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'.

- *'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*, *beth*, *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:

		Old Sogdian	Later Sogdian
swγδyk	'Sogdian'	𐼒𐼈𐼄𐼙𐼋𐼌	𐼼𐼴𐼲𐽀𐼷𐼸
sm'rknδc	'of Samarkand'	𐼒𐼎𐼀𐼙𐼌𐼏𐼙𐼖	𐼼𐼺𐼰𐽀𐼸𐼻𐽀𐼿

- *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 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. 𐼀 is *aleph* and OLD SOGDIAN LETTER ALEPH. For sake of brevity, the descriptor 'OLD SOGDIAN' is dropped when refering 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

Glyph	Character name	Latin
𐼀	OLD SOGDIAN LETTER ALEPH	ʾ
𐼁	OLD SOGDIAN LETTER FINAL ALEPH	ʾ
𐼂	OLD SOGDIAN LETTER BETH	β ; B
𐼃	OLD SOGDIAN LETTER FINAL BETH	β ; B
𐼄	OLD SOGDIAN LETTER GIMEL	γ ; G
𐼅	OLD SOGDIAN LETTER HE	h
𐼆	OLD SOGDIAN LETTER FINAL HE	h

𐼵	OLD SOGDIAN LETTER WAW	w
𐼶	OLD SOGDIAN LETTER ZAYIN	z
𐼷	OLD SOGDIAN LETTER HETH	x ; Ḥ
𐼸	OLD SOGDIAN LETTER YODH	y
𐼹	OLD SOGDIAN LETTER KAPH	k
𐼺	OLD SOGDIAN LETTER LAMEDH	δ ; L
𐼻	OLD SOGDIAN LETTER MEM	m
𐼼	OLD SOGDIAN LETTER NUN	n
𐼽	OLD SOGDIAN LETTER FINAL NUN	n
𐼾	OLD SOGDIAN LETTER FINAL NUN WITH VERTICAL TAIL	n
𐼿	OLD SOGDIAN LETTER SAMEKH	s
𐽀	OLD SOGDIAN LETTER AYIN	ʿ
𐽁	OLD SOGDIAN LETTER ALTERNATE AYIN	ʿ
𐽂	OLD SOGDIAN LETTER PE	p
𐽃	OLD SOGDIAN LETTER SADHE	c
𐽄	OLD SOGDIAN LETTER FINAL SADHE	c ; Ṣ
𐽅	OLD SOGDIAN LETTER FINAL SADHE WITH VERTICAL TAIL	c
𐽆	OLD SOGDIAN LETTER RESH-DALETH-AYIN	r, d, ʿ
𐽇	OLD SOGDIAN LETTER SHIN	š
𐽈	OLD SOGDIAN LETTER TAW	t
𐽉	OLD SOGDIAN LETTER FINAL TAW	t
𐽊	OLD SOGDIAN LETTER FINAL TAW WITH VERTICAL TAIL	t

3.1.1 Notes on letters

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

beth In word-final positions in AL, 𐼂 BETH is written as 𐼃 FINAL BETH, in which the horizontal stroke at the baseline is elongated. See figure 2 for attestations.

gimel See figure 3 for attestations of 𐼄 GIMEL.

daleth The letter 𐼙 *daleth* occurs only in Aramaic heterograms, eg. 𐼒𐼂𐼙𐼜 *ʿBDt* (K 4.1) and 𐼓𐼙 *ʿD* (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 𐡃 ARAMAIC LETTER DALETH and 𐡓 ARAMAIC 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 𐼙 RESH-AYIN-DALETH.

he The letter *he* is represented by both 𐼅 HE and 𐼆 FINAL HE. The 𐼅 HE occurs only in K, eg. 𐼕𐼇𐼙𐼅 *ṢWRH* and (𐼅)𐼇𐼅 *(H)WH* (K 2); 𐼋𐼎𐼛𐼅 *knth* and 𐼈𐼎𐼅 *ZNH* (K 4). The representative glyph is derived from K 2. In AL, *he* is written as 𐼆, eg. 𐼋𐼈𐼌𐼆 *knδh* (AL 2.12), 𐼈𐼈𐼆 *ZNH* (AL 2.10). The 𐼆 FINAL HE is not attested in K, and in contrast to 𐼅, occurs only at the end of a word. It has the variant shape 𐼆, 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 𐼇 WAW.

zayin See figures 7 and 8 for attestations of 𐼈 ZAYIN.

heth See figure 9 for attestations of 𐼉 HETH.

teth An Old Sogdian analogue for Aramaic *teth* does not exist. In K, the *teth* in Aramaic heterograms is represented using 𐼛 TAW: *QṬLt* is written as 𐼋𐼛𐼌𐼜 *KTLt* (K 3.3).

yodh See figure 10 for attestations of 𐼊 YODH.

kaph See figure 11 for attestations of 𐼋 KAPH.

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

mem See figure 13 for attestations of 𐼍 MEM.

nun Occurrences of *nun* are represented using 𐼎 NUN, 𐼏 FINAL NUN, 𐼐 FINAL NUN WITH VERTICAL TAIL (see figure 14). The representative glyph 𐼎 for NUN is derived from K. The final forms occur only in AL. While *nun* has the distinctive shape 𐼎 in K, it has the shape 𐼈 in AL when non-final, which is identical to 𐼈 ZAYIN, eg. 𐼈𐼎𐼅 *ZNH* (K 4.1) and 𐼈𐼈𐼆 *ZNH* (AL 2.10). When word-final in AL, *nun* is written as both 𐼏 and 𐼐, eg. 𐼍𐼏 *MN* (AL 2.2) and 𐼍𐼐 *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 𐼑 SAMEKH occurs as the two-part form 𐼑 in K 4. This archaic form is to be treated as a glyphic variant. See figure 15 for attestations.

ayin The letter *ayin* occurs only in Aramaic heterograms. It has the regular shape 𐼒 and the special shapes 𐼓 and 𐼧 (see figure 16). The regular 𐼒 *ayin* occurs in both K and AL, eg. 𐼒𐼂𐼙𐼜 *ʿBDt* (K 4.1), 𐼒𐼌𐼈𐼋

ʿLZK (AL 2.12), 𐼘𐼌 *ʿL* (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 𐡏 ARAMAIC LETTER AYIN and 𐡓 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 𐼘 RESH-AYIN-DALETH. The letters 𐼒 AYIN and 𐼓 ALTERNATE AYIN occur only in AL for writing the heterogram *ʿD*, eg. 𐼒𐼘 (AL 2.1), 𐼓𐼘 (AL 3 verso), 𐼓𐼘 (AL 3.1), 𐼓𐼘 (AL 5.1). Although 𐼒, 𐼓, 𐼓, 𐼓 could be considered glyphic variants of a single character AYIN, it is appropriate to define two characters on account of their graphical structures. The 𐼓 is a glyphic variant of 𐼓 with an ornate tail; the 𐼓 is a variant with a truncated tail. These three forms are unified as 𐼓 ALTERNATE AYIN, which may be used for representing these special forms in plain text. See figure 25 and § 3.3 for attestations.

pe The letter 𐼔 PE is has the variant 'open' shape 𐼔, which is a glyphic variant (see figure 17).

sadhe This letter is represented using 𐼕 SADHE, 𐼖 FINAL SADHE, and 𐼗 FINAL SADHE WITH VERTICAL TAIL (see figure 18). The final forms occur only in AL. In AL 2, *sadhe* has the shape 𐼗 whenever it occurs at the margin, eg. 𐼀𐼉𐼘𐼈𐼊𐼎𐼎𐼗 *ʾḤRZYnnc* (AL 2.54). In other positions within a line, final *sadhe* is written using 𐼖 FINAL SADHE, eg. 𐼀𐼉𐼘𐼈𐼊𐼎𐼖 *ʾḤRZYnc* (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 𐼖 of 𐼖 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 𐼋 KAPH: *Q̣TLt* is written as 𐼋𐼚𐼌𐼜 *KTLt* (K 3.3). It used to be believed that 𐡒 *qoph* was retained in AL as 𐼦 and reassigned for the number 100. This 𐼦 is now identified as the fraction ½ (Grenet, et al 1998).

resh In AL, the letter 𐼘 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 𐼘 is used ubiquitously for *resh*. Accordingly, *daleth* and *ayin* are unified with *resh* as 𐼘 RESH-AYIN-DALETH. This approach follows the Unicode model for Inscriptional Pahlavi, in which *waw*, *ayin*, *resh* are represented using 𐭥 U+10B65 INSCRIPTIONAL PAHLAVI LETTER WAW-AYIN-RESH; and similarly, *mem* and *qoph* using 𐭬 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 𐼙 SHIN.

taw This letter is represented using 𐼚 TAW, 𐼛 FINAL TAW, 𐼜 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 𐼚 and final *taw* as 𐼜. These archaic two-part forms are to be treated as glyphic variants of TAW and FINAL TAW WITH VERTICAL TAIL, respectively. The 𐼛 FINAL TAW is often written using a glyphic variant with a curved tail 𐼛 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 𐼵 of 𐼀 *aleph* and the final form 𐼶 of 𐼁 *beth* 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 occuring letters such as 𐼅 *waw*, 𐼉 *yodh*, 𐼗 *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 𐼫 *kaph* and 𐼯 *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 𐼴 *sadhe* is written in AL 2 with a vertical tail 𐼳 whenever it occurs at the margin, and with a horizontal tail 𐼵 in other positions within a line (see figure 18). Throughout the AL corpus, 𐼡 *nun* is written using both 𐼢 and 𐼣 at the ends of words and lines. The same applies to the usage of 𐼧 and 𐼦 for final 𐼨 *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 CE, 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 preferences 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 𐼢 / 𐼣 *nun*, 𐼵 / 𐼳 *sadhe*, 𐼧 / 𐼦 *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.

Glyph	Character name	Numeric value
𐼰	OLD SOGDIAN NUMBER ONE	1
𐼱	OLD SOGDIAN NUMBER TWO	2
𐼲	OLD SOGDIAN NUMBER THREE	3
𐼳	OLD SOGDIAN NUMBER FOUR	4
𐼴	OLD SOGDIAN NUMBER FIVE	5
𐼵	OLD SOGDIAN NUMBER TEN	10
𐼶	OLD SOGDIAN NUMBER TWENTY	20
𐼷	OLD SOGDIAN NUMBER THIRTY	30
𐼸	OLD SOGDIAN NUMBER ONE HUNDRED	100
𐼹	OLD SOGDIAN FRACTION ONE HALF	½

Primary units The primary units are expressed using joined repetitions of the sign 𐼰 that are generally grouped in sets of three or four and separated by spaces, eg. 𐼱 for 2, 𐼲 for 3, 𐼳 𐼳 for 8. As the script is non-conjoining, no simple method exists for representing the ligated repetitions of 𐼰. For that reason, the numbers 𐼰 ONE .. 𐼳 FOUR are encoded atomically. This model for ONE .. FOUR follows the Unicode encoding for Inscriptional Parthian, eg. 𐭸 U+10B58 INSCRIPTIONAL PARTHIAN NUMBER ONE .. 𐭻 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 𐼴 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 𐼵 TEN resembles a vertically compressed 𐼪 LAMEDH. The shapes for 𐼶 TWENTY and 𐼷 THIRTY are formed from vertical stacks of 𐼵 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 𐼵𐼶, which is a compound of TWENTY and TEN.

Hundreds The number 100 is written using 𐼸 ONE HUNDRED. The glyph resembles the letter 𐼧 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 𐼝𐼌𐼔 *1LP*, which is represented using the sequence <𐼝 ONE, 𐼌 LAMEDH, 𐼔 PE>. The sequence 𐼝𐼌𐼔 also functions as a unit mark for the thousands. The 𐼝 ONE is an inherent part of the *1LP* unit. Multiples are expressed by prefixing primary numbers before the unit, eg. 2000 is 𐼞 𐼝𐼌𐼔, 3000 is 𐼟 𐼝𐼌𐼔.

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

Fraction The 𐼦 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.

Value	Number	Input string →
4½	𐼠𐼦	<𐼠 FOUR, 𐼦 FRACTION ONE HALF>
5	𐼟 𐼞	<𐼟 THREE, [SP] SPACE, 𐼞 TWO>
5	𐼡	<𐼡 FIVE>
6	𐼟 𐼟	<𐼟 THREE, [SP] SPACE, 𐼟 THREE>
7	𐼠 𐼟	<𐼠 FOUR, [SP] SPACE, 𐼟 THREE>
7½	𐼠 𐼟𐼦	<𐼠 FOUR, [SP] SPACE, 𐼟 THREE, 𐼦 FRACTION ONE HALF>
8	𐼠 𐼠	<𐼠 FOUR, [SP] SPACE, 𐼠 FOUR>
9	𐼟 𐼟 𐼟	<𐼟 THREE, [SP] SPACE, 𐼟 THREE, [SP] SPACE, 𐼟 THREE>
13	𐼢𐼟	<𐼢 TEN, 𐼟 THREE>
15	𐼢𐼡	<𐼢 TEN, 𐼡 FIVE>
30	𐼤	<𐼤 THIRTY>
30	𐼣𐼢	<𐼣 TWENTY, 𐼢 TEN>
32	𐼣𐼢𐼞	<𐼣 TWENTY, 𐼢 TEN, 𐼞 TWO>
100	𐼥	<𐼥 ONE HUNDRED>

200	𐼞 𐼥	<𐼞 TWO, [SP] SPACE, 𐼥 ONE HUNDRED>
500	𐼟 𐼞𐼥	<𐼟 THREE, [SP] SPACE, 𐼞 TWO, 𐼥 ONE HUNDRED>
1000	𐼝𐼌𐼔	<𐼝 ONE, 𐼌 LAMEDH, 𐼔 PE>
2000	𐼞 𐼝𐼌𐼔	<𐼞 TWO, [SP] SPACE, 𐼝 ONE, 𐼌 LAMEDH, 𐼔 PE>
10000	𐼂𐼙𐼊𐼇𐼙	<𐼂 BETH, 𐼙 RESH, 𐼊 YODH, 𐼇 WAW, 𐼙 RESH>

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:

Value	Number	Input string →
2453	𐼞 𐼝𐼌𐼔 𐼠𐼥 𐼤𐼣𐼟	<𐼞 TWO, [SP] SPACE, 𐼝 ONE, 𐼌 LAMEDH, 𐼔 PE, [SP] SPACE, 𐼠 FOUR, 𐼥 ONE HUNDRED, [SP] SPACE, 𐼤 THIRTY, 𐼣 TWENTY, 𐼟 THREE>
2453	𐼞 𐼝𐼌𐼔 𐼠𐼥 𐼣𐼣𐼢𐼟	<𐼞 TWO, [SP] SPACE, 𐼝 ONE, 𐼌 LAMEDH, 𐼔 PE, [SP] SPACE, 𐼠 FOUR, 𐼥 ONE HUNDRED, [SP] SPACE, 𐼣 TWENTY, 𐼣 TWENTY, 𐼢 TEN, 𐼟 THREE>

3.3 Heterogram

The repertoire contains 1 heterogram.

