Title
Proposal to encode the Old Sogdian script in Unicode

Permalink
https://escholarship.org/uc/item/0934973g

Author
Pandey, Anshuman

Publication Date
2016-12-31

Peer reviewed
Proposal to encode the Old Sogdian script in Unicode

Anshuman Pandey
pandey@umich.edu

December 31, 2016

1 Introduction

This is a proposal to encode the ‘Old Sogdian’ script in Unicode. It is a significant revision of the following document:

• L2/15-089 “Preliminary Proposal to Encode the Old Sogdian Script in Unicode”

An ISO proposal summary form is attached. This proposal addresses comments made on previous versions in the following reports:

• L2/16-037 “Recommendations to UTC #146 January 2016 on Script Proposals”
• L2/17-037 “Recommendations to UTC #150 January 2017 on Script Proposals”

A proposed Unicode encoding for the later ‘Sogdian’ script has been presented in:

• L2/16-371R “Revised proposal to encode the Sogdian script in Unicode”

The present proposal has been reviewed by Nicholas Sims-Williams and Yutaka Yoshida, who are leading scholars of Sogdian studies.

2 Background

The proposed Unicode encoding for ‘Old Sogdian’ encompasses a group of related scripts used in the following records for representing Sogdian (ISO 639: sog), an ancient Eastern Iranian language:

• Kultobe inscriptions The oldest Sogdian records are stone inscriptions found at Kultobe, hereafter ‘K’, in modern Kazakhstan (see Sims-Williams & Grenet 1998; Grenet, et al 2007). Fourteen inscriptions have been discovered and studied (see figures 26, 27). They have not been concisely dated, but the archaic features of the script and language indicate that they precede the ‘Ancient Letters’.
Proposal to encode the Old Sogdian script in Unicode

Anshuman Pandey

• ‘Ancient Letters’ The earliest attested Sogdian manuscripts are known as the ‘Ancient Letters’ (see figures 28–35), hereafter ‘AL’. These paper documents were found in 1907 by Aurel Stein in Dunhuang, western China. Based upon internal evidence the ‘AL’ may have been written during 312–314 CE (Sims-Williams 1985; Grenet, et al 1998).

• Upper Indus inscriptions Sogdian text appears on more than 600 rock carvings at Shatial and other sites in the Gilgit region of Pakistan (see figures 36, 37). These ‘Upper Indus inscriptions’, hereafter ‘UII’, have been dated to the 4th–7th centuries CE (Sims-Williams 1989, 2000), and some more precisely to the latter half of the 5th century (Yoshida 2013).

• Short inscriptions on coins and vessels A script resembling that used in AL and UII is used for inscriptions on coins and vessels from the ancient principality of Chach, situated around modern Tashkent, Uzbekistan, and surrounding areas (see figure 39).

The scripts of these records are derived from Imperial Aramaic and exhibit the following features:

• Repertoire Of the 22 letters of the Aramaic alphabet, 20 are attested in the repertoires of these scripts. Analogues for teth and qoph do not exist. Of these 20, 17 have distinctive representations, while 3 share a resemblance. In AL and UII, the shapes of daleth and ayin are in general identical to resh, but may be distinctive in K. The letter taw has a unique final form in K. All 20 letters are exhibited in K 4 and occur collectively in AL. The AL contain additional letters that do not occur in K, such as distinctive final forms of aleph, beh, nun, sadhe, taw; special forms of ayin; and a new form of he (see § 3.1). Numerical signs are attested in AL and UII, but not in K.

• Letterforms The shapes of letters in AL and UII are nearly identical. The letterforms of K are more archaic and reflect constraints imposed by the method and medium of inscription. The shapes of gimel, he, yodh, lamedh, shin in the three varieties differ from the Aramaic originals and corresponding letters in related Iranian scripts. They may be considered characteristically ‘Sogdian’. The special forms of ayin in AL do not occur in K or UII, or in any other script. A comparison of letters in related scripts is shown in table 1 and figure 42.

• Structure Each variety is a non-joining abjad, similar to Hebrew. Letters retain their shapes within a word. Some letters have distinctive word-final forms, but there are no formal conventions for their usage. The strokes of adjacent letters of a word may connect or overlap as the result of cursive writing. This type of conjunction differs from that of later ‘formal’ and ‘cursive’ Sogdian scripts, which possess intrinsic conjoining behaviors similar to Arabic, as shown below:

<table>
<thead>
<tr>
<th>Old Sogdian</th>
<th>Later Sogdian</th>
</tr>
</thead>
<tbody>
<tr>
<td>swγδyk</td>
<td>‘Sogdian’</td>
</tr>
<tr>
<td>smʾrknδc</td>
<td>‘of Samarkand’</td>
</tr>
</tbody>
</table>

• Directionality These old Sogdian varieties are written from right to left in lines that advance from top to bottom. Some UII are written vertically with letters rotated 90° counter-clockwise with lines that advance from left to right (see § 4.5).
These scripts may be considered typologically identical on the basis of their graphical and structural features. For purposes of character encoding they may be unified within a single Unicode script block. Using this approach texts would be represented using the same character set, but the display would be managed through the selection of fonts designed specifically for the K, AL, and UII varieties.

The proposed Unicode block is named ‘Old Sogdian’. This identifier has been selected because proper names do not exist for individual script varieties or for the family. The script of AL has been referred to as “Sogdian Aramaic” (Skjærvø 1996), which may be applied equally to the other two varieties. However, the descriptor ‘Aramaic’ is not used in Unicode names for other scripts descended from Aramaic. The bare name ‘Sogdian’ is used in the catalogue of the International Dunhuang Project for referring to both early and later script varieties. It is, however, practical to reserve this name for a Unicode block for the more well-known ‘formal’ and ‘cursive’ styles, which have been proposed for encoding in a unified ‘Sogdian’ block (see L2/16-371). The designation ‘Old Sogdian’ suitably identifies these early varieties while emphasizing their genetic relationship with later ‘Sogdian’ script styles.

3 Character Repertoire

The proposed repertoire contains 40 characters: 29 letters, 10 numbers, 1 heterogram. Names for letters correspond to those of the ‘Imperial Aramaic’ block. Representative glyphs are based upon forms in the AL unless specified below. The encoded set may differ from traditional and scholarly inventories of script varieties that occur in written and inscriptional sources. Such differences naturally arise from the requirements for digitally representing a script in plain text and for preserving the semantics of characters.

In this document, names in italics refer to scholarly names for graphemes while names in small capitals refer to proposed Unicode characters, eg. Aleph is aleph and old sogdian letter aleph. For sake of brevity, the descriptor ‘old sogdian’ is dropped when referring to Old Sogdian characters, eg. old sogdian letter aleph is referred to as aleph. Characters of other scripts are designated by their full Unicode names. Latin transliteration of Old Sogdian letters follows the scholarly convention. Aramaic heterograms are transliterated using the corresponding uppercase letters, with some exceptions as shown in the table below.

