create duplicate encoding of canonically equivalent forms, which also leads to problems in interoperability.

PAGE 17.

The remainder of the document is divided into sections for each of several scripts, or languages, beginning with Devanagari. Likewise, this document follows that structure with headers for clarity in matching responses to the relevant document sections.

Codepoints given in the sections below are in hexadecimal notation and refer to codepoints as used or suggested. Many of the codepoints under discussion herein are not encoded in the Unicode standard, but are encoding suggestions. Here, the suggested codepoint numbers are retained only for clarity in matching responses and must not be taken to imply any encoding or endorsement by UTC.

GENERAL COMMENTS ON VARIOUS PROPOSED ADDITIONS.

For Bengali and Oriya, the document proposes to add DANDA and DOUBLE DANDA characters. The current practice in Unicode is to use the characters encoded in the

Devanagari block, and therefore no further DANDA and DOUBLE DANDA characters will be added. However, the block introductions should be modified to specify where users are to look for the DANDA and DOUBLE DANDA characters (in the Devanagari block).

The document also proposes the addition of INVISIBLE LETTER in a number of scripts. Unicode does not use an INVISIBLE LETTER. The mechanism for achieving the same effects as ISCII are explained in the text of the Unicode Standard, section 9.1. The mechanism is reversible for compatible transcoding to and from ISCII; therefore the INVISIBLE LETTER is not needed in Unicode. For example, the ISCII sequence of a consonant with halant and INV (C+halant+INV) would be encoded in Unicode with a sequence of consonant with halant and zero width joiner (C+halant+ZWJ).

Throughout, the document suggests a number of explanations and details that could be added to the Unicode block introductions, e.g., for the Konkani language, written with Devanagari. UTC will take these all under advisement and update block introductions and add name list annotations as appropriate. Once documents are ready, the UTC would like to request review of the annotations and changes.

Some discussion between MIT and UTC would be helpful in clearing up the issues surrounding use of the grave, acute, udaatta, anudaatta, guru, laghu, swarita marks in various scripts. The block introductions should then be updated with this information. In particular, UTC would like to discover whether it is reasonable and convenient to unify these new marks across the scripts to avoid encoding many of them. Apparently identical marks are suggested in the document for scripts Gujarati and Kannada. This is in line with the Unicode architecture. Diacritic marks that are shared by multiple scripts are only encoded once. For example, the diaeresis is used in many scripts, but only encoded once, at U+0308. A full cross-script analysis of these marks should be made before encoding them.

For TAMIL and other scripts the document strongly indicates that the Unicode encoding with respect to the two-part vowel signs is considered incorrect. The document suggests not using the split pieces, but the two-part vowel signs exclusively. Normalization form “C” (or NFC) should then be preferred here, and UTC may want to annotate this, and/or deprecate the split-up pieces in some cases. Some discussion of these characters between MIT and UTC would be useful. NFC is the normal form for text on the Web and in much data interchange; in NFC the pieces do not occur at all. (Note: The previous L2 document L2/01-037 submitted by the Directorate of Information Technology, Government of Karnataka, also points out that the various “length marks” have no independent existence; see L2/01-037 page 4 of 17.) NFC is discussed in Unicode Technical Report #15.

DEVANAGARI.

Several additions are candidates for encoding. As with all suggested additions, UTC needs detailed explanations of their usage, form, etc. The document itself does not provide sufficient detail for some of the character additions to be formally proposed. For these, UTC will need to obtain further details and work with MIT experts to solidify the proposed additions with sufficient information for acceptance and publication.

0904. This character may be added. UTC would like to know if the matra form should be added as well.

0955, 0956. UTC is already entertaining proposals for various Vedic accents, and would like to include discussion of these and others (Swarita, Laghu, etc) together with Vedic accent proposals. UTC is considering setting aside a separate block for Vedic accents. A URL for PDF versions of the current proposals is here:

`http://www.evertype.com/standards/iso10646/pdf/vedic/`

The characters shown at 0955 and 0956 are already in that set of proposals.

093A. This character, and all other occurrences of "INV" will not be encoded, as explained above. Unicode has a mechanism for producing the results that ISCII achieves with use of INV.