Glyph	Character name	Value
𐼧	OLD SOGDIAN HETEROGRAM AYIN-DALETH	*ʿD*

Aramaic heterograms are represented as words spelled using conventional letters, eg. 𐼀𐼉𐼙𐼈𐼊 *ʾḤRZY* is written <𐼀 ALEPH, 𐼉 HETH, 𐼙 RESH, 𐼈 ZAYIN, 𐼊 YODH>. The heterogram *ʿD* is the sole exception. Meaning “to”, *ʿD* occurs in the address and salutation of a letter, eg. 𐼒𐼙 𐼂𐼊𐼇 𐼉𐼇𐼛𐼀𐼇 𐼂𐼀𐼙𐼋𐼋𐼇 *ʿD βyw xwtʾw βʾrkkw* “to lord master Barak”. Morphologically, it is comprised of *ayin* and *daleth*. Yet, instead of the expected spelling *𐼙𐼙 <𐼙 RESH-AYIN-DALETH, 𐼙 RESH-AYIN-DALETH>, the *ayin* is written using special forms: 𐼒𐼙, 𐼧, 𐼓𐼙, 𐼓𐼙 (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 *ʿD* (Sims-Williams, personal correspondence, 2016).

There are two ways to analyze these representations of *ʿD*. 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 𐼒 AYIN and 𐼓 ALTERNATE AYIN; the 𐼓 and 𐼓 could be considered glyphic variants of ALTERNATE

AYIN. Accordingly, *ʿD* may be represented as <AYIN | ALTERNATE AYIN, RESH-AYIN-DALETH>. Secondly, *ʿD* is a logographic unit comprised of a ligature or a set of two letters. This unit may be treated as an atomic character, eg. 𐼺 OLD SOGDIAN HETEROGRAM AYIN-DALETH. These approaches are not mutually exclusive and both are practical for character encoding. Depending upon the context, *ʿD* may be spelled using a sequence of letters or represented using an atomic character.

The case of *ʿD* is similar to the Latin '&' ampersand. The '&' represents the Latin word *et* "and". Morphologically, it is a ligation of the Latin letters 'e' and 't', eg.  ,  . 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  U+1F670 SCRIPT LIGATURE ET ORNAMENT.

The character name for HETEROGRAM AYIN-DALETH is derived from the conventional transliteration *ʿD* of the heterogram. The representative glyph 𐼺 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:

𐼰 ALEPH ≪ 𐽆 FINAL ALEPH < 𐼱 BETH ≪ 𐽇 FINAL BETH < 𐼲 GIMEL < 𐼳 HE ≪
𐽈 FINAL HE < 𐼴 WAW < 𐼵 ZAYIN < 𐼶 HETH < 𐼷 YODH < 𐼸 KAPH < 𐼹 LAMEDH <
𐼻 MEM < 𐼼 NUN ≪ 𐽉 FINAL NUN ≪ 𐽊 FINAL NUN WITH VERTICAL TAIL <
𐼽 SAMEKH < 𐼾 AYIN ≪ 𐼿 ALTERNATE AYIN < 𐼺 HETEROGRAM AYIN-DALETH <
𐽀 PE < 𐽁 SADHE ≪ 𐽋 FINAL SADHE ≪ 𐽌 FINAL SADHE WITH VERTICAL TAIL <

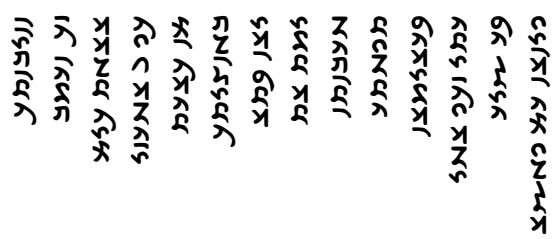
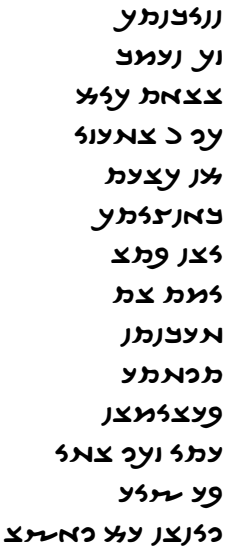
𐼰 RESH-AYIN-DALETH < 𐼩 SHIN < 𐼨 TAW << 𐼪 FINAL TAW <<

𐼫 FINAL TAW WITH VERTICAL TAIL

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)	Horizontal
	

nnyβntk / ZK nrsβ / ʾʾγt kym / kw 10 ʾḤRZY / MN kʾrt / βγncytk / yʾn ptʾ[-] / [-]yst ʾt / xrβntn / twxtr / prʾysʾn / rty ZKw ʾḤY / pr šyr / wynʾn ʿM wyšʾ

"(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 enviroments. 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;;;;;
10F08;OLD SOGDIAN LETTER ZAYIN;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;;;;;
10F10;OLD SOGDIAN LETTER FINAL NUN WITH VERTICAL TAIL;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

Anderson, Deborah; et. al. 2016. "Recommendations to UTC #146 January 2016 on Script Proposals". L2/16-037. http://www.unicode.org/L2/L2016/16037-script-rec.pdf

———. 2017. "Recommendations to UTC #150 January 2017 on Script Proposals". L2/17-037. http://www.unicode.org/L2/L2017/17037-script-ad-hoc.pdf

Grenet, Frantz; Nicholas Sims-Williams; Étienne de La Vaissière. 1998. "The Sogdian Ancient Letter V". *Bulletin of the Asia Institute*, Alexander's Legacy in the East: Studies in Honor of Paul Bernard, New series, vol. 12, ed. Osmund Bopearachchi, Carol Altman Bromberg, and Frantz Grenet, pp. 91–104.

Grenet, Frantz; Nicholas Sims-Williams; Aleksandr Podushkin. 2007. "Les plus anciens monuments de la langue sogdienne: les inscriptions de Kultobe au Kazakhstan". *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres*, 151^{e} année, N. 2, 2007, pp. 1005–1034.

Pandey, Anshuman. 2016. "Revised proposal to encode the Sogdian script in Unicode" (L2/16-371). http://www.unicode.org/L2/L2016/16371-sogdian.pdf

Reichelt, Hans. 1928–31. *Die soghdischen Handschriftenreste des Britischen Museums*. Heidelberg: C. Winter's Universitätsbuchhandlung.

Sims-Williams, Nicholas. 1975. "Notes on Sogdian Palaeography". *Bulletin of the School of Oriental and African Studies, University of London*, vol. 38, no. 1 (1975), pp. 132–139.

———. 1981a. "The Sogdian sound-system and the origins of the Uyghur script". *Journal Asiatique*, pp. 347–360.

———. 1981b. "Remarks on the Sogdian letters *γ* and *x* (with special reference to the orthography of the Sogdian version of the Manichean church-history)", Appendix in W. Sundermann, *Mitteliranische manichäische Texte kirchen-geschichtlichen Inhalts*, Berliner Turfantexte, XI, Berlin: Akademie-Verlag, pp. 194–198.

———. 1985. "Ancient Letters". *Encyclopædia Iranica*, vol. II, fasc. 1, pp. 7–9. http://www.iranicaonline.org/articles/ancient-letters

———. 1989. *Sogdian and Other Iranian Inscriptions of the Upper Indus*. Corpus Inscriptionum Iranicarum, pt. II (Inscriptions of the Seleucid and Parthian Periods and of Eastern Iran and Central Asia), v. III (Sogdian), no. I. London: Published on behalf of Corpus Inscriptionum Iranicarum by School of Oriental and African Studies.

———. 2000. "The Iranian Inscriptions of Shatial". *Indologica Taurinensia*, v. 23–24 (Professor Gregory M. Bongard-Levin Felicitation Volume), pp. 523–541.

Sims-Williams, Nicholas; Frantz Grenet. 2007. "The Sogdian Inscriptions of Kultobe". *Shygys*, 2006, vol. 1 pp. 95–111.

Skjærvø, Prods Oktor. 1996. "Aramaic Scripts for Iranian Languages." *The World's Writing Systems*, edited by Peter T. Daniels and W. Bright, pp. 515–535. New York and Oxford: Oxford University Press.

Waugh, Daniel C. [comp]. 2004. "The Sogdian Ancient Letters", translated by Nicholas Sims-Williams. https://depts.washington.edu/silkroad/texts/sogdlet.html

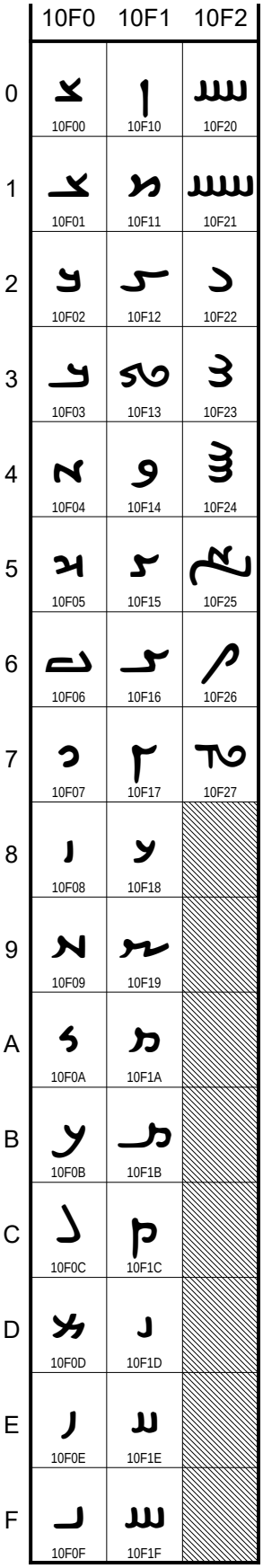
Yoshida, Yutaka. 2002. "In search of traces of the Sogdians 'Phoenicians of the Silk Road"'. *Berlin-Brandenburgische Akademie der Wissenschaften*. Berichte und Abhandlungen, Band 9, p. 185–200. Berlin.

———. 2013. "When Did Sogdians Begin to Write Vertically?". *Tokyo University Linguistic Papers*, vol. 33, pp. 375–394.

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

10F00 𐼀 OLD SOGDIAN LETTER ALEPH
10F01 𐼁 OLD SOGDIAN LETTER FINAL ALEPH
10F02 𐼂 OLD SOGDIAN LETTER BETH
10F03 𐼃 OLD SOGDIAN LETTER FINAL BETH
10F04 𐼄 OLD SOGDIAN LETTER GIMEL
10F05 𐼅 OLD SOGDIAN LETTER HE
10F06 𐼆 OLD SOGDIAN LETTER FINAL HE
10F07 𐼇 OLD SOGDIAN LETTER WAW
10F08 𐼈 OLD SOGDIAN LETTER ZAYIN
10F09 𐼉 OLD SOGDIAN LETTER HETH
10F0A 𐼊 OLD SOGDIAN LETTER YODH
10F0B 𐼋 OLD SOGDIAN LETTER KAPH
10F0C 𐼌 OLD SOGDIAN LETTER LAMEDH
10F0D 𐼍 OLD SOGDIAN LETTER MEM
10F0E 𐼎 OLD SOGDIAN LETTER NUN
10F0F 𐼏 OLD SOGDIAN LETTER FINAL NUN
10F10 𐼐 OLD SOGDIAN LETTER FINAL NUN WITH VERTICAL TAIL
10F11 𐼑 OLD SOGDIAN LETTER SAMEKH
10F12 𐼒 OLD SOGDIAN LETTER AYIN
- used only in the Aramaic heterogram `D
- resh-ayin-daleth is used in other heterograms

10F13 𐼓 OLD SOGDIAN LETTER ALTERNATE AYIN
- used only in the Aramaic heterogram `D
- resh-ayin-daleth is used in other heterograms

10F14 𐼔 OLD SOGDIAN LETTER PE
10F15 𐼕 OLD SOGDIAN LETTER SADHE
10F16 𐼖 OLD SOGDIAN LETTER FINAL SADHE
10F17 𐼗 OLD SOGDIAN LETTER FINAL SADHE WITH VERTICAL TAIL
10F18 𐼘 OLD SOGDIAN LETTER RESH-AYIN-DALETH
10F19 𐼙 OLD SOGDIAN LETTER SHIN
10F1A 𐼚 OLD SOGDIAN LETTER TAW
10F1B 𐼛 OLD SOGDIAN LETTER FINAL TAW
10F1C 𐼜 OLD SOGDIAN LETTER FINAL TAW WITH VERTICAL TAIL

Numbers

10F1D 𐼝 OLD SOGDIAN NUMBER ONE
10F1E 𐼞 OLD SOGDIAN NUMBER TWO
10F1F 𐼟 OLD SOGDIAN NUMBER THREE
10F20 𐼠 OLD SOGDIAN NUMBER FOUR
10F21 𐼡 OLD SOGDIAN NUMBER FIVE
10F22 𐼢 OLD SOGDIAN NUMBER TEN
10F23 𐼣 OLD SOGDIAN NUMBER TWENTY
10F24 𐼤 OLD SOGDIAN NUMBER THIRTY
10F25 𐼥 OLD SOGDIAN NUMBER ONE HUNDRED
10F26 𐼦 OLD SOGDIAN FRACTION ONE HALF