3.1 Letters

<table>
<thead>
<tr>
<th>Glyph</th>
<th>Character name</th>
<th>Latin</th>
</tr>
</thead>
<tbody>
<tr>
<td>ﻁ</td>
<td>old sogdian letter aleph</td>
<td>’</td>
</tr>
<tr>
<td>ﻁ</td>
<td>old sogdian letter final aleph</td>
<td>’</td>
</tr>
<tr>
<td>ﻁ</td>
<td>old sogdian letter beth</td>
<td>β; B</td>
</tr>
<tr>
<td>ﻁ</td>
<td>old sogdian letter final beth</td>
<td>β; B</td>
</tr>
<tr>
<td>ﻁ</td>
<td>old sogdian letter gimel</td>
<td>γ; G</td>
</tr>
<tr>
<td>ﻁ</td>
<td>old sogdian letter he</td>
<td>h</td>
</tr>
<tr>
<td>ﻁ</td>
<td>old sogdian letter final he</td>
<td>h</td>
</tr>
</tbody>
</table>
3.1.1 Notes on letters

*aleph* In word-final positions in AL, *aleph* is written as ◐ FINAL ALEPH, in which the horizontal stroke at the baseline is elongated. The letter ◐ ALEPH has the shape ◐ in K. This form is a glyphic variant. See figure 1 for attestations.
In word-final positions in AL, ꜰbeth is written as Ꜳ final beth, in which the horizontal stroke at the baseline is elongated. See figure 2 for attestations.

gimel See figure 3 for attestations of ꜱgimel.

daleth The letter ꜱdaleth occurs only in Aramaic heterograms, eg. ꜱARAMAIC LETTER DALETH 'BDt (K 4.1) and ꜲARAMAIC LETTER RESH 'D (AL 2.1). In AL, the shape of daleth is identical to resh. In K, there is a possibility that daleth is distinguished from resh. See figure 4 for attestations. The issue regarding the shape of daleth is inherited from Aramaic, in which ꜱARAMAIC LETTER DALETH and ꜲARAMAIC LETTER RESH are nearly identical. Despite the possibility of a distinctive daleth in K, there is insufficient information at this time for defining it as a separate character. For this reason, daleth is unified with resh and is to be represented using ꜳRESH-AYN-DALETH.

he The letter ꜰhe is represented by both ꜱhe and Ꜳfinal he. The ꜱhe occurs only in K, eg. ꜱSWRH and ꜱARAMAIC LETTER HE (H)WH (K 2); ꜱARAMAIC LETTER KINOTH and ꜱARAMAIC LETTER KINOTH ZNH (K 4). The representative glyph is derived from K 2. In AL, he is written as ꜱ, eg. ꜱARAMAIC LETTER KINOTH (AL 2.12), ꜱARAMAIC LETTER KINOTH ZNH (AL 2.10). The Ꜳfinal he is not attested in K, and in contrast to ꜱhe, occurs only at the end of a word. It has the variant shape ꜱ, in which the terminal is truncated or merged back into the baseline stroke. See figure 5 for attestations.

waw See figure 6 for attestations of Ꜷwaw.

zayin See figures 7 and 8 for attestations of ꜱzayin.

heth See figure 9 for attestations of ꜱheth.

teth An Old Sogdian analogue for Aramaic teth does not exist. In K, the teth in Aramaic heterograms is represented using ꜱARAMAIC LETTER TETH QTLS is written as ꜱARAMAIC LETTER TETH KTLS (K 3.3).

yodh See figure 10 for attestations of Ꜷyodh.

kaph See figure 11 for attestations of ꜱkaph.

lamedh The letter ꜱlamedh has the shape ꜱ in K and Ꜳ in AL (see figure 12). The AL form is the representative glyph. In AL 5, lamedh appears as Ꜳ. Differences between Ꜳ, ꜱ, Ꜳ are stylistic, not semantic. The forms ꜱ and Ꜳ are to be treated as glyphic variants of ꜱ.

mem See figure 13 for attestations of ꜱmem.

nun Occurrences of nun are represented using ꜱnun, Ꜳfinal nun, Ꜳfinal nun with vertical tail (see figure 14). The representative glyph ꜱ for nun is derived from K. The final forms occur only in AL. While nun has the distinctive shape ꜱ in K, it has the shape Ꜳ in AL when non-final, which is identical to ꜱzayin, eg. ꜱARAMAIC LETTER KINOTH ZNH (K 4.1) and ꜱARAMAIC LETTER KINOTH ZNH (AL 2.10). When word-final in AL, nun is written as both ꜱ and Ꜳ, eg. ꜱARAMAIC LETTER KINOTH MN (AL 2.2) and ꜱARAMAIC LETTER KINOTH MN (AL 2.6). The regular and final forms are contrastive in AL (see figure 8). They are not glyphic variants. All three characters are required for fully representing nun in plain text.

samekh The letter ꜱsamekh occurs as the two-part form Ꜳ in K 4. This archaic form is to be treated as a glyphic variant. See figure 15 for attestations.

ayin The letter ꜱayin occurs only in Aramaic heterograms. It has the regular shape ꜱ and the special shapes ꜱ and Ꜳ (see figure 16). The regular ꜱayin occurs in both K and AL, eg. ꜱARAMAIC LETTER AYIN 'BDT (K 4.1), ꜱARAMAIC LETTER AYIN D.
'LZK (AL 2.12), ḤRZYnnc (AL 6.6). In AL, the shape of regular ayin is identical to resh (and daleth). In K, there is a possibility that ayin might be a distinctive letter. The similarity between ayin and resh is inherited from Aramaic, compare ʿ ṣ ARAMAIC LETTER AYIN and .CreateCommandLabel y ARAMAIC LETTER RESH. However, there is insufficient information for determining whether or not the differences between ayin and resh in K are semantically significant. Therefore, a separate character for regular ayin is not proposed at present. It is to be represented using  y RESH-AYIN-DALETH. The letters ḥ AJIN and ʾ ALTERNATE AYIN occur only in AL for writing the heterogram 'D, eg. ʾ ṣ (AL 2.1), ʾ ṢO (AL 3 verso), ʾ ᵘ (AL 3.1), ʾ ᶘ (AL 5.1). Although ḥ, ʾO, ʾQ, ʾ could be considered glyphic variants of a single character ḥ AJIN, it is appropriate to define two characters on account of their graphical structures. The ᵘ is a glyphic variant of ʾO with an ornate tail; the ʾO is a variant with a truncated tail. These three forms are unified as ʾ ALTERNATE AYIN, which may be used for representing these special forms in plaintext. See figure 25 and § 3.3 for attestations.

pe The letter ʾ PE is has the variant ‘open’ shape ʾO, which is a glyphic variant (see figure 17).

sadhe This letter is represented using ḥ SADHE, ḥ FINAL SADHE, and ḥ FINAL SADHE WITH VERTICAL TAIL (see figure 18). The final forms occur only in AL. In AL 2, sadhe has the shape ḥ whenever it occurs at the margin, eg. ᵖHRZYnnc ḥHRZYnnc (AL 2.54). In other positions within a line, final sadhe is written using ḥ FINAL SADHE, eg. ᵖHRZYnnc ᶘHRZYnnc (AL 2.34). Such usage suggests a possible convention for the contrastive depiction of sadhe within and at the end of line. A curved variant ḥ of ḥ is attested. All three characters are required for fully representing sadhe in plain text.

qoph An Old Sogdian analogue for Aramaic qoph does not exist. In K, the qoph in Aramaic heterograms is represented using ʿ QAPH: ṢQTLṭ is written as ʿ ṢQTLṭ (K 3.3). It used to be believed that ʿ QOPH was retained in AL as ʿ and reassigned for the number 100. This ʿ is now identified as the fraction ½ (Grenet, et al 1998).

resh In AL, the letter ʿ is used for resh, daleth, and ayin (see figure 19). According to the Unicode character-glyph model, letters with identical glyphic representations are considered variants and are unified as a single character. As the sound [r] represented by resh is phonemic in Sogdian, and those represented by ayin and daleth are not, the letter ʿ is used ubiquitously for resh. Accordingly, daleth and ayin are unified with resh as ʿ RESH-AYIN-DALETH. This approach follows the Unicode model for Inscriptional Pahlavi, in which waw, ayin, resh are represented using ʿ U+10B65 INSCRIPTIONAL PAHLAVI LETTER WAW-AYIN-RESH; and similarly, mem and qoph using ʿ U+10B6C INSCRIPTIONAL PAHLAVI LETTER MEM-QOPH. Despite occurring after daleth and ayin in the alphabetical order, resh is ordered first in the name RESH-AYIN-DALETH because it occurs more frequently in the sources; daleth is ordered before ayin for the same reason.

shin See figure 20 for attestations of ʾ SIN.

taw This letter is represented using ʾ TAW, ʾ FINAL TAW, ʾ FINAL TAW WITH VERTICAL TAIL (see figure 21). Usage of the nominal and final forms is contrastive in both K and AL. In K 4, taw appears as ʾ and final taw as ʾ. These archaic two-part forms are to be treated as glyphic variants of TAW and FINAL TAW WITH VERTICAL TAIL, respectively. The ʾ FINAL TAW is often written using a glyphic variant with a curved tail ʾ in AL. All three characters are required for fully representing taw in plain text.

3.1.2 Note on final forms

Distinctive final forms of aleph, beth, nun, sadhe, taw are included in the repertoire as separate characters. These final forms differ from the nominal forms in the shape of their terminals, which are elongated horizontally or which descend vertically. An analysis of AL indicates that final forms are regularly used at the
end of words, and that some final forms are used specifically at the end of line. The analysis also suggests
an intentional differentiation between nominal and final forms of only these five letters. For instance, the
elongated baseline in the final form א of א aleph and the final form ב ofbeth may be interpreted as a
natural flourish made by the scribe at the end of a word. But, such strokes occur consistently with final forms
of these two letters across the AL corpus, and are not simply stylistic. On the other hand, commonly occur-
ing letters such as ו waw, י yodh, ר resh have curved terminal strokes that present a natural opportunity for
stylistic elongation at end of a word. However, there appears to be deliberate avoidance of such flourishes
when writing these letters in final position. These letters, in turn, may be compared to כ kaph and ט pe,
whose shapes inherently possess an elongated tail that is often extended in final position, and which may be
considered a natural stylistic flourish.

In addition to illustrating distinctive final forms for aleph, beth, nun, sadhe, taw, the available sources also
point to the existence of two types of final forms for nun, sadhe, taw. These three letters occur in word-final
position with either an elongated horizontal stroke or with a descending vertical tail. There is some evidence
to suggest contrastive contextual usage of the two types. For instance, word-final ס sadhe is written in AL 2
with a vertical tail נ whenever it occurs at the margin, and with a horizontal tail ס in other positions within
a line (see figure 18). Throughout the AL corpus, מ nun is written using both מ and נ at the ends of words and
lines. The same applies to the usage of ס and נ for final מ taw. These final forms of nun and taw appear
to be used interchangably and occur on the same line or in close proximity. It may be possible that one form
was intentionally selected over the other based upon spacing requirements along a line. For instance, a scribe
may have chosen the form with a horizontal tail to fill space, or the form with vertical tail to compensate for
lack of space. The usage of both forms within the same source suggests that scribes perceived of a semantic
distinction between the horizontal and vertical final forms for nun, sadhe, taw.

It is difficult to ascertain the nature of Sogdian scribal conventions that were in vogue in the early 4th century
c.e., when the AL were written. There are no sources that provide descriptions of orthographic rules or
explanations for the existence of final forms for only five letters of the repertoire. There are no clues that
offer insights into the development of two final forms for nun, sadhe, taw; or, that specify the rationale for
their usage or the criteria for a scribe’s preference of one form over the other in a given context. The available
sources simply show that both final forms are used for these three letters.

For this reason, the two final forms for nun, sadhe, taw have been included as separate characters in the
proposed repertoire. Without knowledge of the conventions for usage of the two forms, it is impractical to
exclude one set from the repertoire. Moreover, given that there is some evidence to suggest scribal prefer-
ences for a particular form in a given context, it is improper to consider the forms as stylistic variants instead
of semantic alternates. Furthermore, in terms of the Unicode character-glyph model it is difficult to specify
which of final מ/ נ nun, ס/ נ sadhe, מ/ ט taw would be the ‘normative’ final and which is the ‘variant’,
or if such a model may be used in evaluating these forms. Therefore, rather than attempt to define a modern
convention for the representation of final forms for three letters of a historical script, it is advantageous to
develop a repertoire that enables representation of the script in plain text as it appears in the extant corpus.

When developing a Unicode encoding for an ancient script such as Old Sogdian, it is most practical to permit
the extant sources to guide the process. This is especially important when there is an absence of knowledge
regarding a particular orthographic convention in such a script. It is from this perspective that the proposed
repertoire for Old Sogdian has been defined. The proposal author has discussed the issue with scholars, who
will be the primary users of the encoding. These experts have expressed a requirement for representing in
plain text both vertical and final forms of final nun, sadhe, and taw as they occur in the sources in order to
accurately and completely digitize Sogdian records.
3.2 Numbers

The repertoire contains 10 numerical characters. These occur in AL and UII, but not in the extant K sources. See figures 22–24 for attestations.

<table>
<thead>
<tr>
<th>Glyph</th>
<th>Character name</th>
<th>Numeric value</th>
</tr>
</thead>
<tbody>
<tr>
<td>𐼝</td>
<td>OLD SOGDIAN NUMBER ONE</td>
<td>1</td>
</tr>
<tr>
<td>𐼞</td>
<td>OLD SOGDIAN NUMBER TWO</td>
<td>2</td>
</tr>
<tr>
<td>𐼟</td>
<td>OLD SOGDIAN NUMBER THREE</td>
<td>3</td>
</tr>
<tr>
<td>𐼠</td>
<td>OLD SOGDIAN NUMBER FOUR</td>
<td>4</td>
</tr>
<tr>
<td>𐼡</td>
<td>OLD SOGDIAN NUMBER FIVE</td>
<td>5</td>
</tr>
<tr>
<td>但不限</td>
<td>OLD SOGDIAN NUMBER TEN</td>
<td>10</td>
</tr>
<tr>
<td>但不限</td>
<td>OLD SOGDIAN NUMBER TWENTY</td>
<td>20</td>
</tr>
<tr>
<td>但不限</td>
<td>OLD SOGDIAN NUMBER THIRTY</td>
<td>30</td>
</tr>
<tr>
<td>但不限</td>
<td>OLD SOGDIAN NUMBER ONE HUNDRED</td>
<td>100</td>
</tr>
<tr>
<td>但不限</td>
<td>OLD SOGDIAN FRACTION ONE HALF</td>
<td>½</td>
</tr>
</tbody>
</table>

**Primary units**  The primary units are expressed using joined repetitions of the sign 𐼝 that are generally grouped in sets of three or four and separated by spaces, eg. 𐼞 for 2, 𐼟 for 3, 𐼠 𐼠 for 8. As the script is non-conjoining, no simple method exists for representing the ligated repetitions of 𐼝. For that reason, the numbers 但不限 one .. 但不限 four are encoded atomically. This model for one .. four follows the Unicode encoding for Inscriptional Parthian, eg. 𐼝 U+10B58 INSCRIPTIONAL PARTHIAN NUMBER ONE .. 𐼠 U+10B5B INSCRIPTIONAL PARTHIAN NUMBER FOUR. The numbers 5–9 are written using sequences of one .. four arranged in groups separated by spaces. The number 5 may also be represented using the character 但不限 five, which is attested as a single unit in AL 7 (see figure 23). Representations of all primary numbers are shown in the table below.

**Tens**  The 但不限 TEN resembles a vertically compressed 𐼝 LAMEDH. The shapes for 但不限 TWENTY and 但不限 THIRTY are formed from vertical stacks of 但不限 TEN. Multiples of 10 greater than 20 are produced using appropriate repetitions and groupings of ten and twenty. The character thirty is not commonly used in compound numbers. The number 30 may be also represented as 但不限 但不限, which is a compound of twenty and ten.

**Hundreds**  The number 100 is written using 但不限 ONE HUNDRED. The glyph resembles the letter 但不限 GIMEL above a serpentine form, but it is an atomic character. The one hundred also functions as a unit mark for the hundreds. Multiples of hundred are represented by prefixing the appropriate groupings of one .. four before one hundred.
Thousands   The number 1000 is expressed using the Aramaic heterogram 𐤋𐤇𐤍𐤊𐤊 1LP, which is represented using the sequence < weekdays, 𐤋 LAMEDH, 𐤍 PE>. The sequence 𐤋𐤇𐤍𐤊𐤊 also functions as a unit mark for the thousands. The 𐤋 weekdays is an inherent part of the 1LP unit. Multiples are expressed by prefixing primary numbers before the unit, eg. 2000 is 𐤋𐤇𐤍𐤊𐤊 𐤋, 3000 is 𐤋𐤇𐤍𐤊𐤊 𐤋 𐤋 𐤋. 

Ten thousands   The number 10000 is expressed using the Sogdian word 𐤆𐤇𐤌𐤊𐤌 ƅyrwr. There is no distinctive numerical sign for this value.

Fraction   The 𐤉 FRACTION ONE HALF is placed after another numerical character.

3.2.1 Notation system

The ordering of numbers follows the right-to-left directionality of the script. The expression of numbers is additive. Compounds of different units are produced by placing larger units first. However, in some inscriptions on silver coins the units precede the tens (see Livshits 2015: 234), which follows the order of spoken numbers. Spaces are used for separating groups of primary numbers.

<table>
<thead>
<tr>
<th>Value</th>
<th>Number</th>
<th>Input string →</th>
</tr>
</thead>
<tbody>
<tr>
<td>4½</td>
<td>ﻦ‫‪&lt; yyyy FOUR, ﻦ FRACTION ONE HALF&gt;</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ﻦ‫‪&lt; yyyy THREE, ﻦ SPACE, yyyy TWO&gt;</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ﻦ‫‪&lt; yyyy FIVE&gt;</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ﻦ‫‪&lt; yyyy THREE, ﻦ SPACE, yyyy THREE&gt;</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ﻦ‫‪&lt; yyyy THREE, ﻦ SPACE, yyyy THREE&gt;</td>
<td></td>
</tr>
<tr>
<td>7½</td>
<td>ﻦ‫‪&lt; yyyy FOUR, ﻦ SPACE, yyyy THREE, ﻦ FRACTION ONE HALF&gt;</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>ﻦ‫‪&lt; yyyy FOUR, ﻦ SPACE, yyyy FOUR&gt;</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ﻦ‫‪&lt; yyyy THREE, ﻦ SPACE, yyyy THREE, ﻦ SPACE, yyyy THREE&gt;</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>ﻦ‫‪&lt; ﻦ TEN, yyyy THREE&gt;</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ﻦ‫‪&lt; ﻦ TEN, yyyy FIVE&gt;</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>𐤊‫‪&lt; ﻦ THIRTY&gt;</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>𐤊‫‪&lt; ﻦ TWENTY, ﻦ TEN&gt;</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>𐤊‫‪&lt; ﻦ TWENTY, ﻦ TEN, yyyy TWO&gt;</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>ﻦ‫‪&lt; ﻦ ONE HUNDRED&gt;</td>
<td></td>
</tr>
</tbody>
</table>
Proposal to encode the Old Sogdian script in Unicode
Anshuman Pandey

<table>
<thead>
<tr>
<th>Value</th>
<th>Number</th>
<th>Input string →</th>
</tr>
</thead>
<tbody>
<tr>
<td>2453</td>
<td>ƒ ဆ 럼 လင်မ ကြည် သင်ကြည် သင် ခံသည်</td>
<td>&lt; န TWO, igos SPACE, လ ONE, လ LAMEDH, ဝ PE, gos SPACE, င THREE, ယ ONE, လ LAMEDH, ဝ PE&gt;</td>
</tr>
<tr>
<td>2453</td>
<td>ƒ ဆ လင်မ ကြည် သင်ကြည် သင် ခံသည်</td>
<td>&lt; န TWO, igos SPACE, လ ONE, လ LAMEDH, ဝ PE, gos SPACE, င THREE, ယ ONE, လ LAMEDH, ဝ PE&gt;</td>
</tr>
</tbody>
</table>

Attestations for the above numbers are shown in figures 22–24. The repertoire provides for the presentation of any numerical value, even if not attested. For example, the number 2453 could be represented as:

<table>
<thead>
<tr>
<th>Value</th>
<th>Number</th>
<th>Input string →</th>
</tr>
</thead>
<tbody>
<tr>
<td>2453</td>
<td>ʃ ဆ 럼 လင်မ ကြည် သင်ကြည် သင် ခံသည်</td>
<td>&lt; န TWO, igos SPACE, လ ONE, လ LAMEDH, ဝ PE, gos SPACE, င THREE, ယ ONE, လ LAMEDH, ဝ PE&gt;</td>
</tr>
</tbody>
</table>

3.3 Heterogram

The repertoire contains 1 heterogram.

<table>
<thead>
<tr>
<th>Glyph</th>
<th>Character name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>InputGroup: [ais]</td>
<td>OLD SOGDIAN HETEROGRAM AYIN-DALETH</td>
<td>'D</td>
</tr>
</tbody>
</table>

Aramaic heterograms are represented as words spelled using conventional letters, e.g. ʿHRZY is written <Aleph, Heth, Resh, Zayin, Yodh>. The heterogram 'D is the sole exception. Meaning “to”, 'D occurs in the address and salutation of a letter, e.g. ʿD βγw xwt'w βρrkw “to lord master Barak”. Morphologically, it is comprised of ayin and daleth. Yet, instead of the expected spelling ʿD βγw xwt'w βρrkw, the ayin is written using special forms: ʿ, ṭ, ʿo, ʿo, ʿo, ʿo (see figure 25). An explanation for this curious orthography may be that ayin and daleth had disappeared from the script by the time of AL, and the original phonetic values of these letters never existed in Sogdian. Therefore, scribes were unaware of these letters and of the original spelling of the Aramaic word, so they stylized the writing of 'D (Sims-Williams, personal correspondence, 2016).

There are two ways to analyze these representations of 'D. First, as a conventional word comprised of the letters ayin and daleth. These forms of ayin, which occur only in this heterogram, are included in the repertoire as ayin and alternate ayin; the ʿ and ʿo could be considered glyphic variants of alternate...
AYIN. Accordingly, ‘D may be represented as `<AYIN | ALTERNATE AYIN, RESH-AYIN-DALETH>`. Secondly, ‘D is a logographic unit comprised of a ligature or a set of two letters. This unit may be treated as an atomic character, eg. ṭ ః OLD SOGDIAN HETEROGRAM AYIN-DALETH. These approaches are not mutually exclusive and both are practical for character encoding. Depending upon the context, ‘D may be spelled using a sequence of letters or represented using an atomic character.

The case of ‘D is similar to the Latin ‘&’ ampersand. The ‘&’ represents the Latin word et “and”. Morpho-logically, it is a ligation of the Latin letters ‘e’ and ‘t’, eg. ेर, ḍ. The base letters began to be obscured as the ligature became more stylized, eg. े&. The logographic nature of ‘&’ is apparent in the abbreviation “&c” for Latin et cetera “and so forth”, where it masks ‘et’. Latin et can be represented both using the sequence <e, t> and atomic characters, such as ṭ u+1F670 SCRIPT LIGATURE ET ORNAMENT.

The character name for HETEROGRAM AYIN-DALETH is derived from the conventional transliteration ‘D of the heterogram. The representative glyph ṭ is derived from AL 3 and has been selected because it is structurally a ligature. Variant forms may be managed through fonts.

4 Script Details

4.1 Bidirectional model

Old Sogdian may be implemented using the Unicode Bidirectional Algorithm. There are no requirements for shaping.

4.2 Punctuation

Punctuation marks are not attested. Words are separated using spaces in K and AL. Inter-word spacing is inconsistent in the UII.

4.3 Line-breaking

There are no rules for line-breaking. The available sources show line-breaks after the end of a word. Word are not split across lines. Consequently, hyphens or other continuation marks are not attested. In digital layouts, line-breaks may occur after any character.

4.4 Collation

The sort order for Old Sogdian is as follows:

<table>
<thead>
<tr>
<th>ALEPH</th>
<th>FINAL ALEPH</th>
<th>BETH</th>
<th>FINAL BETH</th>
<th>GIMEL</th>
<th>HE</th>
</tr>
</thead>
<tbody>
<tr>
<td>గ</td>
<td>ఘ</td>
<td>ఙ</td>
<td>చ</td>
<td>ఛ</td>
<td>జ</td>
</tr>
<tr>
<td>FINAL HE</td>
<td>WAW</td>
<td>ZAYIN</td>
<td>HETH</td>
<td>YODH</td>
<td>KAPH</td>
</tr>
<tr>
<td>MEM</td>
<td>NUN</td>
<td>FINAL NUN</td>
<td>FINAL NUN WITH VERTICAL TAIL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEKH</td>
<td>AYIN</td>
<td>ALTERNATE AYIN</td>
<td>HETEROGRAM AYIN-DALETH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE</td>
<td>SADHE</td>
<td>FINAL SADHE</td>
<td>FINAL SADHE WITH VERTICAL TAIL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.5 Vertical text

The majority of Old Sogdian records have horizontal orientations. Some UII records are inscribed vertically (Yoshida 2013). There are no formal conventions for text orientation. However, in vertical environments, Old Sogdian text is oriented from top to bottom with lines that advance from left to right. Letters are rotated 90° counter-clockwise from their regular upright shapes.

By default, Old Sogdian may be oriented horizontally in plain text representations. However, support for vertical orientations of the script is required for accurately displaying Old Sogdian text that is natively vertical. Below is a vertical text from Shatial rock 36:38 (see figure 38) and its horizontal representation:

Vertical (rotated 90° CCW)  

```
 Vertical Orientation=R or vo=R, where the value 'R' indicates that the glyphs are rotated in vertical layout. The rotation is 90° counter-clockwise.
```

"(I), Nanai-vandak the (son of) Narisaf have come (here) in/on the (day/year) ten and asked a boon from the spirit of the sacred place Kârt (that) I may arrive at Kharvandan (= Tashkurgan) very quickly and see (my) brother in good (health) with joy." (Yoshida 2013: 379–380).

The “Unicode Technical Report #50: Unicode Vertical Text Layout” describes the Vertical_Orientation (vo) property for specifying the orientation of characters in vertical environments. For Old Sogdian, this property would be defined as: Vertical_Orientation=R or vo=R, where the value ‘R’ indicates that the glyphs are rotated in vertical layout. The rotation is 90° counter-clockwise.
4.6 Character Data

4.6.1 Character properties

In the format of `UnicodeData.txt`:

```
10F00;OLD SOGDIAN LETTER ALEPH;Lo;0;R;;;;;N;;;;;
10F01;OLD SOGDIAN LETTER FINAL ALEPH;Lo;0;R;;;;;N;;;;;
10F01;OLD SOGDIAN LETTER BETH;Lo;0;R;;;;;N;;;;;
10F03;OLD SOGDIAN LETTER FINAL BETH;Lo;0;R;;;;;N;;;;;
10F04;OLD SOGDIAN LETTER GIMEL;Lo;0;R;;;;;N;;;;;
10F05;OLD SOGDIAN LETTER HE;Lo;0;R;;;;;N;;;;;
10F06;OLD SOGDIAN LETTER FINAL HE;Lo;0;R;;;;;N;;;;;
10F07;OLD SOGDIAN LETTER WAW;Lo;0;R;;;;;N;;;;;
10F09;OLD SOGDIAN LETTER HETH;Lo;0;R;;;;;N;;;;;
10F0A;OLD SOGDIAN LETTER YODH;Lo;0;R;;;;;N;;;;;
10F0B;OLD SOGDIAN LETTER KAPH;Lo;0;R;;;;;N;;;;;
10F0C;OLD SOGDIAN LETTER LAMEDH;Lo;0;R;;;;;N;;;;;
10F0D;OLD SOGDIAN LETTER MEM;Lo;0;R;;;;;N;;;;;
10F0E;OLD SOGDIAN LETTER NUN;Lo;0;R;;;;;N;;;;;
10F0F;OLD SOGDIAN LETTER FINAL NUN;Lo;0;R;;;;;N;;;;;
10F11;OLD SOGDIAN LETTER SAMEKH;Lo;0;R;;;;;N;;;;;
10F12;OLD SOGDIAN LETTER AYIN;Lo;0;R;;;;;N;;;;;
10F13;OLD SOGDIAN LETTER ALTERNATE AYIN;Lo;0;R;;;;;N;;;;;
10F14;OLD SOGDIAN LETTER PE;Lo;0;R;;;;;N;;;;;
10F15;OLD SOGDIAN LETTER SADHE;Lo;0;R;;;;;N;;;;;
10F16;OLD SOGDIAN LETTER FINAL SADHE;Lo;0;R;;;;;N;;;;;
10F17;OLD SOGDIAN LETTER FINAL SADHE WITH VERTICAL TAIL;Lo;0;R;;;;;N;;;;;
10F18;OLD SOGDIAN LETTER RESH-AYIN-DALETH;Lo;0;R;;;;;N;;;;;
10F19;OLD SOGDIAN LETTER SHIN;Lo;0;R;;;;;N;;;;;
10F1A;OLD SOGDIAN LETTER TAW;Lo;0;R;;;;;N;;;;;
10F1B;OLD SOGDIAN LETTER FINAL TAW;Lo;0;R;;;;;N;;;;;
10F1C;OLD SOGDIAN LETTER FINAL TAW WITH VERTICAL TAIL;Lo;0;R;;;;;N;;;;;
10F1D;OLD SOGDIAN NUMBER ONE;No;0;R;;;;1;N;;;;;
10F1E;OLD SOGDIAN NUMBER TWO;No;0;R;;;;2;N;;;;;
10F1F;OLD SOGDIAN NUMBER THREE;No;0;R;;;;3;N;;;;;
10F20;OLD SOGDIAN NUMBER FOUR;No;0;R;;;;4;N;;;;;
10F21;OLD SOGDIAN NUMBER FIVE;No;0;R;;;;5;N;;;;;
10F22;OLD SOGDIAN NUMBER TEN;No;0;R;;;;10;N;;;;;
10F23;OLD SOGDIAN NUMBER TWENTY;No;0;R;;;;20;N;;;;;
10F24;OLD SOGDIAN NUMBER THIRTY;No;0;R;;;;30;N;;;;;
10F25;OLD SOGDIAN NUMBER ONE HUNDRED;No;0;R;;;;100;N;;;;;
10F26;OLD SOGDIAN FRACTION ONE HALF;No;0;R;;;;1/2;N;;;;;
10F27;OLD SOGDIAN HETEROGRAM AYIN-DALETH;Lo;0;R;;;;N;;;;;
```

4.6.2 Linebreaking

In the format of `LineBreak.txt`:

```
10F00..10F1C;AL # Lo [29] OLD SOGDIAN LETTER ALEPH..
OLD SOGDIAN LETTER FINAL TAW WITH VERTICAL TAIL
10F1D..10F26;AL # No [10] OLD SOGDIAN NUMBER ONE..OLD SOGDIAN FRACTION ONE HALF
10F27;AL # Lo OLD SOGDIAN HETEROGRAM AYIN-DALETH
```
5 References


6 Acknowledgments

I express my deep gratitude to Nicholas Sims-Williams (SOAS, University of London) for providing detailed comments on earlier versions of this proposal and for informative discussions regarding all facets of the script. I am also thankful to Yutaka Yoshida (University of Kyoto) for reviewing versions of this proposal and for providing valuable feedback. I thank them both for their patient responses to my numerous inquiries and for overlooking my ignorance of the script. I am grateful to Roozbeh Pournader (Google, San Francisco) for discussing Unicode encodings for Iranian scripts and for his feedback on the earliest draft proposal.

The present proposal was funded in part by the Adopt-A-Character Program of the Unicode Consortium. A previous version was made possible in part through a Google Research Award, granted to Deborah Anderson for the Script Encoding Initiative, which funded a post-doctoral research position for me in the Department of Linguistics, University of California, Berkeley during 2015–2016. Preliminary research was made possible through the Script Encoding Initiative at Berkeley. Any views, findings, conclusions or recommendations expressed in this publication do not necessarily reflect those of the Unicode Consortium; the University of California, Berkeley; or Google.
This block unifies the scripts used in the Ancient Letters and the Kultobe and Upper Indus inscriptions.

### Letters

| U+10F00 | U+10F01 | U+10F02 | U+10F03 | U+10F04 | U+10F05 | U+10F06 | U+10F07 | U+10F08 | U+10F09 | U+10F0A | U+10F0B | U+10F0C | U+10F0D | U+10F0E | U+10F0F | U+10F10 | U+10F11 | U+10F12 | U+10F13 | U+10F14 | U+10F15 | U+10F16 | U+10F17 | U+10F18 | U+10F19 | U+10F1A | U+10F1B | U+10F1C | U+10F1D | U+10F1E | U+10F1F |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 𐼀     | 𐼁     | 𐼂     | 𐼃     | 𐼄     | 𐼅     | 𐼆     | 𐼇     | 𐼈     | 𐼉     | 𐼊     | 𐼋     | 𐼌     | 𐼍     | 𐼎     | 𐼏     | 𐼐     | 𐼑     | 𐼒     | 𐼓     | 𐼔     | 𐼕     | 𐼖     | 𐼗     | 𐼘     | 𐼙     | 𐼚     | 𐼛     | 𐼜     | 𐼝     | 𐼞     | 𐼟     | 𐼠     | 𐼡     | 𐼢     | 𐼣     | 𐼤     |

- Old Sogdian Letter Aleph
- Old Sogdian Letter Final Aleph
- Old Sogdian Letter Beth
- Old Sogdian Letter Final Beth
- Old Sogdian Letter Gimel
- Old Sogdian Letter He
- Old Sogdian Letter Final He
- Old Sogdian Letter Waw
- Old Sogdian Letter Zayin
- Old Sogdian Letter Heth
- Old Sogdian Letter Yodh
- Old Sogdian Letter Kaph
- Old Sogdian Letter Lamedh
- Old Sogdian Letter Mem
- Old Sogdian Letter Nun
- Old Sogdian Letter Final Nun
- Old Sogdian Letter Final Nun with Vertical Tail
- Old Sogdian Letter Samekh
- Old Sogdian Letter Ayn
  - Used only in the Aramaic heterogram `D resh-ayin-daleth is used in other heterograms
- Old Sogdian Letter Alternate Ayn
  - Used only in the Aramaic heterogram `D resh-ayin-daleth is used in other heterograms
- Old Sogdian Letter Pe
- Old Sogdian Letter Sadhe
- Old Sogdian Letter Final Sadhe
- Old Sogdian Letter Final Sadhe with Vertical Tail
- Old Sogdian Letter Resh-Ayn-Daleth
- Old Sogdian Letter Shin
- Old Sogdian Letter Taw
- Old Sogdian Letter Final Taw
- Old Sogdian Letter Final Taw with Vertical Tail

### Numbers

<table>
<thead>
<tr>
<th>U+10F1D</th>
<th>U+10F1E</th>
<th>U+10F1F</th>
<th>U+10F20</th>
<th>U+10F21</th>
<th>U+10F22</th>
<th>U+10F23</th>
<th>U+10F24</th>
<th>U+10F25</th>
<th>U+10F26</th>
</tr>
</thead>
<tbody>
<tr>
<td>𐼝</td>
<td>𐼞</td>
<td>𐼟</td>
<td>𐼠</td>
<td>𐼡</td>
<td>𐼢</td>
<td>𐼣</td>
<td>𐼤</td>
<td>𐼥</td>
<td>𐼦</td>
</tr>
</tbody>
</table>

- Old Sogdian Number One
- Old Sogdian Number Two
- Old Sogdian Number Three
- Old Sogdian Number Four
- Old Sogdian Number Five
- Old Sogdian Number Six
- Old Sogdian Number Seven
- Old Sogdian Number Eight
- Old Sogdian Number Nine
- Old Sogdian Number Ten
- Old Sogdian Number Twenty
- Old Sogdian Number Thirty
- Old Sogdian Number One Hundred
- Old Sogdian Fraction One Half

### Heterogram

<table>
<thead>
<tr>
<th>U+10F27</th>
</tr>
</thead>
<tbody>
<tr>
<td>𐼧</td>
</tr>
</tbody>
</table>

- Old Sogdian Heterogram Ayn-Daleth
  - Ligature of the Aramaic heterogram `D
Table 1: Comparison of Old Sogdian letters with those in Unicode blocks for related Iranian scripts and Aramaic. Parenthesis indicate that a letter has been unified with another in the respective encoding. In Inscriptional Pahlavi, ayin and resh are unified with waw, and qoph with mem. For Old Sogdian, daleth and regular ayin are unified with resh.
### Table 2: Comparison of Old Sogdian numerical signs with those in Unicode blocks for related Iranian scripts and Aramaic.

<table>
<thead>
<tr>
<th></th>
<th>Old Sogdian</th>
<th>Inscriptional Pahlavi</th>
<th>Inscriptional Parthian</th>
<th>Imperial Aramaic</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONE</td>
<td>א</td>
<td>י</td>
<td>ל</td>
<td>ו</td>
</tr>
<tr>
<td>TWO</td>
<td>ב</td>
<td>ב</td>
<td>ל</td>
<td>ו</td>
</tr>
<tr>
<td>THREE</td>
<td>ג</td>
<td>ג</td>
<td>ל</td>
<td>ו</td>
</tr>
<tr>
<td>FOUR</td>
<td>ד</td>
<td>ד</td>
<td>ל</td>
<td>—</td>
</tr>
<tr>
<td>FIVE</td>
<td>ה</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>TEN</td>
<td>י</td>
<td>י</td>
<td>א</td>
<td>ח</td>
</tr>
<tr>
<td>TWENTY</td>
<td>ק</td>
<td>ק</td>
<td>ג</td>
<td>ב</td>
</tr>
<tr>
<td>THIRTY</td>
<td>ל</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ONE HUNDRED</td>
<td>מ</td>
<td>מ</td>
<td>נ</td>
<td>ח</td>
</tr>
<tr>
<td>ONE THOUSAND</td>
<td>—</td>
<td>נ</td>
<td>ב</td>
<td>ל</td>
</tr>
<tr>
<td>TEN THOUSAND</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>ח</td>
</tr>
<tr>
<td>ONE HALF</td>
<td>נ</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Proposal to encode the Old Sogdian script in Unicode

Inscriptional, archaic form .xlim of .xlim ALEPH (K 4.1–4).

Written forms of xlim ALEPH (red) and xlim FINAL ALEPH (blue) (AL 2.1–6).

Figure 1: Specimens of aleph.
Inscriptional forms of Ꞁ BETH (K 4.1–2).

Written forms of Ꞁ BETH (red) and Ꞃ FINAL BETH (blue) (AL 2.1–6).

Figure 2: Specimens of beth.
Inscriptional form of ꜡ GIMEL (K 4.6).

Written forms of ꜡ GIMEL (AL 2.7–12).

Written forms of ꜡ GIMEL (AL 3.1–4).

Figure 3: Specimens of gimel
Inscriptional form of daleth in 𐼜𐼘𐼂𐼘 ʿBDt written as 𐤀 (= RESH-AYIN-DALETH) (K 4.1).

Written form of daleth in 𐤀𐤃 ʿD written as 𐤀 (= RESH-AYIN-DALETH) (AL 1.1).

Written form of daleth in 𐤀𐤄 ʿD written as 𐤀 (= RESH-AYIN-DALETH) (AL 2.1).

Written form of daleth in 𐤀𐤅 ʿD written as 𐤀 (= RESH-AYIN-DALETH) (AL 3.1).

The letter daleth in 𐤀𐤆 ʿD written as 𐤀 (= RESH-AYIN-DALETH) (AL 3 verso).

Usage of 𐤀 (= RESH-AYIN-DALETH) for representing daleth (blue), ayin (green), and resh (red) (AL 2.1–12).

Figure 4: Specimens of daleth.
Inscriptional form of א he in $WRH$ and א(ו) (IH)WH (K 2.3–4).

Inscriptional form of א he in זNH, א(ו)תN knth, א(ו)תN TMH (K 4.1–2).

Written forms of א he in זNH and א(ו)תN knδh (AL 2.10, 12).

Ubiquitous usage of ה final he in AL 3.1–6.

Figure 5: Specimens of he.
Proposal to encode the Old Sogdian script in Unicode
Anshuman Pandey

Inscriptional forms of ṭ waw (K 4.2–4).

Written forms of ṭ waw (AL 2.1–5).

Figure 6: Specimens of waw.
Inscriptional form of \( \text{zayin} \) (K 4).

Written form of \( \text{zayin} \) (AL 2.34–36).

Figure 7: Specimens of \( \text{zayin} \). See also figure 8.
Inscriptional forms of ZAYIN and NUN in K 4: 𐾆𐾃 ZNH (line 1) and 𐾆𐾃 GNZ (line 6).

Written forms of ZAYIN (magenta) and NUN at the end of word (AL 2.33–41). Final NUN is represented using both FINAL NUN (green) and FINAL NUN WITH VERTICAL TAIL (blue).

Figure 8: Comparison of zayin and nun. See also figure 14.
Inscriptional forms of ⢚ heth (K 4.3–7).

Written forms of ⢚ heth (AL 3.1–4).

Figure 9: Specimens of heth.
Proposal to encode the Old Sogdian script in Unicode

Anshuman Pandey

Inscriptional forms of 𐼊 yodh (K 4.1–3).

Written forms of 𐼊 yodh (AL 2.1–5).

Figure 10: Specimens of yodh.
Inscriptional forms of ꧄ KAPH (K.4.1–3).

Written forms of ꧄ KAPH (AL.2.1–4).

Figure 11: Specimens of kaph.
Inscriptional, archaic form \(\text{∆} \) of \(\text{∆} \text{LAMEDH} \) (K 4.1).

Written forms of \(\text{∆} \text{LAMEDH} \) (red) and its variant forms \(\text{∆} \) (green) and \(\text{∆} \) (blue) (AL 6.1–8).

Figure 12: Specimens of \textit{lamedh}. 

30
Proposal to encode the Old Sogdian script in Unicode

Inscriptional forms of 𐼍 mem (K 4.1–3).

Written forms of 𐼍 mem (AL 3.1–4).

Figure 13: Specimens of mem.
Proposal to encode the Old Sogdian script in Unicode

Anshuman Pandey

Inscriptional form of j nun (K 4).

Written forms of j nun (red), Temporary j nun (green), 1 final nun with vertical tail (blue) (AL 1.7–12).

Usage of 1 final nun (red) and 1 final nun with vertical tail (blue) in the word MN: ʃ and ʃ (AL 2.2–7).

Figure 14: Specimens of nun. See also figure 8.
Proposal to encode the Old Sogdian script in Unicode

Archaic form ʃ of ū samekh (K 4.1-4).

Written forms of ū samekh (AL 1.7–12).

Figure 15: Specimens of samekh.
The letter *ayin* in 𐼘𐼓 ‘D written using 𐼓 ALTERNATE AYIN (AL 1.1).

The *ayin* in 𐼘𐼓 ‘D written using 𐼓 ALTERNATE AYIN (AL 3.1).

The *ayin* in 𐼘𐼓 ‘D written using the glyphic variant 𐽢 of 𐼓 ALTERNATE AYIN (AL 3 verso).

The *ayin* in 𐼘𐼓 ‘D written using the glyphic variant 𐽢 of 𐼓 ALTERNATE AYIN (AL 5.1).

The letter *ayin* in 𐼘𐼘𐼂𐼘 ‘BDt inscribed as 𐼘 (= RESH-AYIN-DALETH) (K 4.1).

The letter *ayin* in 𐼌𐼘 ‘LZK written using 𐼘 (= RESH-AYIN-DALETH) (AL 2.12).

The letter *ayin* in 𐼘 ‘L written using 𐼘 (= RESH-AYIN-DALETH) (AL 6.6).

Figure 16: Specimens of *ayin*.
Glyphic variant 𒌣 of プ eo (K 4.1–6).

Written forms of プ eo (AL 1.6–12).

Figure 17: Specimens of pe.
Proposal to encode the Old Sogdian script in Unicode

Inscriptional form of šadhe (K 4.1–6).

Written forms of šadhe (red), šadhe (blue), and šadhe with vertical tail (green) in (AL 2).

Curved variant š of šadhe (AL 1.10).

Figure 18: Specimens of šadhe.
Usage of ꔚ for representing daleth (blue), ayin (green), and resh (red) (AL 2.1–12). As shown, ꔚ is most commonly used for resh. The letter ꔚ is proposed for encoding as the unified character RESH-AYIN-DALETH.

Figure 19: Comparison of daleth, ayin, and resh.
Inscriptional forms of  shin (K 4.1–3).

Written forms of  shin (AL 2.1–4).

Written forms of  shin (AL 3.1–3).

Figure 20: Specimens of shin.
Inscriptional, archaic forms 𐽞 (red) and 𐽟 (blue) of 𐽞 TAW and 𐽟 FINAL TAW WITH VERTICAL TAIL (K 4.1–2). The distinction is apparent in 𐽟𐽞𐼇𐼙 šwtt (line 2), which contains both nominal and final forms.

Written forms of 𐽞 FINAL TAW (blue) and 𐽟 FINAL TAW WITH VERTICAL TAIL (red) at the end of word (AL 2.28–36).

Contrasive usage of 𐽞 FINAL TAW and 𐽟 FINAL TAW WITH VERTICAL TAIL in two instances of the word prnxwnt: 𐽞𐽞 and 𐽞𐽞 (AL 1.5–6).

Curved variant 𐽞 of 𐽞 FINAL TAW (AL 1.8–10).

Figure 21: Specimens of taw.
The fraction ½ 𝒑 (AL 5.10).

The number 3 ᵣ (AL 2.32).

The number 4 ᵣ (AL 5.26).

The number 4½ 𝒑ᵣ (AL 5.24).

The number 5 ᵣ (AL 5.10).

The number 7½ 𝒑ᵣ (AL 5.26).

The number 8 ᵣ (AL 2.31).

Figure 22: Examples of numbers in the ‘Ancient Letters’. See also figures 23 and 24.
The number 10 𐼢 (AL 3.26).

The number 13 𐼡𐼢 (AL 2.62).

The number 15 𐼡𐼢 (AL 7.8).

The number 20 𐼣 (AL 5.21).

The number 30 𐼤 (AL 5.32).

The number 32 𐼆𐼣 (AL 2.62).

Figure 23: Additional examples of numbers in the ‘Ancient Letters’. See also figures 22 and 24.
The number 100 𐼕 (AL 2.19).

The number 200 𐼕 𐼙 (AL 7.3).

The number 500 𐼕 𐼙 𐼟 (AL 5.9).

The number 800 𐼕 𐼙 𐼠 (AL 4.3).

The number 1000 𐼔𐼌𐼝 (AL 2.1).

The number 2000 𐼔𐼌𐼝 𐼞 (AL 5.9).

The number 10000 represented using the word 𐼘𐼇𐼊𐼘𐼂 βrywr (AL 2.1).

Figure 24: Further examples of numbers in the ‘Ancient Letters’. See also figures 22 and 23.
The heterogram ‘D written as 𐼘𐼓 alternate ayin, 𐼘 resh-ayin-daleth> (AL 1.1).

The heterogram ‘D written as (𐼘) 𐼓 alternate ayin, (𐼘 resh-ayin-daleth)> (AL 1 verso).

The heterogram ‘D written as 𐼘 𐼓 alternate ayin, 𐼘 resh-ayin-daleth> (AL 2.1).

The heterogram ‘D written as 𐼘 𐼓 alternate ayin, 𐼘 resh-ayin-daleth> (AL 2 verso).

The heterogram ‘D written as 𐼘alternate ayin, 𐼘 resh-ayin-daleth> (AL 3.1).

The heterogram ‘D written as the ligature 𐽡 heterogram ayin-daleth (AL 3 verso).

The heterogram ‘D written as 𐼘 𐼓 alternate ayin, 𐼘 resh-ayin-daleth> (AL 4.1).

The heterogram ‘D written as 𐼘alternate ayin, 𐼘 resh-ayin-daleth> using the glyphic variant 𐽡 of alternate ayin (AL 5.1).

Figure 25: Specimens of the heterogram ‘D.
Figure 26: Two images of Kultobe inscription 4 (KII 26859/1). Top from Sims-Williams 2007; bottom from Grenet, et al 2007.
Figure 27: Kultobe inscriptions 2, 1, 3, 5, 10 (from Grenet, et al 2007).
Figure 28: The ‘Ancient Letter 1’ (British Library, International Dunhuang Project: Or. 8212/92.1 recto 1). “From her daughter, the free-woman Miwnay, to her d[ear] mother [Chatis].” (translation by Sims-Williams in Waugh 2004).
Figure 30: Bottom portion of ‘Ancient Letter 2’ (British Library, International Dunhuang Project: Or. 8212/95 side a). Continued from figure 29.
Figure 32: The ‘Ancient Letter 4’ (British Library: Or. 8212/93 recto; reproduced in Reichelt 1928: plate IV).
Figure 33: The ‘Ancient Letter 5’ (from Grenet, et al. 1998: 94). “To the noble lord, the chief merchant Aspandhāt. [Sent] by your servant [Frī-khwatāw].”
Figure 34: The ‘Ancient Letter 6’ (British Library, International Dunhuang Project: Or. 8212/97).

Figure 35: The ‘Ancient Letter 7’ (British Library: Or. 8212/96 recto; reproduced in Reichelt 1928: plate VII).
Figure 36: Sogdian rock inscription from Shatial (from Sims-Williams 1989: plate 10b) The inscription reads 𐼋𐽘𐼀𐽚𐼀𐽚 nnyʾkk ZK (top line), 𐼐𐼀𐼀𐼙𐼇𐽘 swʾbr (middle), 𐼈 BRY (bottom). Latin transcription from *ibid*: 14. The inscription in the bottom right-hand corner is shown in detail in figure 37.

Figure 37: Sogdian rock inscription from Shatial (from Sims-Williams 1989: plate 10a). The central inscription reads 𐼍𐼍𐼍𐼍 pʾkk (top line), 𐼐𐼀𐼀𐼙𐼇 ZK kwʾʾn (middle), 𐼈 BRY (bottom). Latin transcription from *ibid*: 14. The inscription in the top left-hand corner is shown in detail in figure 36.
Figure 38: Rock at Shatial containing horizontal and vertical inscriptions in the Old Sogdian script (from Sims-Williams 1989: plate 109b). The text of 36:38 is shown in section 4.5.
Figure 39: Silver coin from Chach bearing an Old Sogdian inscription, 3rd–4th c. CE (reproduced in Grenet 2007: 1023). Reverse: profile of human head. Obverse: *tamgha* in the center with the text 

\[ \text{𐼘𐼇𐼉𐽙𐼇𐽙𐼇 𐼕𐽚𐼀𐽙𐽙𐼀𐼕𐼀𐼕} \]

\[ c’c’n’pc wnwnxwr. \]

Figure 40: Reproduction of an Old Sogdian inscription on a silver vessel from the principality of Chach: 

\[ \text{𐼋𐼘𐼊𐼚𐼑 𐼢 𐼣 𐼟 𐼟 𐼟 𐼕𐽚𐼀𐽙𐽙𐼀𐼕𐼀𐼕 𐼎𐼀𐼀𐼉𐼘𐼀𐼊𐼍} \]

\[ my\text{c’c’n’pc 39 sty}rk \]

“Mayārkhan of the nation of Tashkent. 39 stater (ca. 624 g).” (from Yoshida 2002: 191). A *tamgha* appears to the right of the inscription.
Sogdian script

In the Sogdian script used in the “Ancient Letters” (Table 48.2), most of the letters are distinct and do not change shape when joined. In the “formal” and “Uyghur” Sogdian scripts, most of the letters are joined and, owing to the use of a broad pen, are frequently difficult to distinguish. In the earlier form, ‘= is still distinguished from ǚ; but in the later, ‘= n, ‘= n. Some scribes distinguish z from n by not connecting z to the preceding letter, but others make no distinction. In the later, increasingly cursive, form, other letters tend to become indistinguishable as well: γ/δ/δ, r/b/y. Some letters are distinguished only in final position (by some scribes), e.g., n ~ z, x ~ γ.

z is sometimes distinguished from n or z from ẓ by a diacritical point ȧ, and the foreign sound  b was noted as  p.

SAMPLES OF SOGDIAN

PL1  k‘n‘k  w̃α>yynn  k‘rβ  w̃γtwx  wγβ  DOκ–
                    wñγβ  wMXyKZ  YZKYA  ykwnt’sps  wyclmn  MLŚ  rwyrβ
                          ktnβynn  ktnβ  ḅpyx  NM  ṭypse  tryβ

1. Transliteration: OD βγw  xwt‘w  βr‘kk  nnyγβ rw  k‘n‘kk
2. Normalization: at β̄γu  xutāw  βarak  nanē-ōṭār  kanak
3. Gloss: to lord.acc master Barak Nana’s-gift Kanak

1. LP  βywr  ŠLM  nm‘cyw  sp’tz’nwky  AYKZY
2. (ēw-)zār  βrēwar  *āfriwan  namācyu  spätzānūk  kaś-uti
3. thousand ten.thousand greeting(?) reverence.acc bended.knee when-that.and

1. ZKyXMw  βγ̂nw  βyrt  pyşt  MN  xypθ  βntk  nnyβntk
2. wēšanu  βγ̃an(u)  βyart  pišt  con  xēpθ  βantē  nanē-βantē
3. them.obl lords.obl received written from own servant Nana’s-servant

‘To the Divine Master Barak (?) Nanethvar Kanak a thousand, ten thousand greetings, reverently with bended knees when received by their divinities. Written by his own servant Nanevante.’

—From the Old Sogdian “Ancient Letters” found in a mailbag in the Great Wall (AL II, Reichelt 1931: 12 and pl. 2).

Figure 41: Description of the Sogdian script of the ‘Ancient Letters’ (from Skjærvø 1996: 529).
### Table 48.2: Main East Iranian Scripts Developed from Aramaic

<table>
<thead>
<tr>
<th>Aramaic Letters</th>
<th>Sogdian Ancient Sutra Script</th>
<th>Sogdian Manichean Christian</th>
<th>Principal Phonetic Values (Sogdian)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, æ</td>
<td>a</td>
<td>b, β</td>
<td>æ, α</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td>(β)</td>
<td>β</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>g, γ</td>
<td>γ</td>
<td></td>
</tr>
<tr>
<td>(γ)</td>
<td>γ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>d, δ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>h (h)</td>
<td>a, Ø</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>w, ō, ū</td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(j)</td>
<td>ŋ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ž)</td>
<td>ŋ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>h (h)</td>
<td>γ, x, h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>y, ĕ, ĩ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x)</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>l (ł)</td>
<td>ł</td>
<td>ł</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>Ø</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f)</td>
<td>f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>š (c)</td>
<td>č, j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q</td>
<td>k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>š</td>
<td>š</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>t, θ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 42: Table showing various scripts for writing Sogdian (from Skjærvø 1996: 519).
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646


### A. Administrative

| 1. Title: | Proposal to encode the Old Sogdian script in Unicode |
| 2. Requester's name: | Anshuman Pandey <pandey@umich.edu> |
| 3. Requester type (Member body/Liaison/Individual contribution): | Expert contribution |
| 4. Submission date: | 2016-12-31 |
| 5. Requester's reference (if applicable): | |
| 6. Choose one of the following: | |
| (or) More information will be provided later: | Yes |

### B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): Yes
   b. The proposal is for addition of character(s) to an existing block: Old Sogdian

2. Number of characters in proposal: 40

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? 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](http://www.unicode.org) for such information on other scripts. Also see Unicode Character Database ([http://www.unicode.org/reports/tr44/](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.

---

### C. Technical - Justification

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has this proposal for addition of character(s) been submitted before?</td>
<td>No</td>
</tr>
<tr>
<td>If YES explain</td>
<td></td>
</tr>
<tr>
<td>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.)?</td>
<td>Yes</td>
</tr>
</tbody>
</table>
| If YES, with whom?                                                      | Nicholas Sims-Williams <ns5@soas.ac.uk>  
Yutaka Yoshida <yutaka.yoshida@bun.kyoto-u.ac.jp> |
| 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? | Yes      |
| Reference:                                                              | See text of proposal |
| 4. The context of use for the proposed characters (type of use; common or rare) | Common   |
| Reference:                                                              | See text of proposal |
| 5. Are the proposed characters in current use by the user community?    | Yes;     |
| If YES, where? Reference:                                               | Currently used by scholars of Sogdian and Central Asian studies |
| 6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? | N/A      |
| 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? | 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? | 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? | N/A      |
| 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? |          |