0958 - 095F. The document proposes to discourage the use of these precomposed characters with nuktas. By putting them into the composition exclusions list, UTC has already excluded them from the NFC normalization. Annotations and cautionary statements could also be added to that effect, with whatever degree of strength is appropriate.

094D. A change in the character name is suggested. UTC might want to make an annotation, since a name change would be a violation of the stability policy.

0970. An annotation or explanation as suggested will be added.

0974. DEVANAGARI LETTER SHORT YA is proposed for use in Sindhi. This character is a good candidate for encoding. The UTC would like more detailed explanation of this character's usage and forms. Also UTC would like to clarify, for information only, whether the Government of India intends to add this character to ISCII?

BENGALI.

09BD. "Avagrah" for Bengali is proposed. This is a well-understood character, and no further information is required about instances of avagraha in other scripts. It will be added to the list of proposed characters.

09CE, 09CF, 09DE. The document suggests adding signs for Bengali YA, RA, and LA. This would apparently be a change to the model for Bengali (which would be a violation of the stability policy), and it needs to be considered in detail. UTC requests some detailed explanation to understand the proposal and how it affects the current model for Bengali.

09F4, 09F5, 09F6. Changes in namelist annotation will be added to the list of proposed additions.

GURMUKHI.

0A01, 0A03. Gurmukhi sign "adak bindi" and a visarga will be added to the list of proposed additions. Further clarification or documentation of these would be useful for the block introduction.

0A50. The suggestion for codepoint 0A50 amounts to moving the existing character U+0A74 to a new location. It would be a violation of the stability policy to move this character, and the suggested annotation is already made in the name list for U+0A74. No action is needed. UTC would like to inquire whether a note is in order referring to the obsolescence of this character?

0A4E, 0A4F. These two proposed additions for RA and HA subjoined signs for Gurmukhi seem likely to change the encoding model for Gurmukhi. UTC would like to request detailed explanation to decide whether they are reasonable additions that would not change the current model, and to learn how the proposed characters are used, and how the additions differ (if they do) from use of halant (virama) plus the nominal consonants for producing subjoined sequences.

0A78. "Gurmukhi Sign Khanda" suggested here is identical to the character already encoded as U+262C. There is no need to add this character, but the block introduction will point out that the sign "Khanda" is encoded at U+262C and may be used in Gurmukhi text.

0A33. Moving the nukta to the opposite side of the glyph will be accomplished by an erratum note to the Unicode standard and should be corrected in the next edition.

GUJARATI.

0A8C. Gujarati vocalic L is a well-understood letter and will be added to the list of proposed character additions.

0AD1, 0AD2, 0AD3, 0AD4. These four marks should be treated in conjunction with the equivalent marks for Kannada. See the "General Comments" section above.

0AE1, 0AE2, 0AE3. Additions of Vocalic L, LL will be added to the list of proposed additions without further explanation, as they are well-understood letters.

0AF1. UTC will add this rupee sign for Gujarati to the list of proposed additions, since the symbol is not made from pieces that are already encoded Gujarati characters. The form of this character is very Gujarati-like, and it will be proposed for encoding at this location, rather than in the Currency Symbols block.

ORIYA.

0B0A, 0B0B, 0B48, 0B4C. The document proposes changing shapes of these five codepoints, but there is no explanation. UTC will need details of the changes and motivations before a determination can be made. Are these simply font differences between what UTC used to print the standard and what Oriya fonts are available to MIT for their document? Or is there some deeper problem here?

0B66. The shape/size of the representative glyph for the "zero" character will be changed; the document gives some detail as to why it should be smaller than 0B20, avoiding confusion.

The document suggests removing the annotation under character 0B2C; probably because it also suggests the addition of an Oriya "va" character at 0B35. UTC will remove the annotation, and put "va" on the list of proposed additions for Oriya.

TAMIL.

0B83. The document indicates that this is not a combining character at all, but an independent character. It has "Mc" category in the Unidata. UTC has already agreed to make this change to "Lo", and remove the dotted circle. This is already posted as an erratum on the Unicode web site, effective immediately. WG2 has also already taken corresponding action to correct this in ISO/IEC 10646 as well.

0B82. The document suggests that this is not used in Tamil. Presumably, this means that the Tamil language itself does not use it. UTC would like to clarify whether this should be annotated "for use in Sanskrit"?

Under the heading "TAMIL" the item numbered (3) is a serious issue: "Tamil letter sequencing as in the Unicode Standard 3.0 is also not acceptable. New code-set is being worked out." This looks like groundwork to ask for a replacement of the Tamil encoding by something else. UTC would like to work with the experts at MIT and INFITT to show that the current Unicode Tamil encoding can represent all Tamil syllables, and that the Unicode Collation Algorithm can be used, with the appropriate tailoring, to correctly order Tamil words. If other encodings of Tamil are developed in the future, the UTC would work together with the appropriate organizations to develop precise mapping tables between those encodings and Unicode.

Additionally, UTC would like to mention that eight Tamil symbols are currently being proposed for addition to the standard. They are: Tamil Day Sign, Tamil Month Sign, Tamil Year Sign, Tamil Debit Sign, Tamil Credit Sign, Tamil Ditto Sign, Tamil Rupee Sign, and Tamil Number Sign. These correlate with the Tamil 99 (Phonetic) Keyboard standard.

TELUGU.

0C3C, 0C3D. These two characters, Avagrah and Nukta, will be added to the list of proposed additions for Telugu.

0C0D. The character proposed for 0C0D is already encoded at 0C10 so no action is needed.

0C11. UTC would like to request clarification of the usage of this character.

0C34. UTC would like to request clarification and explanation of this character. It should perhaps be encoded elsewhere, but UTC needs to know the intent for cross-mapping purposes.

KANNADA.

0CBC, 0CBD. The document requests Nukta and Avagraha to be added for Kannada. These are well-understood additions and will be put onto the list of proposed additions without requiring any further information.

0CF8. UTC would like to request clarification and explanation of this character.

0CD2, 0CD1, 0CF9, 0CD3, 0CD4. Several additional diacritics are suggested, and these are treated above in the “General Comments” section.

MALAYALAM.

A number of additions are requested. UTC would like to request explanation and documentation before they can be added to the list of proposed additions for Malayalam. The document also suggests changing the representative shape of 0D4C, but UTC would like to request confirmation and an explanation of the motivation for the proposed changes before taking any action.

The names of seventeen consonants are suggested to be changed, but this request cannot be accommodated as a matter of policy.

The document also suggests removing the character U+0D57 as a duplicate of U+0D4C. The mark U+0D4C is actually the right hand piece of U+0D57, not the entire vowel sign.

This is known to have no independent existence, as pieces of other two-part vowel signs (see the “General Comments” above). Moving his character would be a violation of the stability policy.

ARABIC.

The document suggests adding three characters at 0656, 0657, 0658, since they are used in Urdu.

Ulta Pesh is proposed for encoding at 0656. Ulta Pesh is a combining character (diacritic) which is used to indicate the length of the [u] vowel in Urdu. The proposed symbol is commonly used in everyday writing. It is used in some Iranian Korans. The High Council of Informatics of Iran (HCI) is currently preparing a proposal for its encoding under the name ARABIC TURNED DAMMA.

Jazm is proposed for encoding at 0657. Jazm is the Urdu combining character (diacritic) that is analogous to Arabic Sukun (U+0652). It is considered a glyphic variant of Arabic Sukun with no semantic difference. In a bilingual Arabic-Urdu context, it could be necessary to use Sukun for Arabic and Jazm for Urdu. UTC would like to solicit the opinion of MIT on this matter.

Khari Zer is proposed for encoding at 0658. Zer (Subscript Alef) is a combining character (diacritic) that is used in Urdu in a manner similar to Kasr (U+0650). HCI is preparing a proposal for its encoding, with a name of ARABIC LETTER SUBSCRIPT ALEF.

The Combining Hamza proposed for encoding at 0659 is already encoded at 0654.

“Bat” was proposed for encoding at 0690. UTC would like to request further explanation of this character and information on its usage. In principle there is no objection to encoding it.

Dal with 4 dots above (U+0690). It is known that this character is not used for contemporary standard Urdu, but it was used in the past. The four dots were occasionally used in some Urdu texts as a substitute for the “Small Tah” diacritic that appears, for example, in U+0688.

UTC will consider adding annotations for the variant digit shapes. Likewise, similar text for Sindhi could be included at 06F4-06F7.

The document suggests removing one annotation for 0690 and several other annotations as well. UTC will investigate the proposed annotations.