Heterogram

10F27 𐼧 OLD SOGDIAN HETEROGRAM AYIN-DALETH
- ligature of the Aramaic heterogram `D

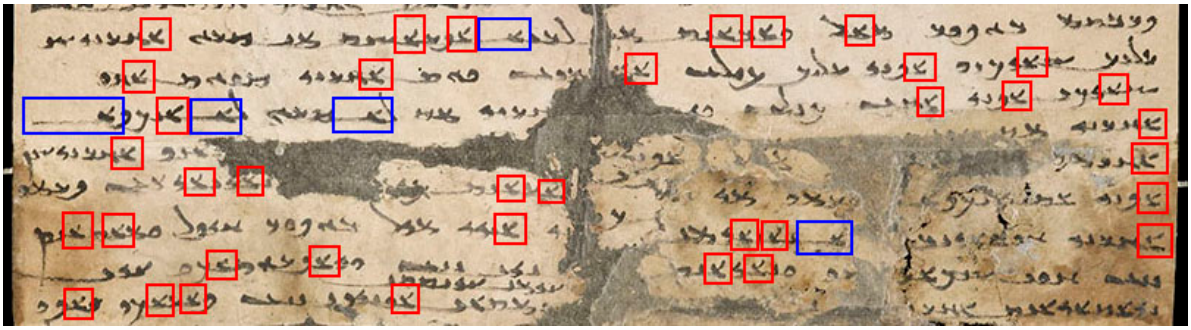
	Old Sogdian	Inscriptional Pahlavi	Inscriptional Parthian	Imperial Aramaic
aleph	𐼰, 𐼱	𐭠	𐭠	𐡀
beth	𐼲, 𐼳	𐭡	𐭡	𐡁
gimel	𐼴	𐭢	𐭢	𐡂
daleth	(𐼵)	𐭣	𐭣	𐡃
he	𐼶, 𐼷	𐭤	𐭤	𐡄
waw	𐼸	𐭥	𐭥	𐡅
zayin	𐼹	𐭦	𐭦	𐡆
heth	𐼺	𐭧	𐭧	𐡇
teth	—	𐭨	𐭨	𐡈
yodh	𐼻	𐭩	𐭩	𐡉
kaph	𐼼	𐭪	𐭪	𐡊
lamedh	𐼽	𐭫	𐭫	𐡋
mem	𐼾	𐭬	𐭬	𐡌
nun	𐼿, 𐽀, 𐽁	𐭭	𐭭	𐡍
samekh	𐽂	𐭮	𐭮	𐡎
ayin	𐽃, 𐽄, (𐼵)	(𐭥)	𐭯	𐡏
pe	𐽅	𐭰	𐭰	𐡐
sadhe	𐽆, 𐽇, 𐽈	𐭱	𐭱	𐡑
qoph	—	(𐭬)	𐭲	𐡒
resh	𐽉	(𐭥)	𐭳	𐡓
shin	𐽊	𐭲	𐭴	𐡔
taw	𐽋, 𐽌, 𐽍	𐭵	𐭵	𐡕

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*.

	Old Sogdian	Inscriptional Pahlavi	Inscriptional Parthian	Imperial Aramaic
ONE	𐼱	𐭡	𐭡	𐡡
TWO	𐼲	𐭢	𐭢	𐡢
THREE	𐼳	𐭣	𐭣	𐡣
FOUR	𐼴	𐭤	𐭤	—
FIVE	𐼵	—	—	—
TEN	𐼶	𐭥	𐭩	𐡩
TWENTY	𐼷	𐭦	𐭪	𐡪
THIRTY	𐼸	—	—	—
ONE HUNDRED	𐼹	𐭧	𐭫	𐡫
ONE THOUSAND	—	𐭨	𐭬	𐡬
TEN THOUSAND	—	—	—	𐡭
ONE HALF	𐼺	—	—	—

Table 2: Comparison of Old Sogdian numerical signs with those in Unicode blocks for related Iranian scripts and Aramaic.

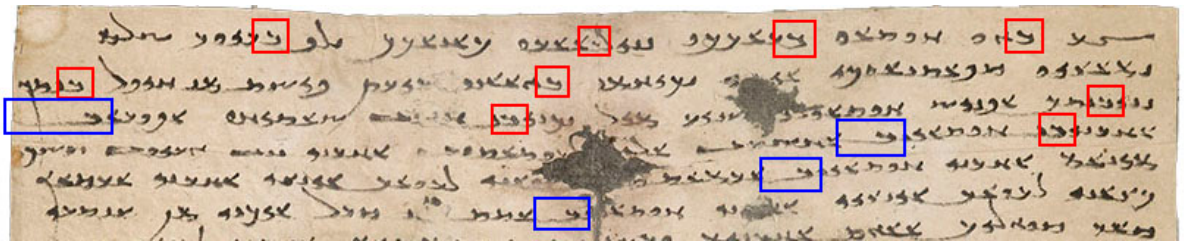
Inscriptional, archaic form 𐼀 of 𐽀 ALEPH (K 4.1–4).



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

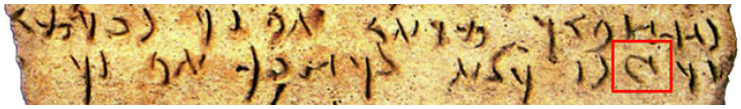
Figure 1: Specimens of *aleph*.

Inscriptional forms of 𐼁 BETH (K 4.1–2).

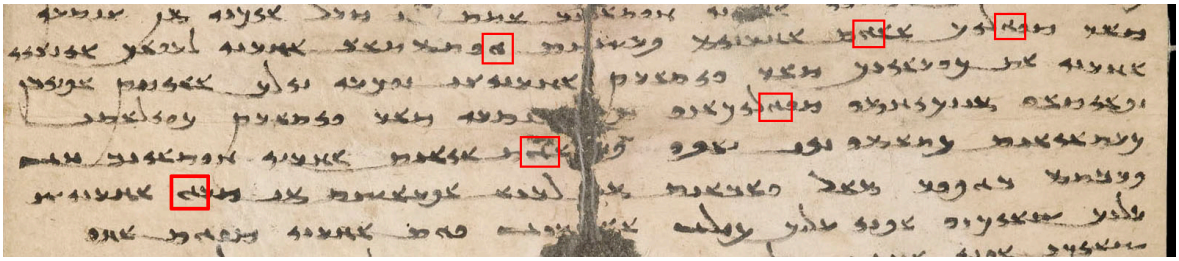


Written forms of 𐼁 BETH (red) and 𐼂 FINAL BETH (blue) (AL 2.1–6).

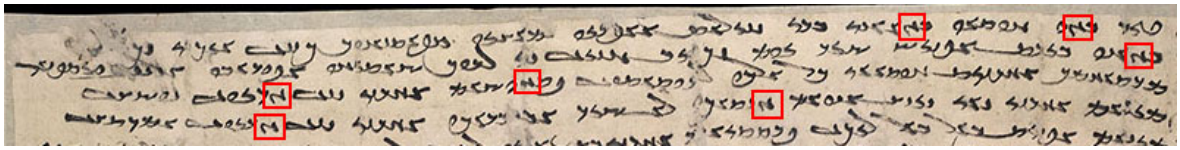
Figure 2: Specimens of *beth*.



Inscriptional form of 𐼇 GIMEL (K 4.6).



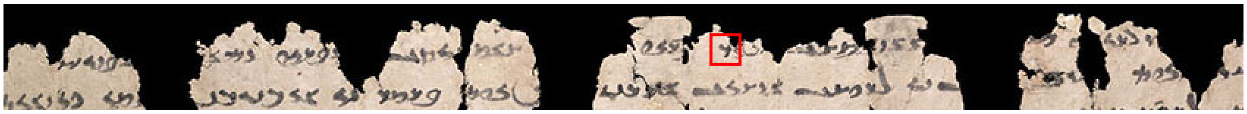
Written forms of 𐼇 GIMEL (AL 2.7–12).



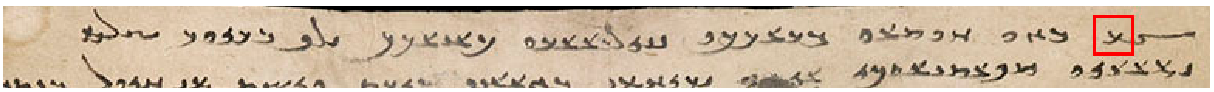
Written forms of 𐼇 GIMEL (AL 3.1–4).

Figure 3: Specimens of *gimel*

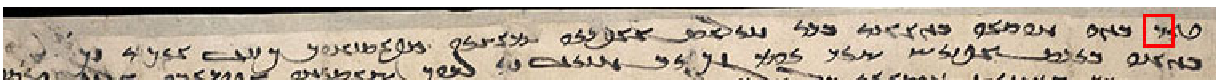
Inscriptional form of *daleth* in 𐼒𐼂𐼘𐼚 ʿ*BDt* written as 𐼘 (= RESH-AYIN-DALETH) (K 4.1).



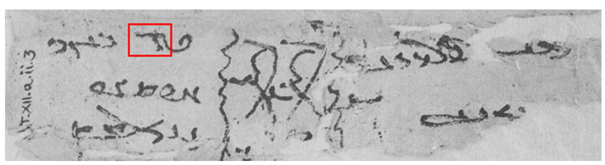
Written form of *daleth* in 𐼒𐼘 ʿ*D* written as 𐼘 (= RESH-AYIN-DALETH) (AL 1.1).



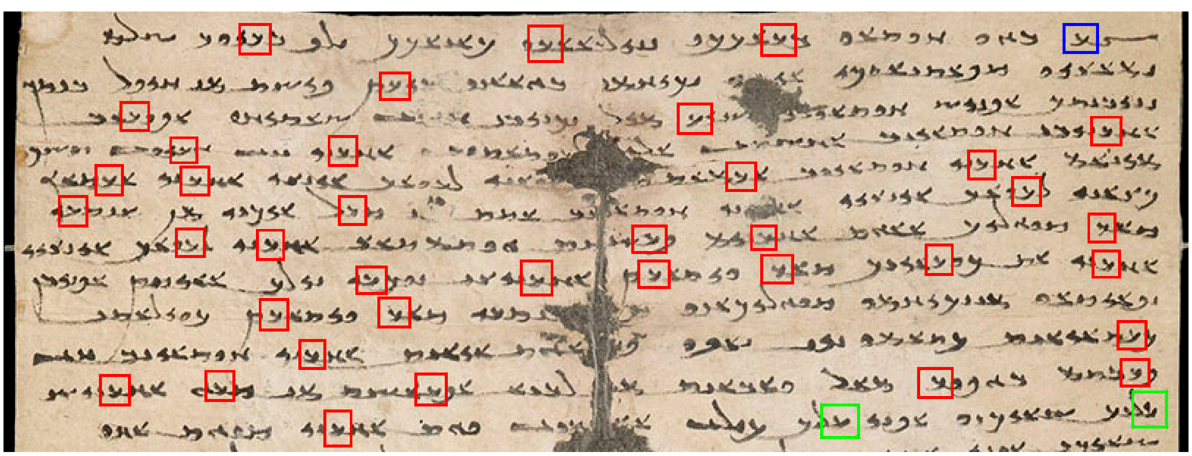
Written form of *daleth* in 𐼓𐼘 ʿ*D* written as 𐼘 (= RESH-AYIN-DALETH) (AL 2.1).



Written form of *daleth* in 𐼒𐼘 ʿ*D* written as 𐼘 (= RESH-AYIN-DALETH) (AL 3.1).



The letter *daleth* in 𐼒𐼘 ʿ*D* written as 𐼘 (= RESH-AYIN-DALETH) (AL 3 verso).



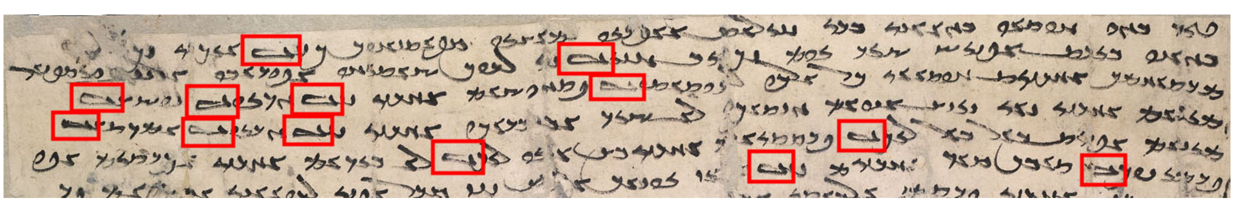
Usage of 𐼘 (= RESH-AYIN-DALETH) for representing *daleth* (blue), *ayin* (green), and *resh* (red) (AL 2.1–12).

Figure 4: Specimens of *daleth*.

Inscriptional form of 𐼅 HE in 𐼗𐼈𐼚𐼅 *ṢWRH* and (𐼅)𐼈𐼅 *(H)WH* (K 2.3–4).

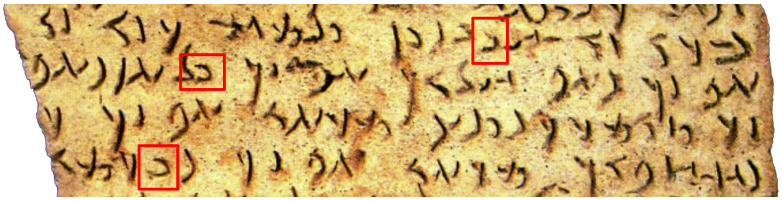
Inscriptional form of 𐼅 HE in 𐼊𐼐𐼅 *ZNH*, 𐼍𐼐𐼜𐼅 *knth*, 𐼜𐼏𐼅 *TMH* (K 4.1–2).

Written forms of 𐼆 FINAL HE in 𐼊𐼐𐼆 *ZNH* and 𐼍𐼐𐼚𐼆 *knδh* (AL 2.10, 12).

