

Proposal to encode Kaithi Number Sign Above

Benjamin Yang
 Script Encoding Initiative
 University of California, Berkeley
 Berkeley, California, U.S.A.
 yang@panlex.org

April 29, 02016

1. Introduction

- The Kaithi script was proposed for encoding in L2/08-194 (N3389) (prepared by Anshuman Pandey), with section 5.11 of this document explaining the KAITHI ENUMERATION SIGN. As written, L2/08-194 indicates that the KAITHI ENUMERATION SIGN does not require separate code points for the 3 possible uses of the character.
- However, in document L2/08-370 (N3544), Deborah Anderson has noted that two possible uses of the character should be disambiguated and assigned separate code points, with 110BC KAITHI ENUMERATION SIGN assigned to the inline usage of the character and the format control character being encoded as 110BD KAITHI NUMBER SIGN.
- In L2/08-400, Deborah Anderson noted the two uses of the 110BD KAITHI NUMBER SIGN and stated:
 - “If the two uses are distinct and can take multiple digits, a second KAITHI NUMBER SIGN seems warranted and should be proposed at the next UTC.”
- Currently, the following characters are encoded:

110BC		KAITHI ENUMERATION SIGN
110BD		KAITHI NUMBER SIGN
- In 2015, Kamal Mansour of Monotype has contacted the Script Encoding Initiative with a recommendation to further disambiguate between the two combining uses of 110BD KAITHI NUMBER SIGN (combining above and combining below), with each use being assigned its own code points due to the semantic distinction between the two use cases.

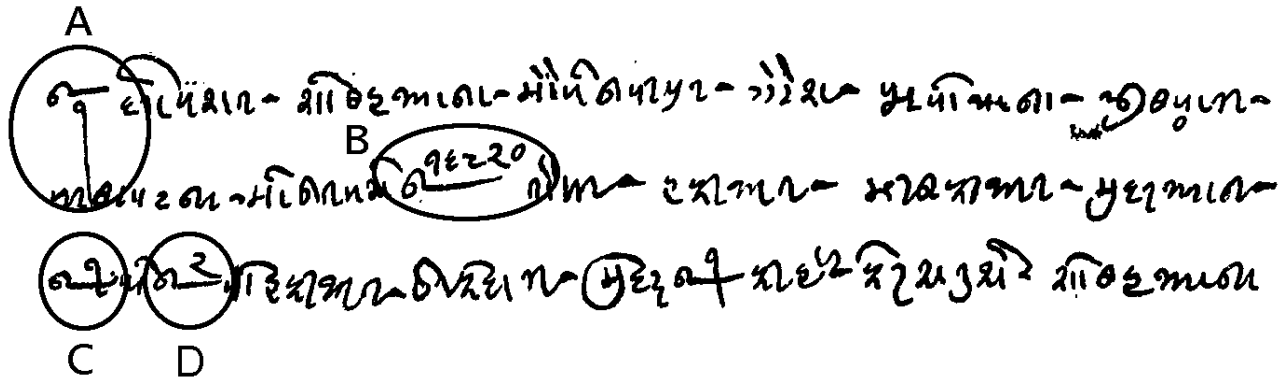
2. Request

- This proposal requests the addition of one new control format character, with the proposed name KAITHI NUMBER SIGN ABOVE at U+110CD.

3. Justification

- In section 5.11 of the Kaithi proposal (L2/08-194), it is indicated that there are two possible positions for the Kaithi number sign when combining with digits, with the placement of the sign reflecting two different usages.
- In the image below (from p. 34 of the Kaithi proposal), both usages are present.
 - If the sign is written above the digit or digits, it indicates a number in an itemized list, similar to U+2116 NUMERO SIGN ("A" in the example).
 - If the sign is written below the digit or digits, it indicates a numerical reference ("B", "C", and "D" in the example).

- In the example below, "A" is the first number in a listing, "B" is the court-case number, and "C" and "D" represent the plaintiff numbers (Plaintiff 1 and Plaintiff 2).



