To: UTC and ISO/IEC JTC1/SC2 WG2

Title: Proposal to encode ORIYA SIGN OVERLINE in the UCS

From: Lorna Priest Evans (SIL International)

Date: 21 June 2019

We wish to propose the addition of ORIYA SIGN OVERLINE in the "Oriya (Odia)" block at the expected (preserved) position for length marks: U+0B55. The character is a combining mark and the glyph should appear as a straight macron above the base character.

The Kuvi [kxv] language of India is spoken by 158,000 [Ethnologue]. The Kuvi language is written using the Oriya (Odia) script. However, length is quite important to distinguish in the Kuvi orthography. At this point there is no ordinary "LENGTH MARK" in the Oriya script (there are two characters 0B56 ◌ ORIYA AI LENGTH MARK and 0B57 ◌ ORIYA AU LENGTH MARK which have a different purpose and different appearance, and are described in The Unicode Standard, Chapter 12 South and Central Asia-I: Official Scripts of India).

The length mark proposed here has this appearance:

In the Table below, one can see the same words listed without and with the length mark. Transliteration is marked, and the definition of the word is included. It is quite clear that the length mark is needed in order to understand the writing system. There have been some language materials printed using this orthography since at least 1987, and those examples will be shown below. (The language community is considering the use of a nukta instead of the overline. However, since the overline has a long history of use we believe it should be encoded for historical reasons even if the use of the nukta is embraced.)

No	Without macron		With macron	
1	ଅଡୁ	support	ଅ̄ଡୁ	share
	[oɖu]		[oːɖu]	
	କଡି	stone	କ̄ଡି	cow
	[koɖi]		[koːɖi]	
2	ଆକୁ	grand father	ଆ̄କୁ	leaf
	[aku]		[aːku]	
	ବାଲା	seed	ବା̄ଲା	spear
	[bala]		[baːla]	
3	ଏସ୍କି	up/top	ଏ̄ସ୍କି	thirsty
	[eski]		[eːski]	
	ମେଣ୍ଟା	knee	ମେ̄ଣ୍ଟ	sheep
	[meɳa]		[meːɳa]	

The ORIYA SIGN OVERLINE acts like a secondary modifier. ଅ is a short vowel and ଅ with the length mark is the long vowel. The ORIYA SIGN OVERLINE would always immediately follow the vowel sign. In the above examples, the codepoints would be:

U+0B05 U+0B55 U+0B21 U+0B41 (share) ଅଡୁ

U+0B15 U+0B55 U+0B21 U+0B3F (cow) କଡି

U+0B06 U+0B55 U+0B15 U+0B41 (leaf) ଆକୁ

U+0B2C U+0B3E U+0B55 U+0B32 U+0B3E (spear) ବାଲା

U+0B0F U+0B55 U+0B38 U+0B4D U+200C U+0B15 U+0B3F ଏସ୍କି (thirsty)

U+0B2E U+0B47 U+0B55 U+0B23 U+0B4D U+0B24 (sheep) ମେଣ୍ଟ

This character could be considered the same as U+0304; COMBINING MACRON. The appearance is very similar. However, in Ganesan's document, there is a relevant quote from Ken Whistler regarding the similar case of encoding a Tamil nukta:

> The problem which needs to be addressed is how to represent
> the diacritic dot (or dots) below that are manifestly present in
> Badaga written text (and in several other minority language
> orthographies) written with the Tamil script in Tamil Nadu.
> The engineering requirements that I see falling out from this are:
> 1. The diacritic must be a combining mark with ccc=7.
> This comes from the general Indic rendering system
> implementation requirements, as suggested by Martin.
> It takes 0323 COMBINING DOT BELOW (ccc=220) off the
> table.
> 2. The diacritic must not unnecessarily break script runs
> in Tamil.
> That follows from the fact that the *rest* of the text in question
> is going to be using Tamil characters. Encoding this with
> a character that breaks script runs will do nobody any favors.
> 3. The diacritic must be easily available on Badaga
> language keyboards and display correctly with fonts
> that support Badaga text display.
> 4. The single dot form and the double dot form should not
> be unified as a single character for representation.
> Unification of those two would be an unusual step and lead
> to confusion in use and data representation, I think. In
> that respect I agree with Michael Everson -- so the requirement
> from the data is for *two* distinct characters that meet
> requirements #1 and #2, although there is only a single
> *function* involved here, and no single orthography seems
> to need to distinguish a single dot diacritic from a double dot
> diacritic systematically.

This case is similar in that if the combining macron were to be used, it would not be supported in the general Indic rendering system implementation requirement. If the combining macron were used, script runs in Oriya would break.

The ORIYA SIGN OVERLINE should immediately follow whatever it comes after. If there is a vowel sign, it should immediately follow the vowel sign. If there is no vowel sign it should immediately follow the base consonant.

The Kuvi language does not use any of the following signs with the overline:

- U+0B3F ◌ି
- U+0B4C �6◌ୌ
- U+0B56 ◌ୖ
- U+0B57 ◌ୗ

The following signs do occur with the overline:

- U+0B3E ◌ା
- U+0B47 6◌

Examples of these are shown in the figures below (for example, Figure 1).

Thus, we wish to request the encoding of ORIYA SIGN OVERLINE as a length mark.

Character Name and annotations

0B55	ORIYA SIGN OVERLINE
	• Kuvi language

Unicode Character Properties

Properties for the new character:

`0B55; ORIYA SIGN OVERLINE ;Mn;0;NSM;;;;;N;;;;;`

This character has no decomposition.

This character can sort immediately before U+0B56 ORIYA AI LENGTH MARK.

Indic Properties

```
# Indic_Positional_Category=Top
# Indic_Syllabic_Category=Vowel_Dependent
```
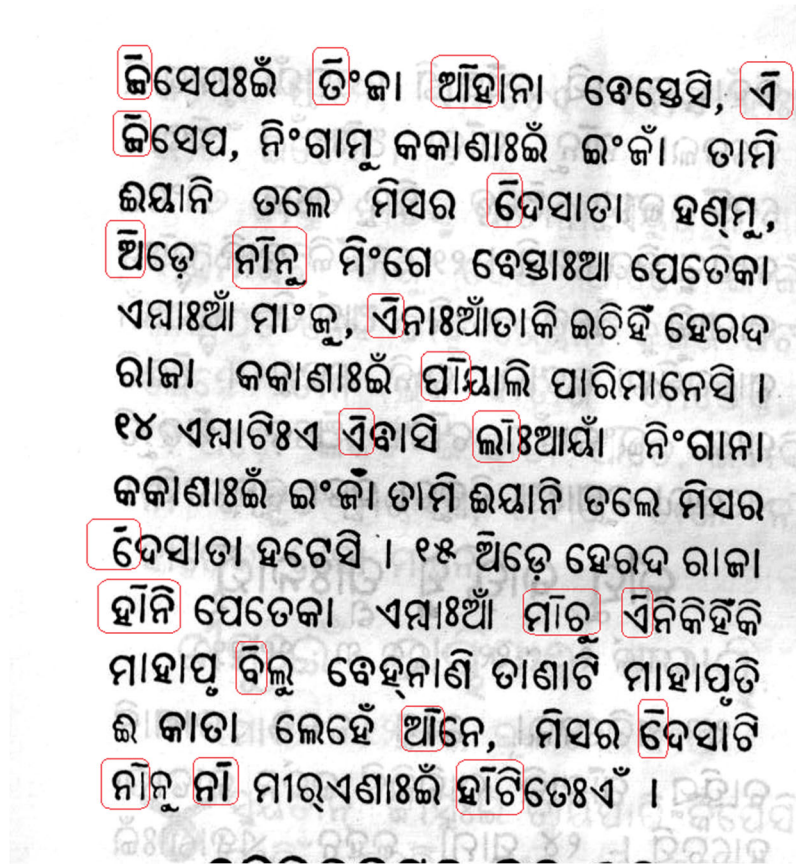
Figures

Figure 1. ORIYA SIGN OVERLINE (*BSI, 1987*).

Figure 2. ORIYA SIGN OVERLINE (*BSI, 1987*).

Figure 3. ORIYA SIGN OVERLINE (*BSI, 1987*).

ବୈବାଣାୟଇଁ ନିନ୍ଦା କିତେର ।

ଜୀସୁ ହାଁନାୟ୍ଡ଼ି

(ମାତି ୨୭ : ୪୫-୫୬; ଲୂକ ୨୩ : ୪୪-୪୯;
ଜହ ୧୯ : ୨୮-୩୦)

୩୩ ବାଁର ଗଣ୍ଡା ବେଲାଟିୟଏ ଦିସା
ବାରେ ଆଦେରି ଆଁହା ହାଜାନା ତୀନି ଗଣ୍ଡା
ପେତେକା ଆଦେରି ଆଁହାନା ଡ଼ୟିତେ ।
୩୪ ତୀନି ଗଣ୍ଡା ବେଲାତା ଜୀସୁ କାଜା ଗିୟଁା
ତଲେ ହାଁଟିହଁ ଏଲେ ଇତେସି, ଏଲହି, ଏଲହି,
ଲେମା ସାବାକ୍ତାନି ଇତିହଁ ଏ ନୀ ମାହାପୁ,
ଏ ନୀ ମାହାପୁ, ନୀନୁ ନାଂଗେ ବୈନାୟଆଁ
ତାକି ପିଷ୍ଟାତି । ୩୫ ଏୟାୟଁା ନିଗ୍ମାତି ଲିକୁ
ବିତ୍ରାଟି ଏଚର ଜାଣା ଏଲେ ଇତେରି,
ମେହ୍ଦୁ, ବୈବାସି ଏଲିୟୟଇଁ ହାଟିମାନେସି ।
୩୬ ଇଂଜାଁ ରୟଅସି ହଟି ହଟିହଁ ହାଜାନା ରି
ଡ଼ାୟଲିତା ପୁଲା ବୈସୁ ନେଂଜ କିହାନା ବୈଦାୟଁା
ରି ବାଡ଼୍ଗାତା ଦସାନା ବୈବାଣାକି ଗସାଲି

Figure 4. ORIYA SIGN OVERLINE (*Gospel of Mark* in Kuvi).

References

BSI. 1987. *The New Testament: Kuvi - O.V (Ref)*. The Bible Society of India. Printed in Calcutta, India.

Ganesan N. 4-20-2016 Proposal to Encode Single-dot Sign at U+1133B Comment on M. Hosken's Badaga dot sign Proposal (L2/15-256) (https://www.unicode.org/L2/L2016/16084-badaga-single-dot.pdf accessed 30-Oct-2018)

Simons, Gary F. and Charles D. Fennig (eds.). 2018. Ethnologue: Languages of the World, Twenty-first edition. Dallas, Texas: SIL International. (http://www.ethnologue.com accessed 30-Oct-2018).

The Unicode Consortium. *The Unicode Standard*, Version 11.0.0, (Mountain View, CA: The Unicode Consortium, 2018. ISBN 978-1-936213-19-1) (http://www.unicode.org/versions/Unicode11.0.0/ accessed 28-Nov-18).

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for
guidelines and details before filling this form.
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest *Roadmaps*.

A. Administrative

1. **Title:** Proposal to encode ORIYA SIGN OVERLINE in the UCS
2. Requester's name: *Lorna Priest Evans*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *21-June-2019*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later: *No*

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Oriya (Odia)*
2. Number of characters in proposal: *1*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary   x   B.1-Specialized (small collection)   B.2-Specialized (large collection)
   C-Major extinct   D-Attested extinct   E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic   G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines"
      in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
      Michael Everson
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):

6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources)
      of proposed characters attached? *Yes*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input,
   presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist
in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties
are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths
etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up
contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at
http://www.unicode.org for such information on other scripts. Also see Unicode Character Database
(http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the
Unicode Technical Committee for inclusion in the Unicode Standard.

---

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

Lorna Priest Evans   June 21, 2019

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	*No*
If YES explain	
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	*Yes*
If YES, with whom?	Kuvi language speakers
If YES, available relevant documents:	*Examples in document were provided by them.*
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	*Yes*
Reference:	*See proposal*
4. The context of use for the proposed characters (type of use; common or rare)	*Yes*
Reference:	*See proposal*
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference:	*See proposal*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	Yes
If YES, is a rationale provided?	*Yes*
If YES, reference:	*See proposal*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	*n/a*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	*No*
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	*No*
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?	*Yes*
If YES, is a rationale for its inclusion provided?	*Yes*
If YES, reference:	*See proposal*
11. Does the proposal include use of combining characters and/or use of composite sequences?	*Yes, combining*
If YES, is a rationale for such use provided?	*Yes*
If YES, reference:	*See proposal*
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	*No*
If YES, reference:	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	*No*
If YES, describe in detail (include attachment if necessary)	
13. Does the proposal contain any Ideographic compatibility characters?	*No*
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	

Lorna Priest Evans   June 21, 2019