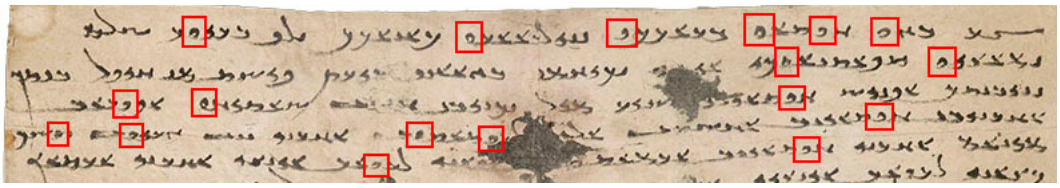


Ubiquitous usage of 𐼆 FINAL HE in AL 3.1–6.

Figure 5: Specimens of *he*.



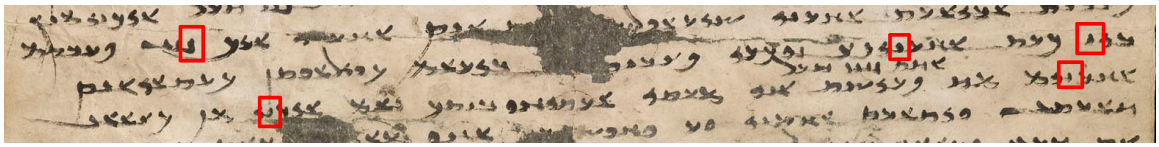
Inscriptional forms of 𐼅 WAW (K 4.2–4).



Written forms of 𐼅 WAW (AL 2.1–5).

Figure 6: Specimens of *waw*.

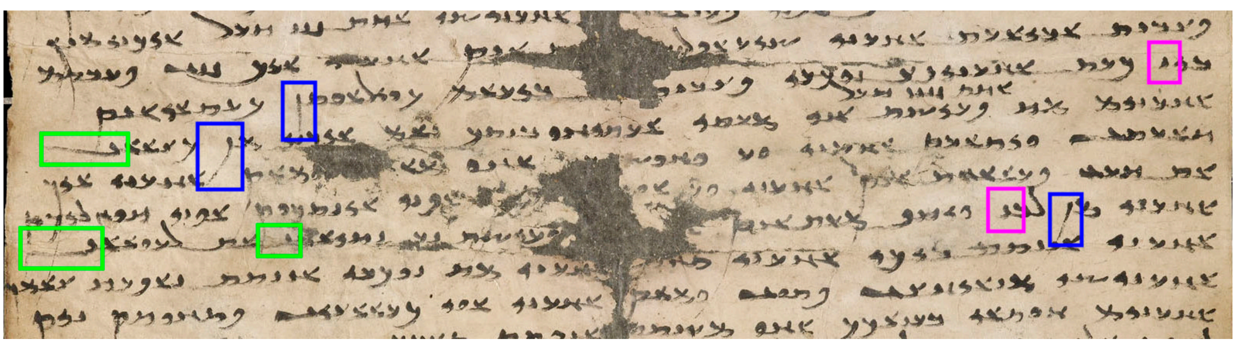
Inscriptional form of 𐼙 ZAYIN (K 4).



Written form of 𐼙 ZAYIN (AL 2.34–36).

Figure 7: Specimens of *zayin*. See also figure 8.

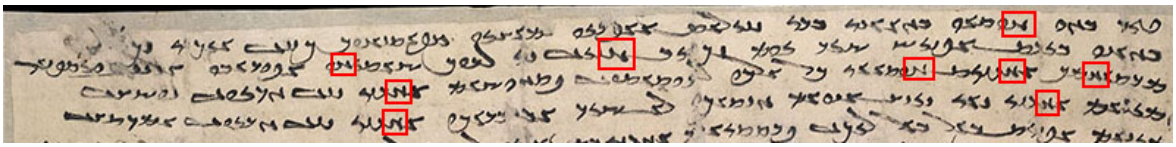
Inscriptional forms of 𐼈 zayin and 𐼎 nun in K 4: 𐼈𐼎𐼅 *ZNH* (line 1) and 𐼄𐼎𐼈 *GNZ* (line 6).



Written forms of 𐼈 *zayin* (magenta) and *nun* at the end of word (AL 2.33–41). Final *nun* is represented using both 𐼏 final nun (green) and 𐼐 final nun with vertical tail (blue).

Figure 8: Comparison of *zayin* and *nun*. See also figure 14.

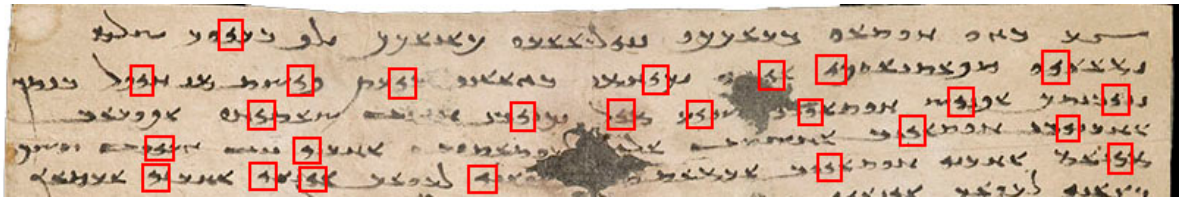
Inscriptional forms of 𐼇 HETH (K 4.3–7).



Written forms of 𐼇 HETH (AL 3.1–4).

Figure 9: Specimens of *heth*.

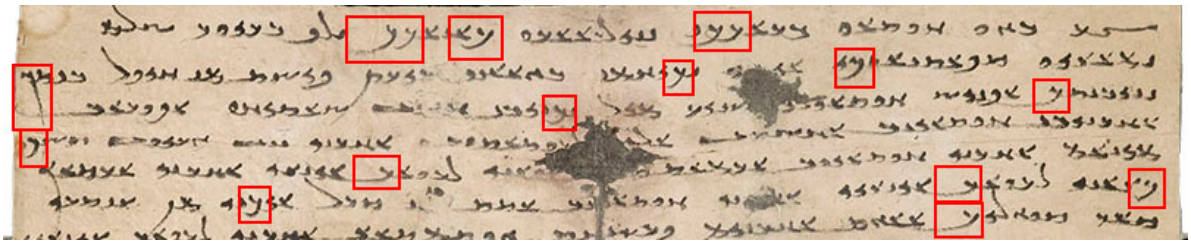
Inscriptional forms of 𐼹 YODH (K 4.1–3).



Written forms of 𐼹 YODH (AL 2.1–5).

Figure 10: Specimens of *yodh*.

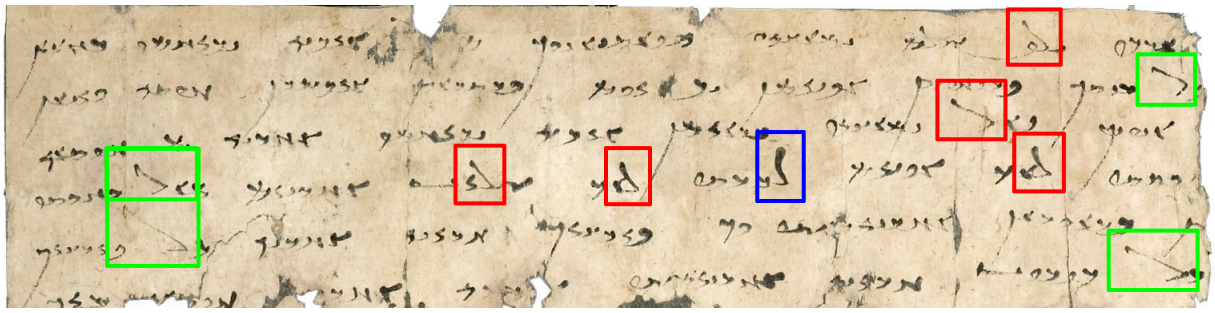
Inscriptional forms of 𐼫 KAPH (K 4.1–3).



Written forms of 𐼫 KAPH (AL 2.1–4).

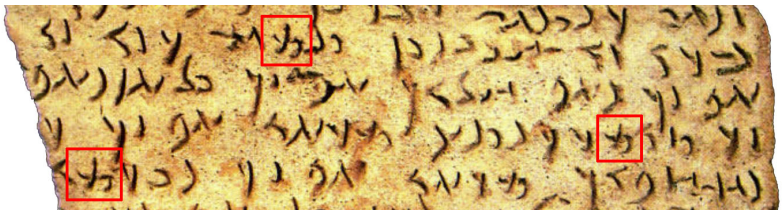
Figure 11: Specimens of *kaph*.

Inscriptional, archaic form 𐼛 of 𐼚 LAMEDH (K 4.1).

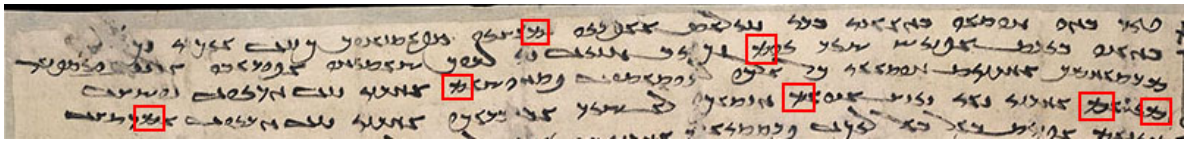


Written forms of 𐼚 LAMEDH (red) and its variant forms 𐼚 (green) and 𐼛 (blue) (AL 6.1–8).

Figure 12: Specimens of *lamedh*.

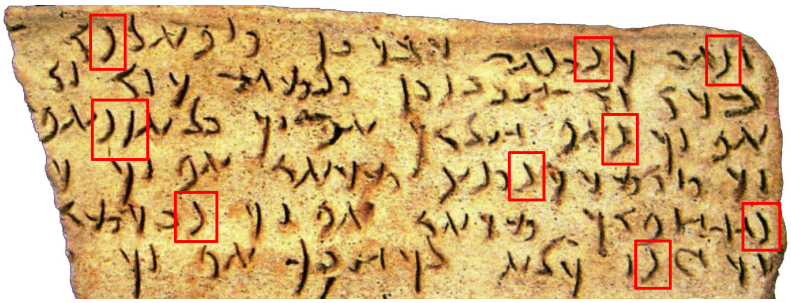


Inscriptional forms of 𐼧 MEM (K 4.1–3).

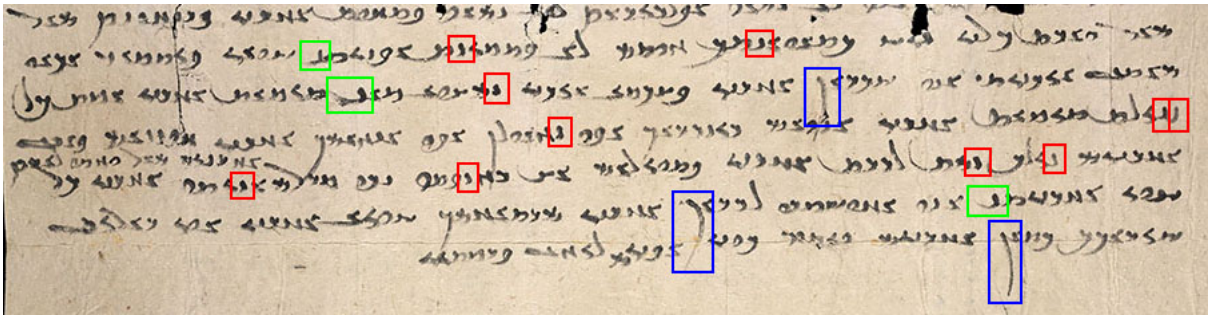


Written forms of 𐼧 MEM (AL 3.1–4).

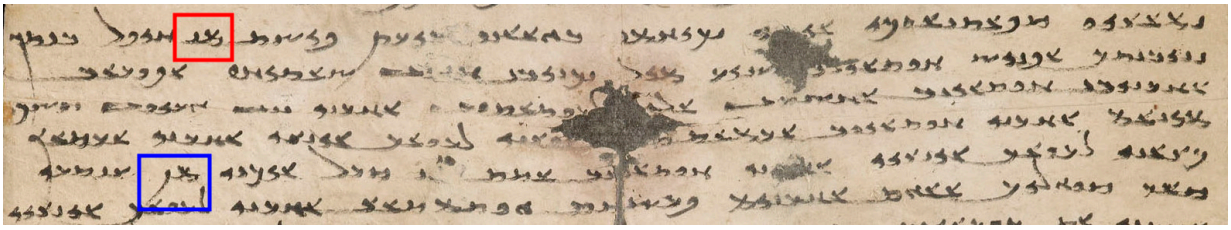
Figure 13: Specimens of *mem*.



Inscriptional form of 𐼍 NUN (K 4).



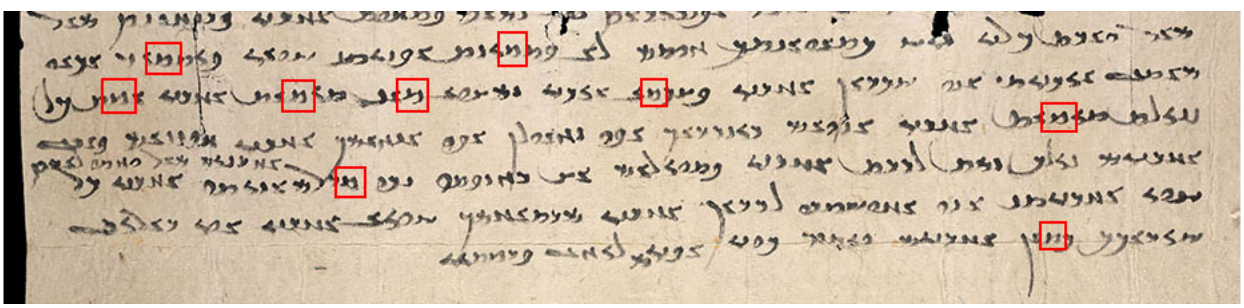
Written forms of 𐼍 NUN (red), 𐼎 FINAL NUN (green), 𐼏 FINAL NUN WITH VERTICAL TAIL (blue) (AL 1.7–12).



Usage of 𐼎 FINAL NUN (red) and 𐼏 FINAL NUN WITH VERTICAL TAIL (blue) in the word *MN*: 𐼀𐼎 and 𐼀𐼏 (AL 2.2–7).

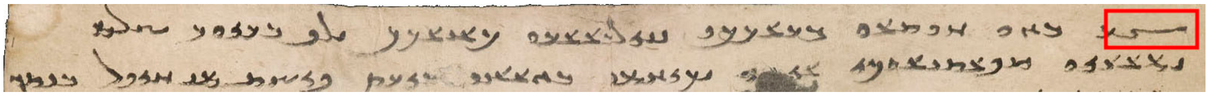
Figure 14: Specimens of *nun*. See also figure 8.

Archaic form 𐼓 of 𐼓 SAMEKH (K 4.1-4).



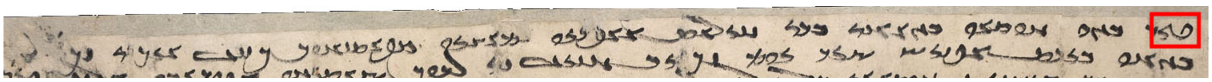
Written forms of 𐼓 SAMEKH (AL 1.7–12).

Figure 15: Specimens of *samekh*.



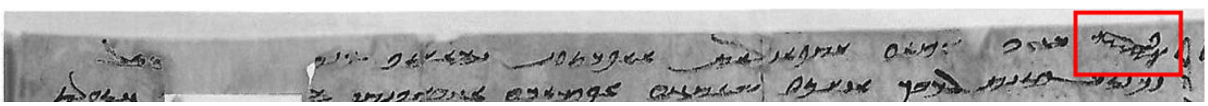
The letter *ayin* in 𐼒𐼘 *ʿD* written using 𐼒 AYIN (AL 2.1).

The *ayin* in 𐼓𐼘 *ʿD* written using 𐼓 ALTERNATE AYIN (AL 1.1).



The *ayin* in 𐼓𐼘 *ʿD* written using 𐼓 ALTERNATE AYIN (AL 3.1).

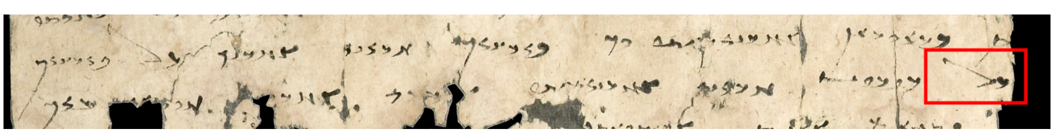
The *ayin* in 𐼓𐼘 *ʿD* written using the glyphic variant 𐼓 of 𐼓 ALTERNATE AYIN (AL 3 verso).



The *ayin* in 𐼓𐼘 *ʿD* written using the glyphic variant 𐼓 of 𐼓 ALTERNATE AYIN (AL 5.1).

The letter *ayin* in 𐼘𐼂𐼘𐼛 *ʿBDt* inscribed as 𐼘 (= RESH-AYIN-DALETH) (K 4.1).

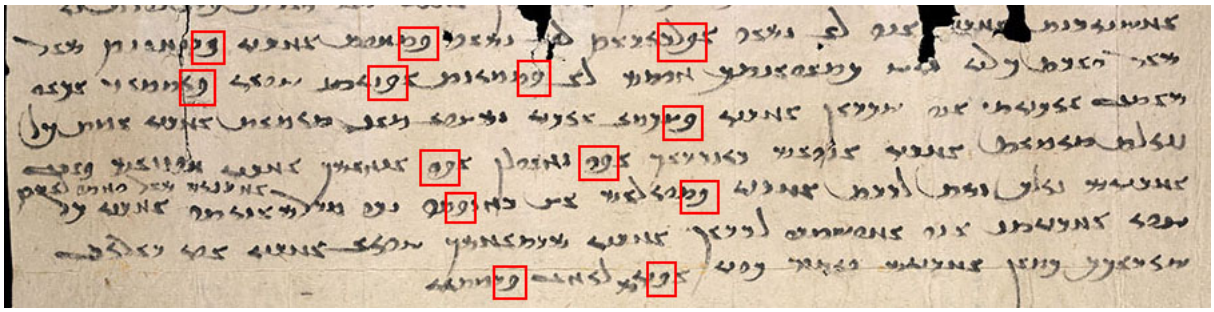
The letter *ayin* in 𐼘𐼌𐼈𐼋 *ʿLZK* written using 𐼘 (= RESH-AYIN-DALETH) (AL 2.12).



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

Figure 16: Specimens of *ayin*.

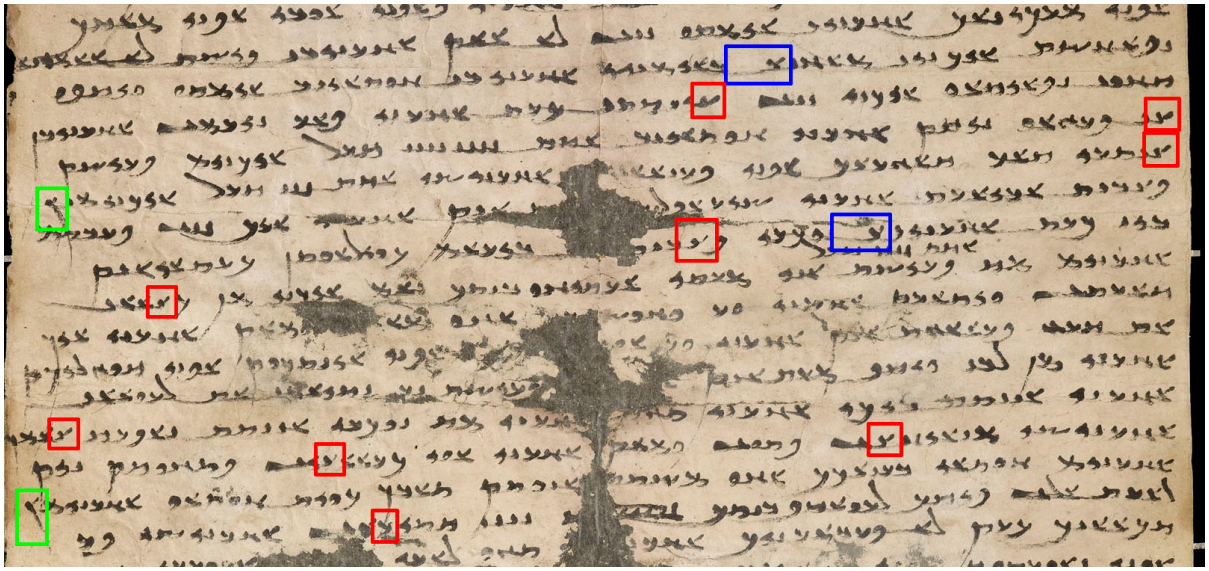
Glyphic variant 𐼟 of 𐼝 PE (K 4.1–6).



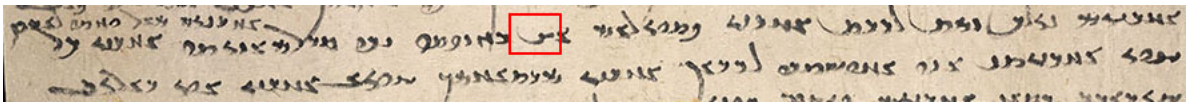
Written forms of 𐼝 PE (AL 1.6–12).

Figure 17: Specimens of *pe*.

Inscriptional form of 𐼴 SADHE (K 4.1–6).

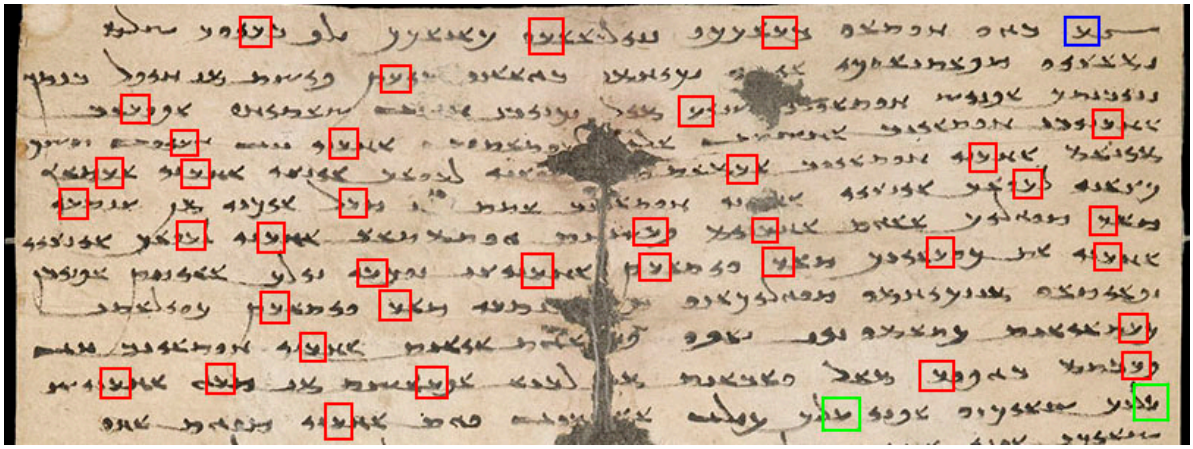


Written forms of 𐼴 SADHE (red), 𐼵 FINAL SADHE (blue), and 𐼶 FINAL SADHE WITH VERTICAL TAIL (green) in (AL 2)



Curved variant 𐽅 of 𐼵 FINAL SADHE (AL 1.10).

Figure 18: Specimens of *sadhe*.

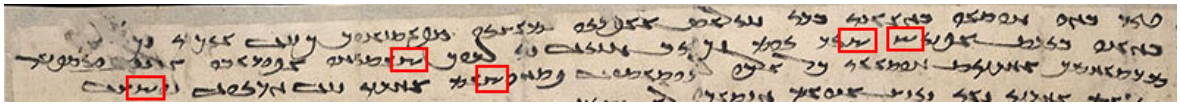


Usage of 𐼙 for representing *daleth* (blue), *ayin* (green), and *resh* (red) (AL 2.1–12). As shown, 𐼙 is most commonly used for *resh*. The letter 𐼙 is proposed for encoding as the unified character RESH-AYIN-DALETH.

Figure 19: Comparison of *daleth*, *ayin*, and *resh*.

Inscriptional forms of 𐼴 SHIN (K 4.1–3).

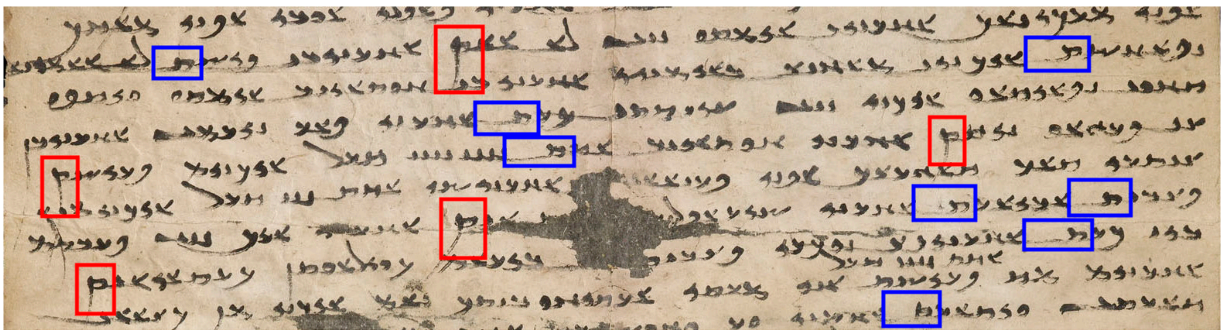
Written forms of 𐼴 SHIN (AL 2.1–4).



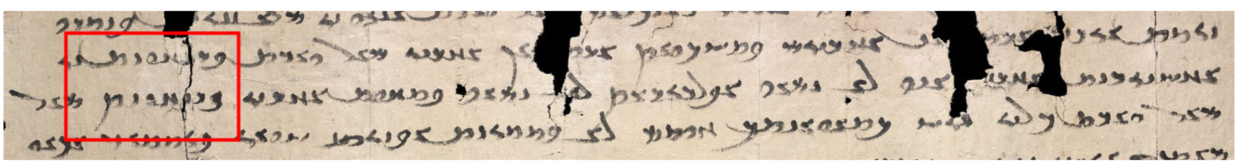
Written forms of 𐼴 SHIN (AL 3.1–3).

Figure 20: Specimens of *shin*.

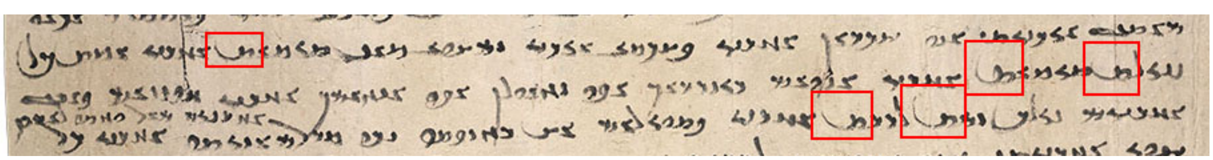
Inscriptional, archaic forms 𐼚 (red) and 𐼜 (blue) of 𐼚 TAW and 𐼜 FINAL TAW WITH VERTICAL TAIL (K 4.1–2). The distinction is apparent in 𐼙𐼇𐼚𐼜 *šwtt* (line 2), which contains both nominal and final forms.



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

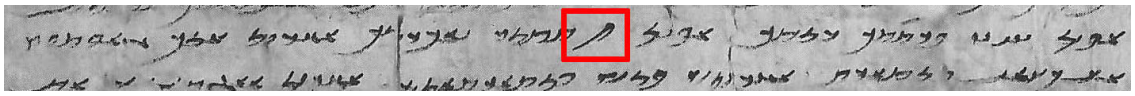


Contrasive usage of 𐼛 FINAL TAW and 𐼜 FINAL TAW WITH VERTICAL TAIL in two instances of the word *prnxwnt*: 𐼔𐼘𐼎𐼉𐼇𐼎𐼛 and 𐼔𐼘𐼎𐼉𐼇𐼎𐼜 (AL 1.5–6).



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

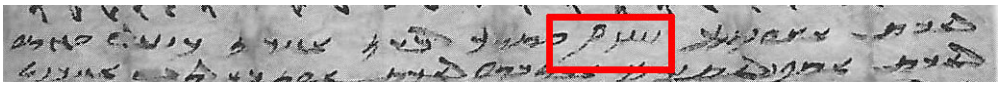
Figure 21: Specimens of *taw*.



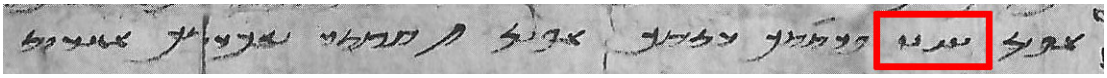
The fraction ½ 𐼰 (AL 5.10).

The number 3 𐼳 (AL 2.32).

The number 4 𐼳 𐼴 (AL 5.26).



The number 4½ 𐼳𐼰 (AL 5.24).

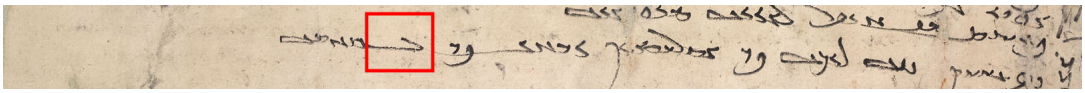


The number 5 𐼳 𐼴 (AL 5.10).

The number 7½ 𐼳𐼰 𐼳 (AL 5.26).

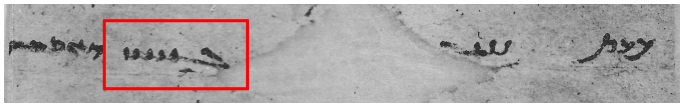
The number 8 𐼳 𐼳 (AL 2.31).

Figure 22: Examples of numbers in the ‘Ancient Letters’. See also figures 23 and 24.



The number 10 𐼴 (AL 3.26).

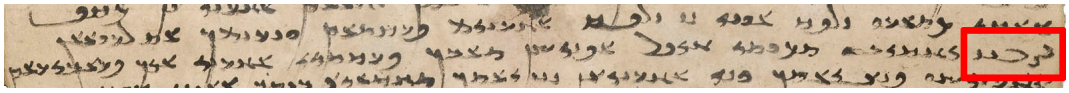
The number 13 𐼴𐼲 (AL 2.62).



The number 15 𐼴𐼲𐼲 (AL 7.8).

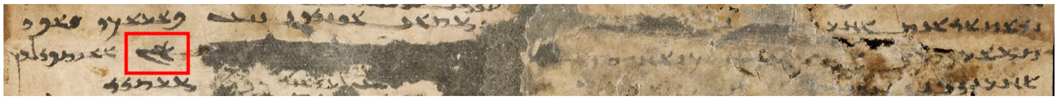
The number 20 𐼳 (AL 5.21).

The number 30 𐼷 (AL 5.32).

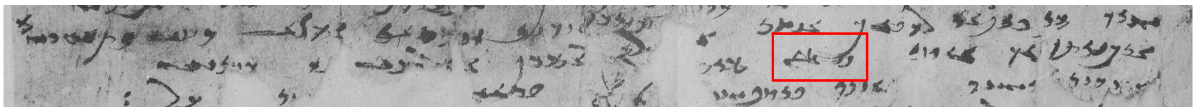


The number 32 𐼲𐼴𐼳 (AL 2.62).

Figure 23: Additional examples of numbers in the ‘Ancient Letters’. See also figures 22 and 24.

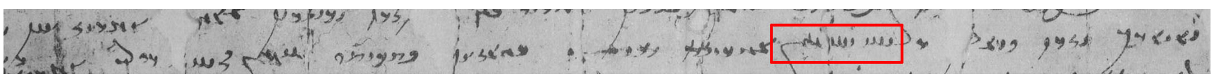


The number 100 𐼥 (AL 2.19).

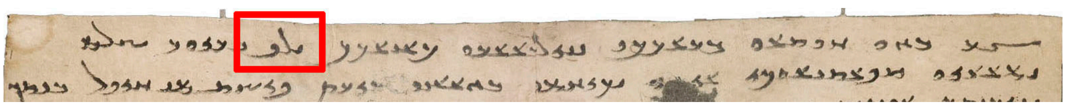


The number 200 𐼞𐼥 (AL 7.3).

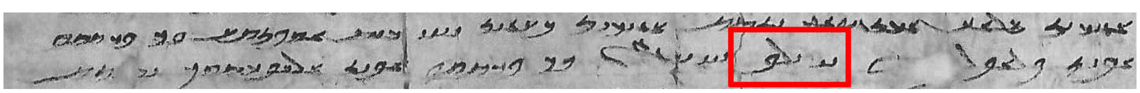
The number 500 𐼟𐼞𐼥 (AL 5.9).



The number 800 𐼠𐼠𐼥 (AL 4.3).



The number 1000 𐼧 (AL 2.1).

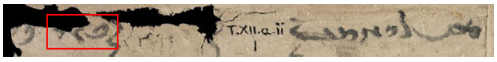


The number 2000 𐼞𐼧 (AL 5.9).

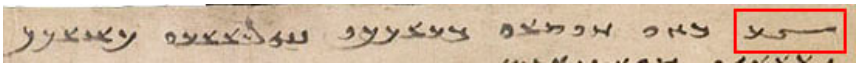
The number 10000 represented using the word 𐼂𐼘𐼊𐼇𐼘 *βrywr* (AL 2.1).

Figure 24: Further examples of numbers in the 'Ancient Letters'. See also figures 22 and 23.

The heterogram *ʿD* written as 𐼴𐼰 <𐼰 ALTERNATE AYIN, 𐼴 RESH-AYIN-DALETH> (AL 1.1).

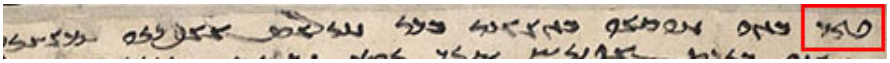


The heterogram *ʿD* written as (𐼴)𐼰 <𐼰 ALTERNATE AYIN, (𐼴 RESH-AYIN-DALETH)> (AL 1 verso).

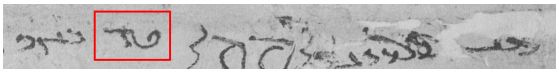


The heterogram *ʿD* written as 𐼴𐼲 <𐼲 AYIN, 𐼴 RESH-AYIN-DALETH> (AL 2.1).

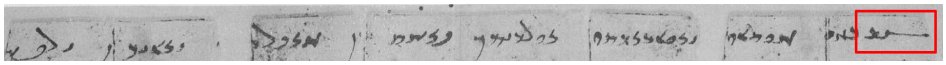
The heterogram *ʿD* written as 𐼴𐼲 <𐼲 AYIN, 𐼴 RESH-AYIN-DALETH> (AL 2 verso).



The heterogram *ʿD* written as 𐼴𐼰 <𐼰 ALTERNATE AYIN, 𐼴 RESH-AYIN-DALETH> (AL 3.1).



The heterogram *ʿD* written as the ligature 𐼶 HETEROGRAM AYIN-DALETH (AL 3 verso).



The heterogram *ʿD* written as 𐼴𐼲 <𐼲 AYIN, 𐼴 RESH-AYIN-DALETH> (AL 4.1).

The heterogram *ʿD* written as 𐼴𐼰 <𐼰 ALTERNATE AYIN, 𐼴 RESH-AYIN-DALETH> using the glyphic variant 𐼯 of ALTERNATE AYIN (AL 5.1).

Figure 25: Specimens of the heterogram *ʿD*.

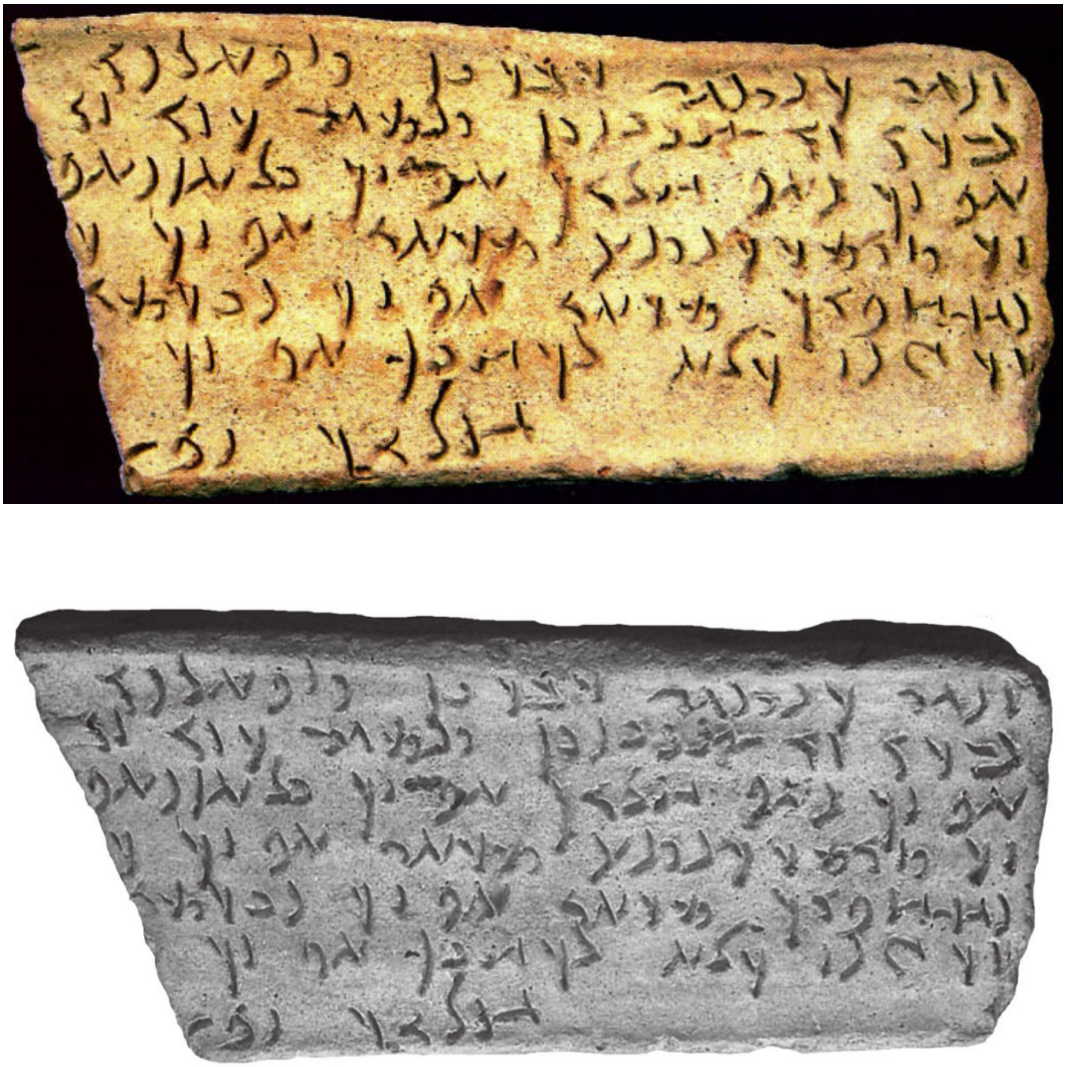


Figure 26: Two images of Kultobe inscription 4 (KII 26859/1). Top from Sims-Williams 2007; bottom from Grenet, et al 2007.

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).

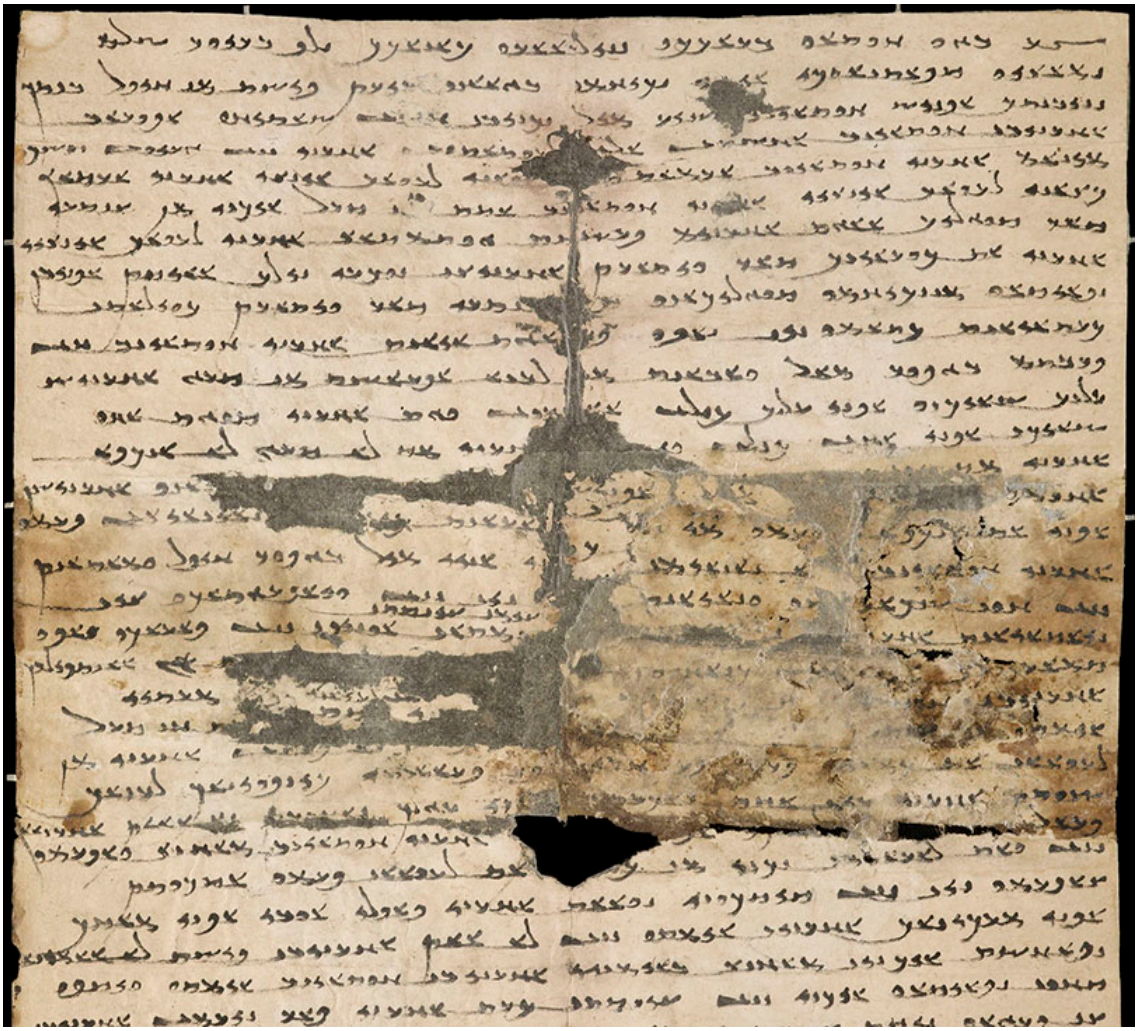


Figure 29: Top portion of 'Ancient Letter 2' (British Library, International Dunhuang Project: Or. 8212/95 side a). "To the noble lord Varzakk (son of) Nanai-thvar (of the family) Kanakk. Sent [by] his servant Nanai-vandak." (translation by Sims-Williams in Waugh 2004). Continued in figure 30.

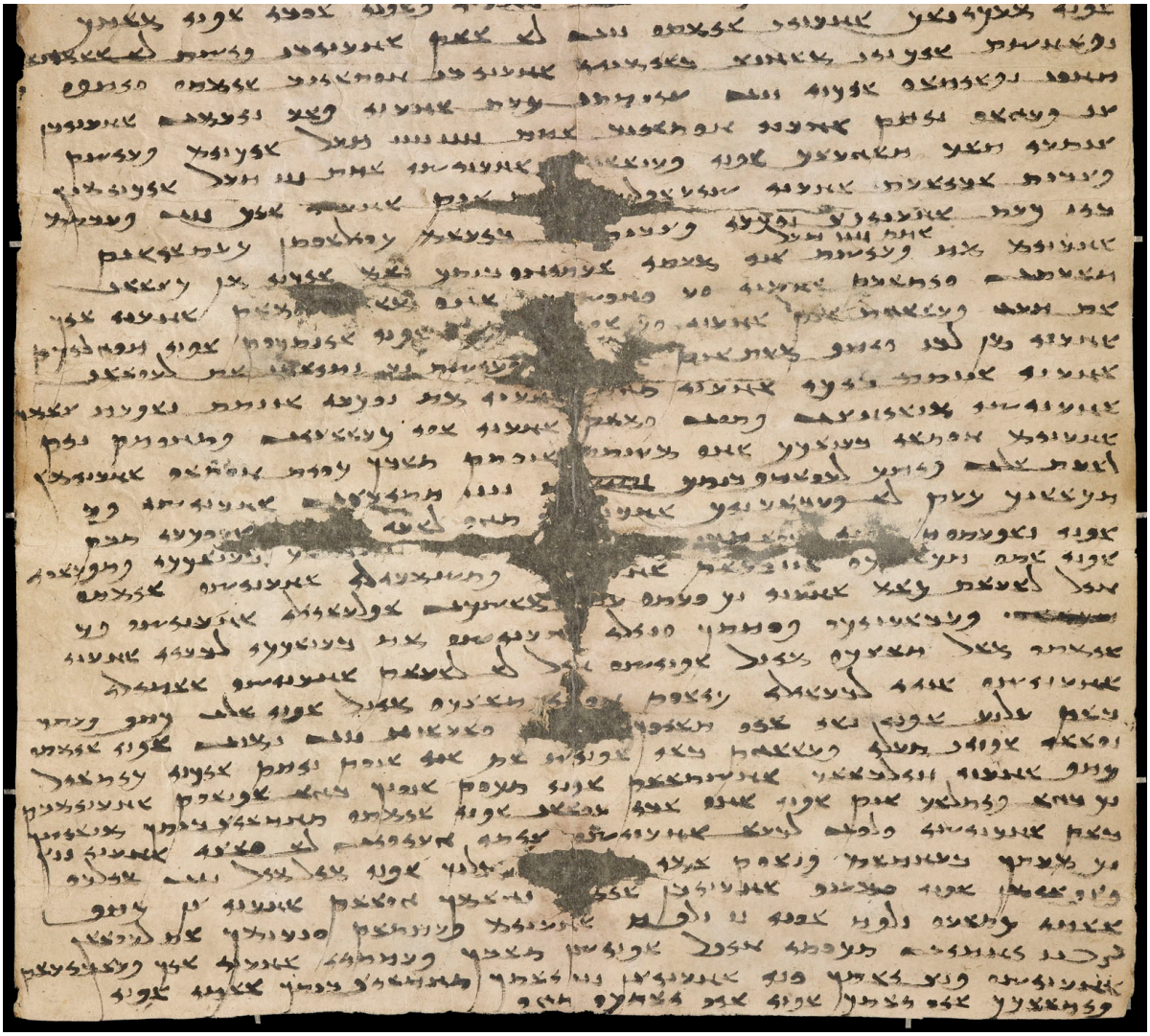


Figure 30: Bottom portion of ‘Ancient Letter 2’ (British Library, International Dunhuang Project: Or. 8212/95 side a). Continued from figure 29.

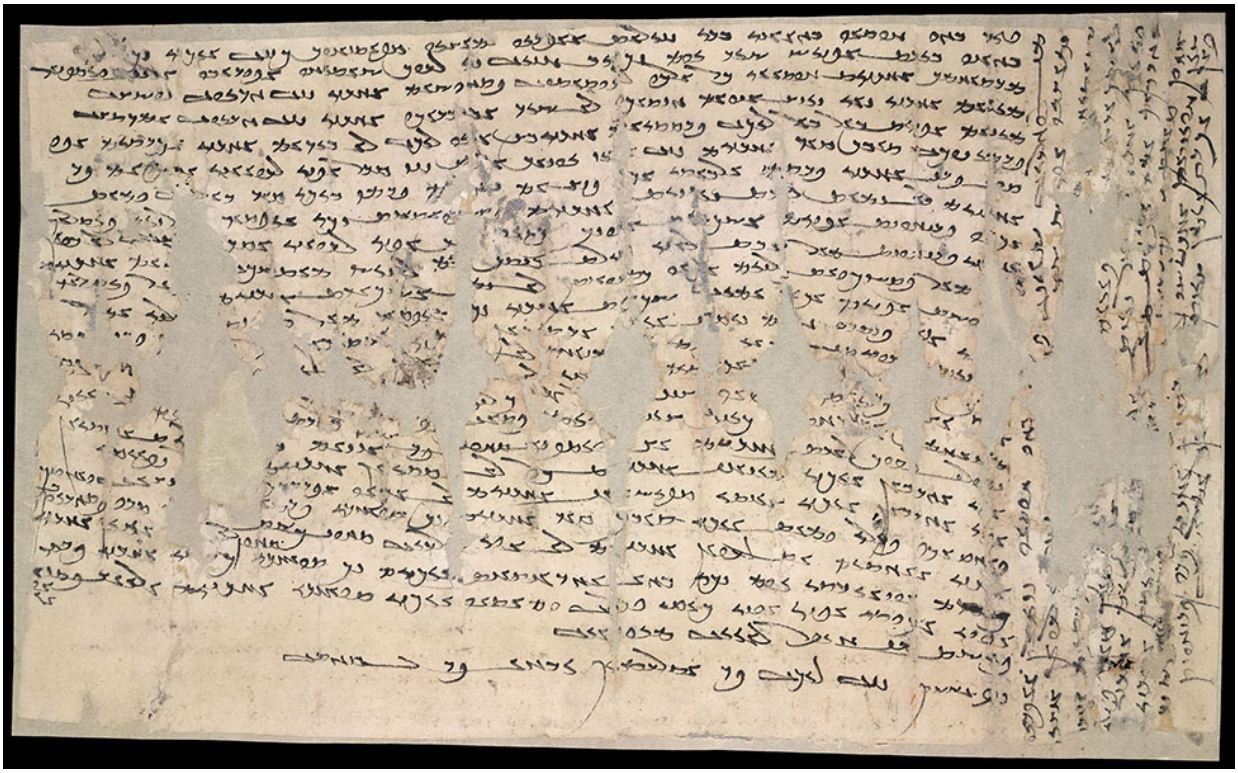


Figure 31: The ‘Ancient Letter 3’ (British Library, International Dunhuang Project: Or. 8212/98 recto 1). “To (my) noble lord (and) husband Nanai-dhat.” (translation by Sims-Williams in Waugh 2004).

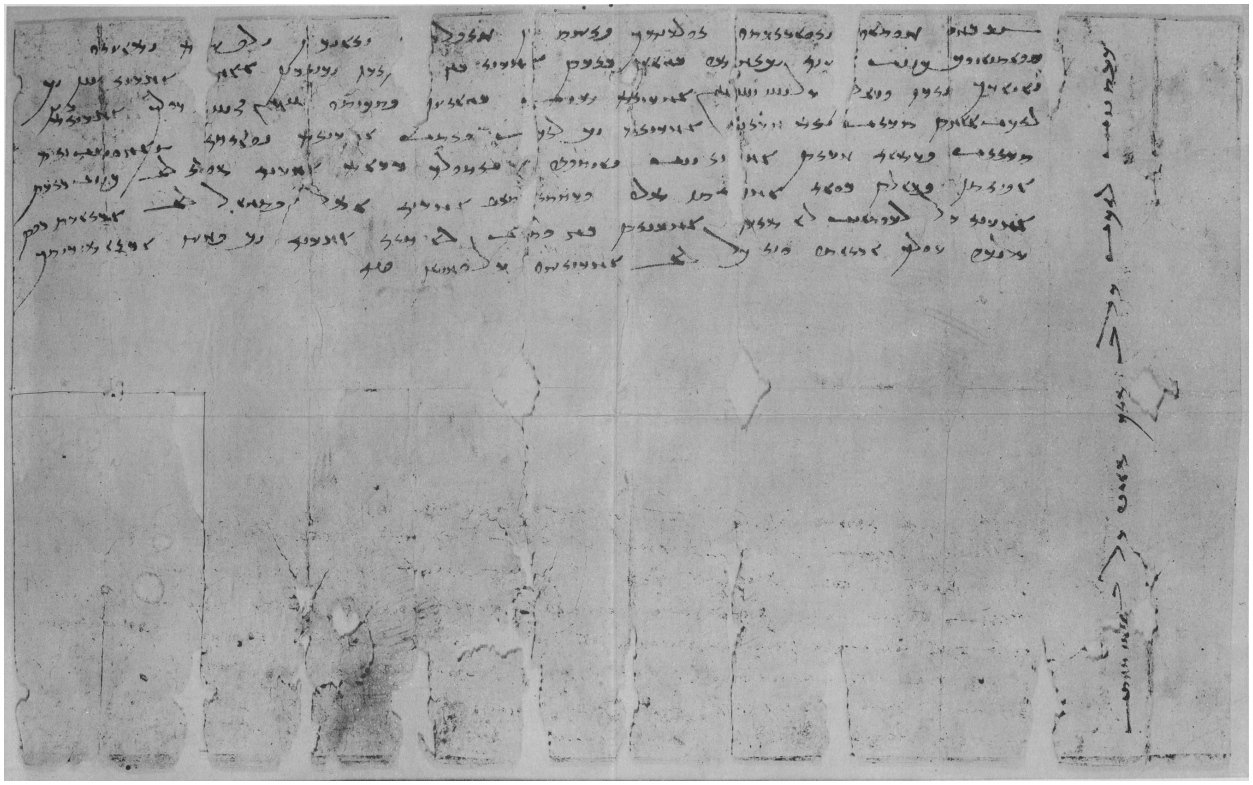


Figure 32: The ‘Ancient Letter 4’ (British Library: Or. 8212/93 recto; reproduced in Reichelt 1928: plate IV).

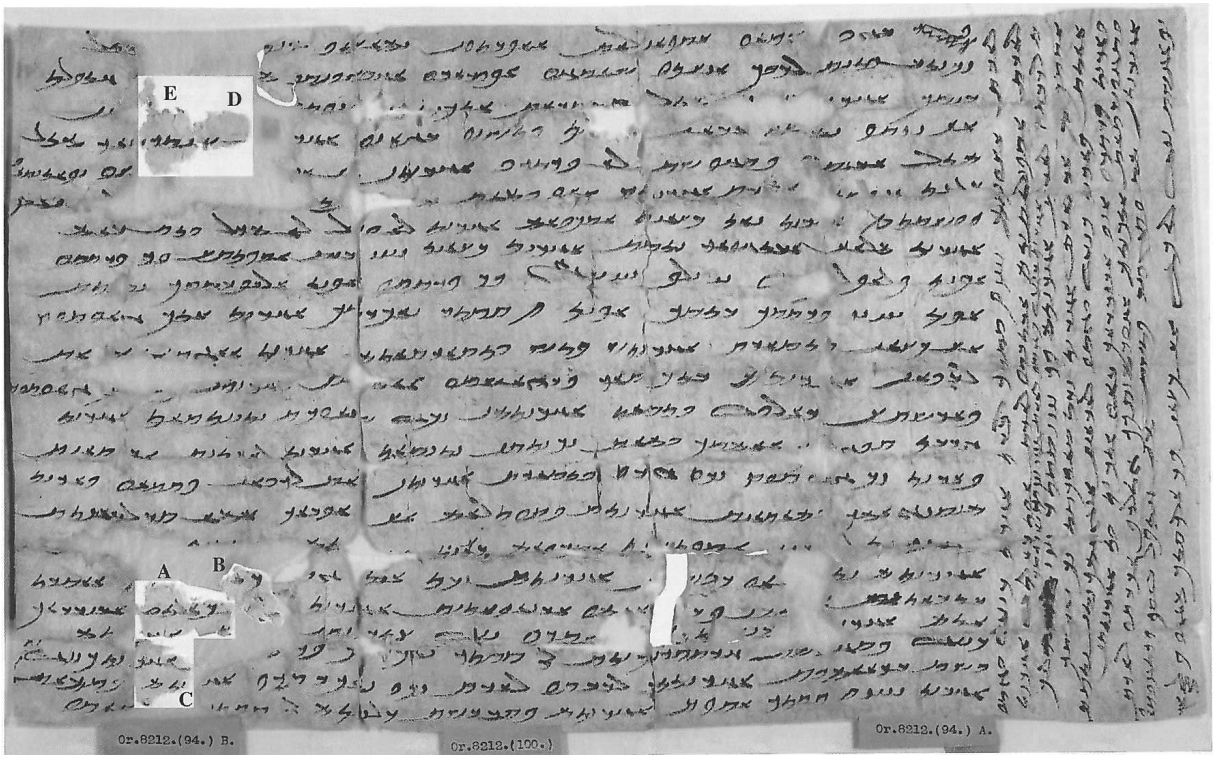


Figure 33: The ‘Ancient Letter 5’ (from Grenet, et al. 1998: 94). “To the noble lord, the chief merchant Aspandhāt. [Sent] by your servant [Frī-khwatāw].”

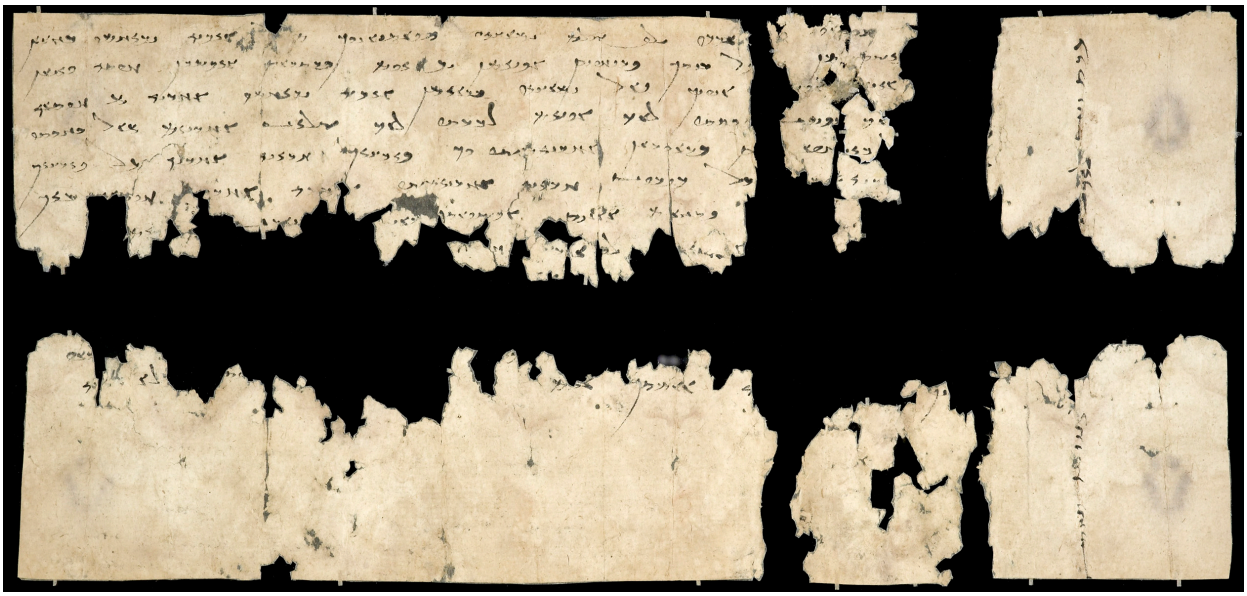


Figure 34: The ‘Ancient Letter 6’ (British Library, International Dunhuang Project: Or. 8212/97).

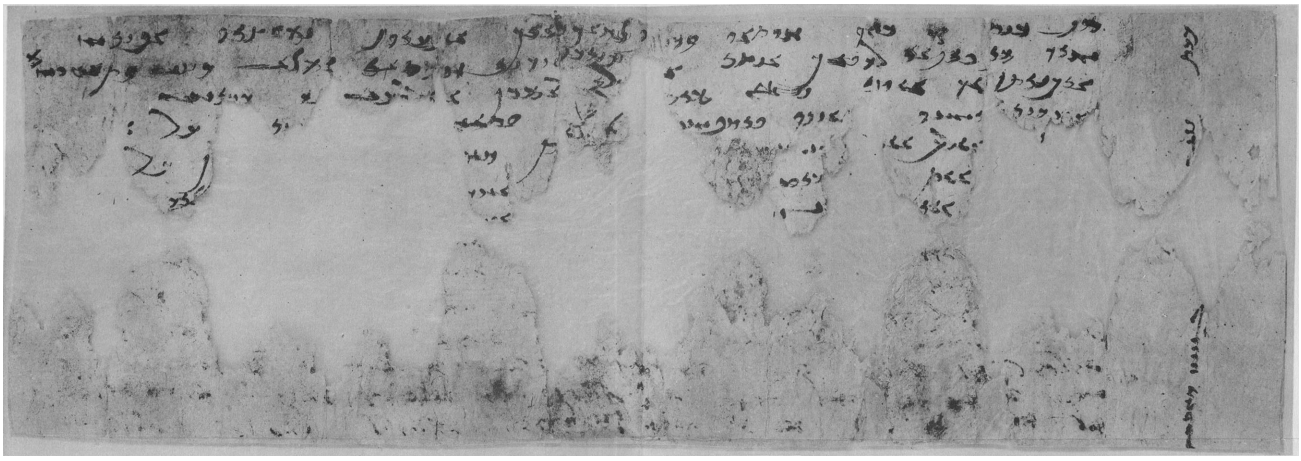


Figure 35: The ‘Ancient Letter 7’ (British Library: Or. 8212/96 recto; reproduced in Reichelt 1928: plate VII).

Figure 36: Sogdian rock inscription from Shatial (from Sims-Williams 1989: plate 10b) The inscription reads 𐼏𐼏𐼊𐼀𐼋𐼋 𐼈𐼋 *nnyʾkk ZK* (top line), 𐼒𐼇𐼀𐼂𐼙 *swʾbr* (middle), 𐼂𐼙𐼊 *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 𐼕𐼀𐼕𐼀𐼋𐼋 *pʾpʾkk* (top line), 𐼈𐼋 𐼋𐼇𐼚𐼀𐼀𐼏 *ZK kwšʾʾn* (middle), 𐼂𐼙𐼊 *BRY* (bottom). Latin transcription from *ibid*: 14. The inscription in the top left-hand corner is shown in detail in figure 36.

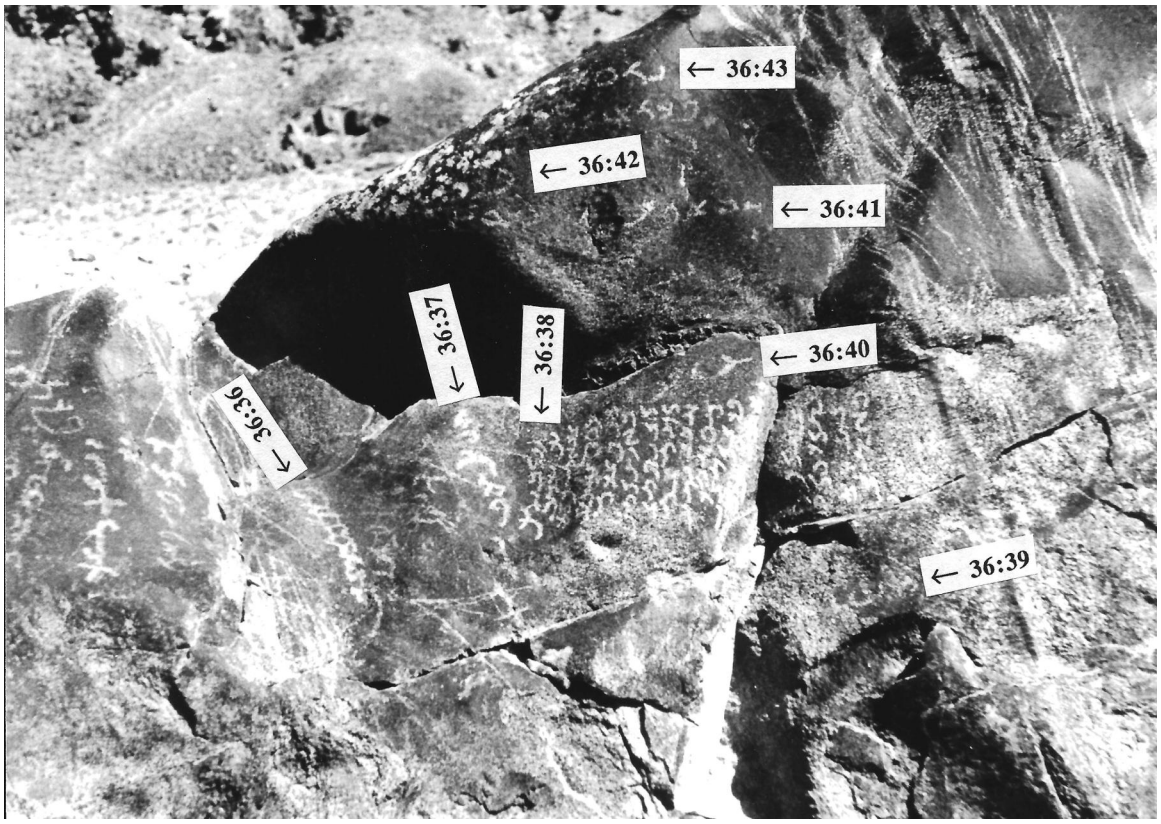


Figure 38: Rock at Shatial containing horizontal and vertical inscriptions in the Old Sogdian script (from Sims-Williams 1989: plate 109b). The text of 36:38 is shown in section 4.5.

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 *c'c'nn'pc wnwnxwr*.

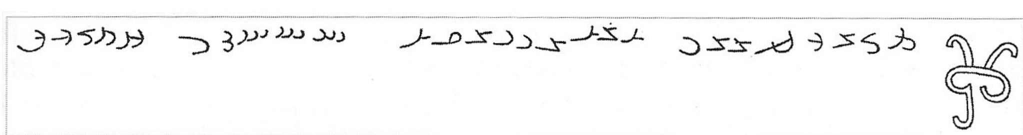


Figure 40: Reproduction of an Old Sogdian inscription on a silver vessel from the principality of Chach: *my'rx''n c'c'nn'pc 39 styrk* "Mayārkhān of the nation of Tashkent. 39 stater (ca. 624 g)." (from Yoshida 2002: 191). A *tamgha* appears to the right of the inscription.

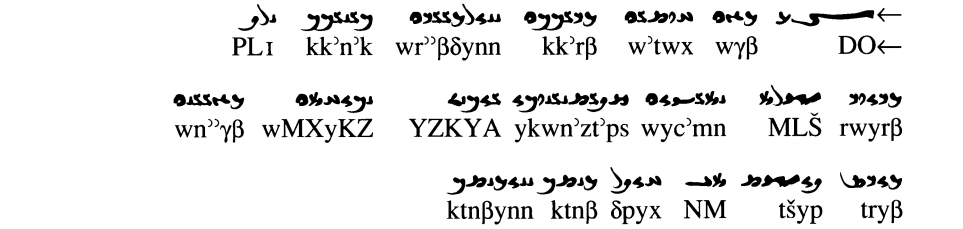
Sogdian script

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, ʾ is still distinguished from **n**; but in the later, ʾ = **n**, ʾ**n** = **n**ʾ. 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/x/s/š**, **r/β/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 **ž** by a diacritical point ◌̇, and the foreign sound *b* was noted as **ṗ**.

SAMPLES OF SOGDIAN

ANCIENT LETTERS



1. *Transliteration:*	OD	βγw	xwtʾw	βrʾkk	nnyδβʾʾrw	kʾnʾkk	
2. *Normalization:*	at	βaγu	xutāw	βarak	nanē-θβār	kanak	
3. *Gloss:*	to	lord.ACC	master	Barak	Nana's-gift	Kanak	

1. ILP	βrywr	ŠLM	nmʾcyw		spʾtzʾnwky	AYKZY
2. (ēw-)zār	βrēwar	*āfrīwan	namācyu		spātzānūk	kaδ-uti
3. thousand	ten.thousand	greeting(?)	reverence.ACC		bended.knee	when-that.and

1. ZKyXMw	βγʾʾnw	βyrt	pyšt	MN	xypθ	βntk	nnyβntk	
2. wēšanu	βaγān(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*

Aramaic	*Sogdian Ancient Letters*	*Sogdian sutra script*	*Manichean Sogdian*	*Christian Sogdian*	*Principal Phonetic Values (Sogdian)*
ʾ					a, ā
b					b, β
(β)					β
g					g, γ
(γ)					γ
d					d, δ
h (h̲)					a, Ø
w					w, ŏ, ŭ
z					z
(j)					ž
(ž)					ž
ḥ (h)					γ, x, h
ṭ					t
y					y, ĕ, ĭ
k					k
(x)					x
l (δ)					δ
m					m
n					n
s					s
ʿ					Ø
p					p
(f)					f
ṣ (c)					č, ǰ
q					k
r					r
š					š
t					t, θ

Figure 42: Table showing various scripts for writing Sogdian (from Skjærvø 1996: 519).

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 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:
   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): *Yes*
      Proposed name of script: *Old Sogdian*
   b. The proposal is for addition of character(s) to an existing block:
      Name of the existing block:
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 (X) 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 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? *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?
    If YES, reference: