L2/17-012
WG2 N4812

# Additional Characters for Kharoṣṭhī Script

Andrew Glass, Microsoft | Stefan Baums

andrew dot glass at microsoft dot com | baums at lmu dot de

January 17th 2017

## Introduction

The Kharoṣṭhī script was added to Unicode in version 4.1 based on the proposal by Glass, Baums, and Salomon 2002 (L2/02-203-R2). Three fonts supporting the Unicode encoding for Kharoṣṭhī with the requisite OpenType shaping tables are known to exist:

- Kharosthi Unicode (a private font used for the Kharoṣṭhī code chart)
- Noto Sans Kharoshthi (https://www.google.com/get/noto/#sans-khar)
- Segoe UI Historic (included in Windows 10)

Rendering Kharoṣṭhī based on the Unicode encoding makes use of the Universal Shaping Engine and compatible engines.

The original encoding proposal for Kharoṣṭhī was based on studies of the then known manuscripts, inscriptions, documents, and coins in the Gāndhārī language and Kharoṣṭhī script, and, in particular, the work *A Preliminary Study of Kharoṣṭhī Manuscript Paleography* (Glass 2000). Since that time new discoveries and ongoing research on the Kharoṣṭhī materials have continued to expand our understanding of the script. This proposal aims to bring the encoding for Kharoṣṭhī up to date through the addition of signs not included in the original proposal back in 2002.

In particular, the discovery of two documents that record sign inventories of the Kharoṣṭhī script have improved our understanding of particular details of the primary signs of the script (see Salomon 2004, Strauch 2008: 121–3, Melzer 2015). Consequently, we feel that it is both necessary and timely to update the encoding of Kharoṣṭhī with additional characters. The traditional sign inventory of Kharoṣṭhī is called the Arapacana syllabary after the first five signs. The established form is as follows (after Salomon 2004: 47, with signs added using the Segoe UI Historic font where available)

| The Arapacana Syllabary (based on Salomon 2004: 47) | | | | | |
|---|---|---|---|---|---|
| 1. *a* 𐨀 | 2. *ra* 𐨪 | 3. *pa* 𐨨 | 4. *ca* 𐨌 | 5. *na* 𐨤 | 6. *la* 𐨪 |
| 7. *da* 𐨣 | 8. *ba* 𐨩 | 9. *ḍa* 𐨦 | 10. *ṣa* 𐨯 | 11. *va* 𐨷 | 12. *ta* 𐨟 |
| 13. *ya* 𐨁 | 14. *ṭha* 𐨡 | 15. *ka* 𐨐 | 16. *sa* 𐨮 | 17. *ma* 𐨫 | 18. *ga* 𐨒 |

| | | | | | |
|---|---|---|---|---|---|
| 19. *tha* 𐨠 | 20. *ja* 𐨡 | 21. *śpa* 𐨢 | 22. *dha* 𐨣 | 23. *śa* 𐨤 | 24. *kha* 𐨥 |
| 25. *kṣa* 𐨦 | 26. *sta* 𐨧 | 27. *ña* 𐨨 | 28. *ʈa* 𐨩* | 29. *bha* 𐨪 | 30. *cha* 𐨫 |
| 31. *spa* 𐨬 | 32. *vha* 𐨭* | 33. *tsa* 𐨮 | 34. *gha* 𐨯 | 35. *ʈha* 𐨰 | 36. *ṇa* 𐨱 |
| 37. *pha* 𐨲 | 38. *ḱa* 𐨳 | 39. *za* 𐨴 | 40. *cā* 𐨵 | 41. *ṭa* 𐨶 | 42. *ḍha* 𐨷 |

*The signs *ʈa* and *vha* are included in this proposal. All the other characters can be rendered correctly using the existing encoding.

In this document references to catalog numbers for Kharoṣṭhī items (CK-) are based on Baums and Glass 2002– b.

## Proposed characters

| Glyph | Code | Character name |
|---|---|---|
| 𐨩 | 10A34 | KHAROSHTHI LETTER TTTA |
| 𐨭 | 10A35 | KHAROSHTHI LETTER VHA |
| 𐩷 | 10A48 | KHAROSHTHI FRACTION ONE HALF |

## Description

### *ʈA*

The recent discovery of an acrostic text that preserves the complete form of the traditional Arapacana syllabary proves that two signs which had been considered variants of the same sign (Glass 2000: 69) were in fact distinct within the tradition. The two signs are now transliterated *ṭa* and *ʈa* (Salomon 2004: 46–7). The former is by far the more common, and is the form illustrated in the Unicode chart (𐨶 U+10A1A). The second form is being proposed here for encoding.

#### Examples

| | | |
|---|---|---|
| *ʈa* 𐨩 |  | *a**ʈa**ragavaśeṇa* (CKM 268 l.55; unpublished) |

| | | |
|---|---|---|
| *ṭ́a* 𐨟 |  | ***ṭ́a,*** ‘(the syllable) *ṭ́a*’ (CKI 512 l. 3; Salomon 2004) |

Example of the standard *ṭa*

| | | |
|---|---|---|
| ṭa 𐨠 |  | ***ṭaṭa*** (CKM 268 l.81; unpublished) |

## Attested forms

| *ṭa* | *ṭi* |
|---|---|
| 𐨟 | 𐨟𐨿 |

## *VHA*

The character *vha* (KHAROSHTHI LETTER VHA) is well attested in the corpus and should have been proposed as a distinct character in the original proposal to encode Kharoṣṭhī. However, due to an desire to minimize code points it was omitted in favor of rendering this form via the sequence KHAROSHTHI LETTER VA + KHAROSHTHI VIRAMA + KHAROSHTHI LETTER HA. The Arapacana syllabary contains several signs that are represented as Consonant Virama Consonant sequences as they are understood to represent or correspond closely to conjunct consonant groups in their cognate Sanskrit forms. This is not the case for *vha* which is understood to be an aspirated form of *va,* and therefore warrants separate encoding. Separate encoding would facilitate sorting and other text processing purposes. Note that contemporary practice does not use the Arapacana order as a sort order for Kharoṣṭhī.

## Examples

| | | |
|---|---|---|
| vha 𐨯 |  | *la**vha**di* ‘obtains’ (CKM 17 31r l. 21; unpublished) |

| | | |
|---|---|---|
| vha 𑀯 |  | *ṇa**vha**pati* ‘Navhapati’ (a title) (CKI 249 1b; Bailey 1980) |

## Attested forms

| *vha* | *vhaṃ* | *vhi* | *vhiṃ* | *vhu* | *vhe* | *vho* | *vhra* | *vhri* | *vhrya* |
|---|---|---|---|---|---|---|---|---|---|
| 𑀯 | 𑀯𑀁 | 𑀯𑀺 | 𑀯𑀺𑀁 | 𑀯𑀼 | 𑀯𑁂 | 𑀯𑁄 | 𑀯𑁆𑀭 | 𑀯𑁆𑀭𑀺 | 𑀯𑁆𑀭𑁆𑀬 |

½

A sign for a fraction was recently discovered in two inscriptions (CKI 721 and CKI 727) published by Harry Falk (Falk 2001). Falk has observed, that this sign also occurs in several Niya documents (CKD 131, CKD 437, CKD 595, CKD 702). The editors of the Niya documents identified this sign as an allograph of the digit 1, but the contexts also support the reading ½ and this interpretation is more likely since there is good basis to support the allograph for one. The sign consists of an additional stroke added to the left side of one. In one inscription and in the Niya documents this stroke points to the bottom left at about 45°. In the other case the stroke points up. Based on the current data it is reasonable to suppose that the upward stroke also indicates the same fraction since this is otherwise unattested. Unfortunately, no images of the Niya documents in question are available, but the sign is illustrated in Rapson’s paleography chart (Boyer, Rapson, and Senart 1920–29: pl. 14), and clearly described in his discussion.

In addition to the ½ sign there may also be a sign for 1½ (CKD 211), however images of this document are not available and the sign is not illustrated. In this case, the same downward stroke is applied to the digit 2. It is reasonable to infer this also indicates subtraction of a half since this is how half units are expressed linguistically, for example, *adha-trodaśa* ‘twelve and a half’ (Baums and Glass 2002– a, s.v.), literally, ‘half-thirteen’. This then reminds us of the subtracting half stroke found in Tibetan, e.g., 𑦤 ‘one’ and 𑦥 ‘half’; 𑦢 ‘two’ and 𑦣 ‘one and a half’, such that it is tempting to see a connection between these two Central Asian notations, albeit their respective attestations are separate by hundreds of years.

## Examples

| | | |
|---|---|---|
| ½ 𑁡 |  | ½ (CKI 727 l. 1; Falk 2001: 317) |
| ½ 𑁡 |  | ½ (CKI 721 l. 1; Falk 2001: 314) |

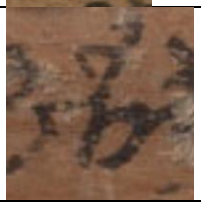
| | | |
|---|---|---|
| ½ 𐨁 |  | ½ (Boyer, Rapson, and Senart 1920–29: pl. 14) |

## Proposed sequence

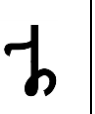
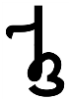
### *Subjoined -Ý-*

The sign *-ý-* occurs in 163 documents in the Niya collection. The status of this sign has been reexamined in connection with ongoing work on the Gāndhārī Dictionary (Baums and Glass 2002– a). At the time of the original proposal for Kharoṣṭhī the status of this sign was unclear as two interpretations were in use (see Boyer et al. 1920–29: 318-19; Burrow 1937: 11; Glass 2000:126–7). Since then one of the possible interpretations, i.e., *-ṕ-*, has been resolved to be simply a cursive form of the subjoined consonant *pa*, and therefore those cases do not warrant separate encoding. The remaining cases, that cannot be *-pa*, are now understood to be a subjoined *y* sign. This sign occurs almost exclusively in loan words into Gandhari, most likely from Tocharian A, where the sequence *-ly-* occurs regularly. The most frequent cases of such loans are proper names occurring in the Niya Documents. The acute accent is used in transliteration to differentiate this sign from the regular subjoined *ya* (U+10A3F U+10A29), but otherwise there is no firm ground to that it was pronounced differently. Because the shape of the sign clearly resembles the full *ya* shape (^) cursively attached to the base of the preceding *la* (𐨪) we propose to use a ZWJ VIRAMA sequence to invoke an explicit medial form, LA ZWJ VIRAMA YA.

#### Examples

| | | |
|---|---|---|
| *lýa* 𐨪𐨿𐨯 |  | *alýaya*, ‘Alýaya’ a proper name (CKD 214 l. 2) |
| *lýa* 𐨪𐨿𐨯 |  | *lýa* (CKD 341 l. 3) |
| *lýi* 𐨪𐨿𐨯𐨰 |  | *palýi, cf.* Skt. *bali,* ‘tax’ (CKD 275 l. 2) |
| *lýo* 𐨪𐨿𐨯𐨱 |  | lýokmana ‘an item of clothing?’ (CKD 318 cr.1) |

### Attested forms

| *lýa* | *lýaṃ* | *lýi* | *lýo* |
|---|---|---|---|
|  |  |  |  |

## Character data

### Core data (Unicode data.txt)

```
10A34;KHAROSHTHI LETTER TTTA;Lo;0;R;;;;;N;;;;;
10A35;KHAROSHTHI LETTER VHA;Lo;0;R;;;;;N;;;;;
…
10A48;KHAROSHTHI FRACTION ONE HALF;No;0;R;;;;1/2;N;;;;;
```

### Line breaking (LineBreak.txt)

```
10A19..10A35;AL # Lo [29] KHAROSHTHI LETTER NYA..KHAROSHTHI LETTER VHA
10A40..10A48;AL # No  [9] KHAROSHTHI DIGIT ONE..KHAROSHTHI FRACTION ONE HALF
```

### Syllabic category (IndicSyllabicCategory.txt)

```
# Indic_Syllabic_Category=Consonant

10A19..10A35 ; Consonant # Lo [29] KHAROSHTHI LETTER NYA..KHAROSHTHI LETTER VHA

# Indic_Syllabic_Category=Number
10A40..10A48 ; Number # No  [9] KHAROSHTHI DIGIT ONE..KHAROSHTHI FRACTION ONE HALF
```

### Collation order (allkeys.txt)

```
10A44 ;       # KHAROSHTHI NUMBER TEN
10A45 ;       # KHAROSHTHI NUMBER TWENTY
10A46 ;       # KHAROSHTHI NUMBER ONE HUNDRED
10A47 ;       # KHAROSHTHI NUMBER ONE THOUSAND
10A48 ;       # KHAROSHTHI FRACTION ONE HALF
…
10A00 ;       # KHAROSHTHI LETTER A
10A01 ;       # KHAROSHTHI VOWEL SIGN I
10A02 ;       # KHAROSHTHI VOWEL SIGN U
10A03 ;       # KHAROSHTHI VOWEL SIGN VOCALIC R
10A05 ;       # KHAROSHTHI VOWEL SIGN E
10A06 ;       # KHAROSHTHI VOWEL SIGN O
10A0C ;       # KHAROSHTHI VOWEL LENGTH MARK
10A10 ;       # KHAROSHTHI LETTER KA
```

```
10A32 ;        # KHAROSHTHI LETTER KKA
10A11 ;        # KHAROSHTHI LETTER KHA
10A12 ;        # KHAROSHTHI LETTER GA
10A13 ;        # KHAROSHTHI LETTER GHA
10A15 ;        # KHAROSHTHI LETTER CA
10A16 ;        # KHAROSHTHI LETTER CHA
10A17 ;        # KHAROSHTHI LETTER JA
10A19 ;        # KHAROSHTHI LETTER NYA
10A1A ;        # KHAROSHTHI LETTER TTA
10A34 ;        # KHAROSHTHI LETTER TTTA
10A1B ;        # KHAROSHTHI LETTER TTHA
10A33 ;        # KHAROSHTHI LETTER TTTHA
10A1C ;        # KHAROSHTHI LETTER DDA
10A1D ;        # KHAROSHTHI LETTER DDHA
10A1E ;        # KHAROSHTHI LETTER NNA
10A1F ;        # KHAROSHTHI LETTER TA
10A20 ;        # KHAROSHTHI LETTER THA
10A21 ;        # KHAROSHTHI LETTER DA
10A22 ;        # KHAROSHTHI LETTER DHA
10A23 ;        # KHAROSHTHI LETTER NA
10A24 ;        # KHAROSHTHI LETTER PA
10A25 ;        # KHAROSHTHI LETTER PHA
10A26 ;        # KHAROSHTHI LETTER BA
10A27 ;        # KHAROSHTHI LETTER BHA
10A28 ;        # KHAROSHTHI LETTER MA
10A29 ;        # KHAROSHTHI LETTER YA
10A2A ;        # KHAROSHTHI LETTER RA
10A2B ;        # KHAROSHTHI LETTER LA
10A2C ;        # KHAROSHTHI LETTER VA
10A35 ;        # KHAROSHTHI LETTER VHA
10A2D ;        # KHAROSHTHI LETTER SHA
10A2E ;        # KHAROSHTHI LETTER SSA
10A2F ;        # KHAROSHTHI LETTER SA
10A30 ;        # KHAROSHTHI LETTER ZA
10A31 ;        # KHAROSHTHI LETTER HA
10A3F ;        # KHAROSHTHI VIRAMA
```

N.B., the order of two existing signs (10A32 KHAROSHTHI LETTER KKA and 10A33 KHAROSHTHI LETTER TTTHA) has been corrected in the above sequence.

## References


Bailey, H. W. 1980. “A Kharoṣṭrī Inscription of Seṇavarma, King of Oḍi.” *The Journal of the Royal Asiatic Society of Great Britain and Ireland:* 21–9.

Baums, Stefan and Andrew Glass. 2002– a. *A Dictionary of Gāndhārī.* <https://gandhari.org/dictionary>

Baums, Stefan and Andrew Glass. 2002– b. *Catalog of Gāndhārī Texts.* <https://gandhari.org/catalog>

Boyer, A.-M, E. J. Rapson, E. Senart, and P. S. Noble. 1920–29. *Kharoṣṭhī Inscriptions Discovered by Sir Aurel Stein in Chinese Turkestan.* Oxford: Clarendon Press.

Burrow, Thomas. 1937. *The Language of the Kharoṣṭhi Documents from Chinese Turkestan.* Cambridge: University Press.

Ching, Chao-jung (慶昭蓉). 2013. “Qiūcí shíkū xiàncún tíjì zhōng de Guīzī guó wáng 龜茲石窟現存題記中的龜茲國王.” *Dūnhuáng Tǔlǔfān yánjiū* 敦煌吐魯番研究 13: 387–418.

———.2014. “Kèzīěr chūtǔ Dé cáng Qūlú wén Qiūcí wáng zhàoyù yǔ qìyuē wénshū yánjiū 克孜尔出土德藏佉卢文龟兹王诏谕与契约文书研究.” *Xīyù wénshǐ* 西域文史 9: 51–73.

Falk, Harry. 2001. “Names and Weights Inscribed on Some Vessels from the Silver Hoard.” *Journal des savants*: 308–19.

Glass, Andrew. 2000. “A Preliminary Study of Kharoṣṭhī Manuscript Paleography.” MA Thesis. University of Washington. <http://andrewglass.org/ma.php>.

Glass, Andrew, Stefan Baums, and Richard Salomon. 2002. “Proposal to Encode Kharoṣṭhī in Plane 1 of ISO/IEC 10646” L2/02-203-R2 http://www.unicode.org/L2/L2002/02203r2-kharoshthi.pdf.

Melzer, Gudrun. 2015. "Ein Alphabet-Akrostichon aus Gandhāra." *Akademie aktuell: Zeitschrift der Bayerischen Akademie der Wissenschaften* 53: 29–33.

Salomon, Richard. 2004. “An Arapacana Abecedary from Kara Tepe (Termez, Uzbekistan).” *Bulletin of the Asia Institute* 18: 43–51.

Strauch, Ingo. 2008. “The Bajaur Collection of Kharoṣṭhī Manuscripts – A Preliminary Survey.” *Studien zur Indologie und Iranistik* 25: 103–36.

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://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 Additional Characters for Kharoṣṭhī Script**
2. Requester's name: Andrew Glass, Stefan Baums
3. Requester type (Member body/Liaison/Individual contribution): Individual contribution
4. Submission date:
5. Requester's reference (if applicable):
6. Choose one of the following:
   - This is a complete proposal: Complete
   - (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: 10A00–10A5F
     - Name of the existing block: KHAROSHTHI
2. Number of characters in proposal: 3
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 | D | 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?
     Andrew Glass
   - b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
     asg @ uw . edu
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
     Shaping

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? 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? Stefan Baums
   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? No
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) Rare
   Reference:
5. Are the proposed characters in current use by the user community? Yes
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? No
   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)? Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? Yes
   If YES, is a rationale for its inclusion provided? Yes
   If YES, reference: See proposal document
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? No
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? No
    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? 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: