

Proposal to encode the Khojki letter SHORT I in Unicode

Anshuman Pandey


pandey@umich.edu
pandey.github.io/unicode

May 21, 2021




1 Introduction






This is a proposal to encode a new character in the Khojki block of the Unicode standard:









GLYPH	CODE	CHARACTER NAME
	11240	KHOJKI LETTER SHORT I

The character is named KHOJKI LETTER SHORT I because the name KHOJKI LETTER I is already assigned to  (U+11202). The representative glyph has been designed to conform to the letterforms in the Khojki code chart, which are based on the ‘Khojki Jiwa’ font designed by Pyarali Jiwa.

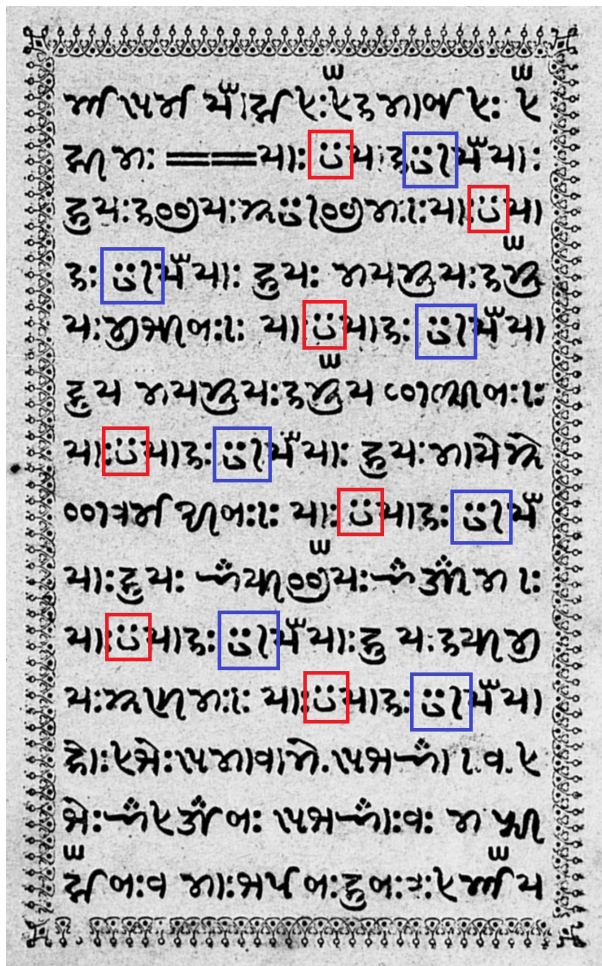
2 Description

Conventional Khojki lacks distinctive letters for independent forms of the vowels *i* and *ī*. Both of these vowels are represented using  U+11202 KHOJKI LETTER I. It does, however, have distinctive signs for dependent forms of these vowels:  U+1122D KHOJKI VOWEL SIGN I for *i* and  U+1122E KHOJKI VOWEL SIGN II for *ī*.

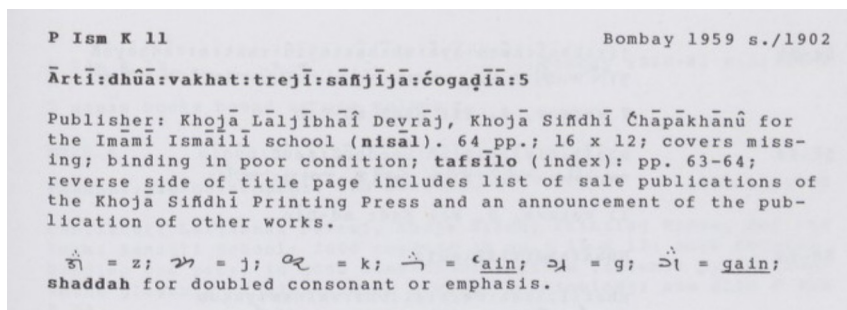
When Khojki printers encountered a text in which independent *i* and *ī* needed to be distinguished, they used  specifically for *ī* and represented *i* using , a letter that is not part of the conventional repertoire. The origins of  are unknown and it is not described in any detail in the scholarly literature, apart from its inclusion in charts. It could be a revival of a historical, and previously abandoned, letter for *i*. There is some evidence that  is an alternate form of , used in some manuscripts for both short and long forms of the vowel. It appears this letter was repurposed to enable distinctions between *i* and *ī*.

Palaeographically,  is a valid letterform. It is related to letters such as  U+11185 SHARADA LETTER I. Within Khojki graphemics, it is  U+11202 KHOJKI LETTER I with the right-side element  removed. The structure of these two letters adheres to an Indic graphemic paradigm for short and long vowels, in which a modifier stroke is added to the short vowel letter to represent the long vowel. Theoretically,  would be the base form and  would be the derived by adding , which resembles the dependent sign for the long vowel, ie.  U+1122E KHOJKI VOWEL SIGN II.

The 𑈍 is used, for example, in a printed edition of the book *Ārtī dhūā vakhat trejī sāñjijā ʿogaḍīā* 5, p. 25, published in 1902 by the Khoja Sindhi Printing Press, Bombay. This book contains liturgies composed by Pīr Ṣadr ad-Dīn, who founded the Khoja Ismaili community in the 14th century. The excerpt below shows contrastive usage of 𑈍 (red) for *i* and 𑈍̄ (blue) for *ī* in the phrase 𑀗: 𑀘𑀙𑀚: 𑀘𑀙𑀚𑀛: *lā ilaha illalā*, which is a transliteration of the Arabic phrase لا إله إلا الله *lā ʾilāha ʾillā-llāh* “there is no god, but the god”:



The above excerpt is taken from the *Ārtī dhūā vakhat trejī sāñjijā ʿogaḍīā* 5 in the Harvard collection. However, this letter is not mentioned in the entry for this text in *The Harvard Collection of Ismaili Literature in Indic Languages* (p. 283), compiled by Ali S. Asani (1992):



The above excerpt from the text clearly shows distinctive usage of **𑂔**. The absence of the letter from the enumeration of special characters in the entry is a rare omission by the catalogue editor. Indeed, Asani lists **𑂔** as a documented variant in an accompanying table of scripts:

Roman	Urdu	Sindhi	Gujarati	Khojki#
a	ا	ا	અ	𑂔 𑂕
ā	آ	آ	આ	𑂔𑂕 𑂔𑂕
i*	ا	ا	ઈ	𑂔 [i:] (𑂔 𑂔?)
ī*	ای	ای	ई	𑂔 [i:] (𑂔 𑂔?)
u*	و	و	ઉ	𑂕 [u]
ū*	و	و	ऊ	𑂕 [u:] (𑂕?)
o	او	او	ઓ	𑂕 [o, ɔ] (𑂕?)
e	ے	ای	એ	𑂕 [e:] (𑂕?)
ai	ای	ای	ૐ	[𑂕 𑂔, 𑂔 𑂔]
au	او	او	ૐ	[𑂕 𑂕, 𑂕 𑂔]

Usage of **𑂔** (red) for both long and short forms of the independent vowel in manuscripts is presented by Gulshan Khakee (1981, p. 147). The table also shows usage of **𑂔𑂕** (blue) for both in a Khojki script primer.

Roman	Devana-gari	Gujarati	Sindhi	D MS 1815	Kx MS 1737	Khojaki Primer 1932
a	अ	અ	ا	अ	𑂔	𑂔
ā	आ	આ	آ	आ	𑂔𑂕	𑂔𑂕
i	इ	ઈ	ى	ई	𑂔 ¹	𑂔𑂕
ī	ई	ई	ى	ई	—	—

The evidence suggests that **𑂔** and **𑂔𑂕** may have been used historically in Khojki for *i* and *ī*, respectively. At some point in time, a single letter began to be used for both vowels. In some sources that letter was **𑂔**, while in others it was **𑂔𑂕**. Eventually, **𑂔𑂕** became the conventional letter. When the need arose to distinguish short and long forms of the vowel /i/ at the beginning of words, Khojki users simply looked to the history of their script to find a practical and palaeographically valid solution.

As Khojki is a liturgical script of Ismaili communities of South Asia, it is important that the Unicode repertoire for Khojki provide all characters required for the distinctive representation of records in digital plain text. The **𑂔** KHOJKI LETTER SHORT I should be encoded in the standard to enhance such support.

3 Collation

The ્ KHOJKI LETTER SHORT I should be sorted after ્ U+11201 KHOJKI LETTER AA and before ્ U+11202 KHOJKI LETTER I.

4 Discouraged Combinations

The ્ KHOJKI LETTER SHORT I should not be used with ્ U+1122E KHOJKI VOWEL SIGN II for representing ્ U+11202 KHOJKI LETTER I.

There are other discouraged combinations for representing Khojki vowel letters, which should be added to the core specification:

For	Use	Do Not Use
ଁ	ଁ U+11201 KHOJKI LETTER AA	ଁ U+11200 KHOJKI LETTER A + ્ U+112CC KHOJKI VOWEL SIGN AA
ଁ	ଁ U+11202 KHOJKI LETTER I	ଁ U+11240 KHOJKI LETTER SHORT I + ્ U+112CC KHOJKI VOWEL SIGN II
ଁ	ଁ U+11205 KHOJKI LETTER AI	ଁ U+11200 KHOJKI LETTER A + ્ U+11231 KHOJKI VOWEL SIGN AI
ଁ	ଁ U+11207 KHOJKI LETTER AU	ଁ U+11200 KHOJKI LETTER A + ્ U+11233 KHOJKI VOWEL SIGN AU

5 Input Methods

On digital keyboards with touch-and-hold features, the ્ KHOJKI LETTER SHORT I should be added to the character palette for the ્ U+11202 KHOJKI LETTER I key. On other keyboards it may be placed on a SHIFT, CTRL, or ALT layer, mapped to the same key as U+11202 KHOJKI LETTER I.

6 Character Data

Character Properties: UnicodeData.txt

```
11240;KHOJKI LETTER SHORT I;Lo;0;L;;;;;N;;;;;
```

Linebreaking Properties: LineBreak.txt

```
11240;AL          # Lo          KHOJKI LETTER SHORT I
```

Indic Syllabic Category: IndicSyllabicCategory.txt

```
11240          ; Vowel_Independent # Lo  [8] KHOJKI LETTER SHORT I
```

7 References

Asani, Ali S. 1992. *The Harvard Collection of Ismaili Literature in Indic Languages: A Descriptive Catalog and Finding Aid*. Boston: G. K. Hall & Co.

Khakee, Gulshan. 1981. “The Dasa Avatara of Pir Shams as Linguistic and Literary Evidence of the Early Development of Ismailism in Sind.” In *Sind Through the Centuries*. Proceedings of an International Seminar held in Karachi in Spring 1975 by the Department of Culture, Government of Sind. Pages 143–155. Edited by Hamida Khuhro. Karachi: Oxford University Press.

Pandey, Anshuman. 2011. “Final Proposal to Encode the Khojki Script in ISO/IEC 10646” (L2/11-021) <https://www.unicode.org/L2/L2011/11021-khojki.pdf>

Pīr Ṣadr ad-Dīn. Ārtī:dhûā:vakhat:tredjī:sāñjījā:ćogađīā:5. Bombay: Khojā Lālġībhāī Devrāj, Khojā Siñdhā Ćhāpākhānū, 1902. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

**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 the Khojki Letter SHORT I in Unicode
2. Requester's name:	Anshuman Pandey <pandey@umich.edu>
3. Requester type (Member body/Liaison/Individual contribution):	Expert contribution
4. Submission date:	2021-05-21
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:	Yes
Name of the existing block:	Khojki
2. Number of characters in proposal:	1
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary <input checked="" type="checkbox"/> B.1-Specialized (small collection) <input type="checkbox"/> B.2-Specialized (large collection) <input type="checkbox"/>	
C-Major extinct <input type="checkbox"/> D-Attested extinct <input type="checkbox"/> E-Minor extinct <input type="checkbox"/>	
F-Archaic Hieroglyphic or Ideographic <input type="checkbox"/> G-Obscure or questionable usage symbols <input type="checkbox"/>	
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?	Anshuman Pandey
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	Anshuman Pandey
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)?	Yes

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? If YES explain	No
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.)? If YES, with whom? If YES, available relevant documents:	No
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Reference:	Yes <i>See text of proposal</i>
4. The context of use for the proposed characters (type of use; common or rare) Reference:	Common <i>See text of proposal</i>
5. Are the proposed characters in current use by the user community? If YES, where? Reference:	Yes; <i>The Khoja Ismaili religious community and scholars</i>
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? If YES, is a rationale provided? If YES, reference:	N/A
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? If YES, is a rationale for its inclusion provided? If YES, reference:	No
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? If YES, is a rationale for its inclusion provided? If YES, reference:	No
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? If YES, is a rationale for its inclusion provided? If YES, reference:	No
11. Does the proposal include use of combining characters and/or use of composite sequences? 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:	No N/A
12. Does the proposal contain characters with any special properties such as control function or similar semantics? If YES, describe in detail (include attachment if necessary)	No
13. Does the proposal contain any Ideographic compatibility characters? If YES, are the equivalent corresponding unified ideographic characters identified? If YES, reference:	No