- Due to the fact that the position of the number sign cannot be anticipated, it is proposed that 110BD (as proposed in L2/08-370) be assigned to the numerical reference usage (mark below digit) and a new code point 110CD be assigned to the itemized list usage (mark above digit). (110BC will remain the code point for the in-line usage of the sign, as indicated in L2/08-370).
- 110BD is assigned the numerical reference usage as it is likely to be the more common usage of the character.
- The glyph for KAITHI NUMBER SIGN ABOVE will be represented as the glyph for KAITHI ENUMERATION SIGN placed high in a dotted box to represent the fact that the mark is placed above the numerals. This is consistent with glyph images for other control format characters:

Subtending marks

0600		ARABIC NUMBER SIGN
0601		ARABIC SIGN SANAH
0602		ARABIC FOOTNOTE MARKER
0603		ARABIC SIGN SAFHA
0604		ARABIC SIGN SAMVAT

- used for writing Samvat era dates in Urdu

Supertending mark

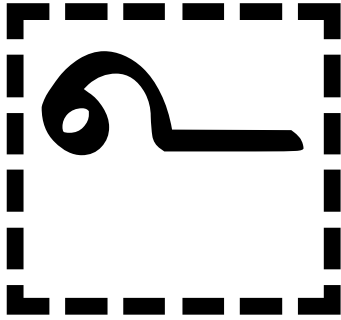
0605		ARABIC NUMBER MARK ABOVE
------	--	--------------------------

- may be used with Coptic Epact numbers

- The existing glyph for KAITHI NUMBER SIGN is appropriate for the numerical reference usage as it is already represented as the glyph for KAITHI ENUMERATION SIGN placed low in a dotted box, reflecting the fact that the mark will be placed below the numerals.

4. Character Data

4.1. Glyph



4.2. Character properties

In the format of UnicodeData.txt:

```
110CD;KAITHI NUMBER SIGN ABOVE;Cf;0;L;;;;;N;;;;;
```

4.3. Linebreaking

In the format of LineBreak.txt:

```
110CD;AL          # Cf          KAITHI NUMBER SIGN ABOVE
```

5. Acknowledgements

I would like to thank Kamal Mansour of Monotype for initially pointing out the need to disambiguate these two usages and Deborah Anderson for editing this proposal.

This project was made possible in part through a Google Research Award, granted to Deborah Anderson for the Script Encoding Initiative, and a grant from the United States National Endowment for the Humanities (PR-50205-15), which funds the Universal Scripts Project (part of the Script Encoding Initiative at the University of California, Berkeley). Any views, findings, conclusions or recommendations expressed in this publication do not necessarily reflect those of Google or the National Endowment for the Humanities.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://std.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://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest Roadmaps.

A. Administrative

1. Title:	Proposal to encode Kaithi Number Sign Above
2. Requester's name:	Benjamin Yang, Script Encoding Initiative, UC Berkeley
3. Requester type (Member body/Liaison/Individual contribution):	Liaison contribution
4. Submission date:	28 March 02016
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:	

B. Technical – General

1. Choose one of the following:	
a. This proposal is for a new script (set of characters):	
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	
Name of the existing block:	Kaithi
2. Number of characters in proposal:	1
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary ☐	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?	
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?	see L2/08-194 and L2/08-370
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	see L2/08-194 and L2/08-370
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)?	
Yes, please also see L2/08-194	

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)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	<i>no</i>
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.)?	<i>no</i>
If YES, with whom?	
If YES, available relevant documents:	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	<i>see L2/08-194</i>
Reference: <i>Awadhi, Bhojpuri, Magahi, and Maithili speakers; as well as linguists, historians, legal scholars working with sources from colonial South Asia.</i>	
4. The context of use for the proposed characters (type of use; common or rare)	<i>common</i>
Reference: <i>Court records from colonial India, pedagogical materials from north India, commercial and accounting records; religious and literary texts; bibles printed in north India during the 19th and early 20th century.</i>	
5. Are the proposed characters in current use by the user community?	<i>see L2/08-194</i>
If YES, where? Reference: <i>In India, the United States, and other localities</i>	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	<i>No</i>
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	<i>n/a</i>
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	<i>no</i>
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?	<i>no</i>
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?	<i>yes</i>
If YES, is a rationale for its inclusion provided?	
If YES, reference: <i>see L2/08-194 and L2/08-370</i>	
11. Does the proposal include use of combining characters and/or use of composite sequences?	<i>no</i>
If YES, is a rationale for such use provided?	
If YES, reference:	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	
If YES, reference:	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	<i>yes</i>
If YES, describe in detail (include attachment if necessary)	
<i>See L2/08-194 and L2/08-370</i>	
13. Does the proposal contain any Ideographic compatibility characters?	<i>no</i>
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	