L2/15-089R
2015-11-03

Preliminary proposal to encode the Old Sogdian script in Unicode

Anshuman Pandey
Department of Linguistics
University of California, Berkeley
Berkeley, California, U.S.A.
anshuman.pandey@berkeley.edu

November 3, 2015

1 Introduction

This is a preliminary proposal for a Unicode encoding for the scripts used in the Sogdian 'Ancient Letters' and in Sogdian inscriptions of the Upper Indus. These scripts are tentatively designated as 'Old Sogdian' (see section 4.1). The intent is to enable users to input, display, and exchange content in the Old Sogdian script on digital devices.

This document describes an encoding model and character repertoire, and provides specimens of sources in which the script appears. In the proposed repertoire each graphical element of the script that possesses a distinctive semantic value is defined as a Unicode 'character'. The exception are letters with identical glyphs, which are merged into a single character in accordance with the character-glyph model of the Unicode standard. This character-encoding repertoire may differ from traditional scholarly analyses of the script. Such differences naturally arise from the requirements for representing the script electronically as opposed to writing it by hand. The representative glyphs shown here are illustrative and are not intended to be typographically aesthetic.

The author seeks feedback from scholars regarding the proposed encoding model, character repertoire, and representative glyphs. Issues requiring further discussion are enumerated in section 7. A formal proposal to encode the script is forthcoming, pending comments from experts.

2 Background

The Old Sogdian script was used primarily for writing Sogdian (ISO 639: sog), an Eastern Iranian language that was spoken in the region between the Amu Darya and the Syr Darya, which roughly corresponds to the modern provinces of Samarqand and Bokhara in Uzbekistan and the Sughd province of Tajikistan. The language and script were carried by Sogdian merchants from eastern Iran to western China.

The earliest manuscripts containing the Old Sogdian script are known as the 'Ancient Letters' (see figures 1–5). These paper documents were found in 1907 by Aurel Stein in Dunhuang, now in Gansu province, western China. Based upon evidence contained within the documents, it has been suggested that the 'Ancient Letters' were written in 312–313 CE (Sims-Williams 1985). A system of writing similar to that used in the 'Ancient

Letters' appears in some 600 rock carvings at Shatial in the Gilgit region of Pakistan. These 'Upper Indus inscriptions' have been dated to the 4th–7th centuries CE (Sims-Williams 1989, 2000; see figures 6 and 7 here). Based upon the structural and graphical similarities between the scripts of the 'Ancient Letters' and the 'Upper Indus inscriptions', they may be considered to be typologically identical.

The script bears resemblance to Aramaic, Inscriptional Parthian, and Inscriptional Pahlavi. A comparison of these scripts is provided in table 5.

The Old Sogdian script is the focus of active scholarship. The International Dunhuang Project at the British Library is currently digitizing texts in Old Sogdian, namely the 'Ancient Letters'. The ability to represent the contents of these documents in plain text for scholarly exchange, cataloguing, and the production of corpora requires an encoding for Old Sogdian in Unicode.

3 Roadmap for Sogdian scripts

More than a decade ago, in an overview on the encoding of additional Semitic scripts it was suggested that the Aramaic block would be expected to encompass 'Sogdian' (Everson 2001). The document did not define 'Sogdian'. It may be inferred that the reference was most likely to the 'formal' Sogdian script, which is the most well known. There are, however, several varieties of typologically related scripts that are known by designation 'Sogdian':

1. the script of the oldest Sogdian inscriptions at Kultobe
2. the script of the 'Ancient Letters' from Dunhuang
3. the script of the Upper Indus inscriptions at Shatial
4. the 'cursive' script of documents from Mount Mug
5. the formalized 'cursive' script used in Buddhist texts, known as the 'formal' or 'sūtra' script
6. the latest cursive form, known as the 'Sogdian Uyghur' or 'Turco-Sogdian' script

In relation to the above 'proper Sogdian' scripts, mention may be made here to a derivative of 'Sogdian Uyghur' known as 'Old Uyghur', which is written vertically. The Sogdian language was also written in Syriac, Manichaean, and Brahmi-based scripts. All of these scripts are distinctive writing systems in their own right and some have already been encoded in Unicode.

None of the scripts belonging to the 'proper Sogdian' family are encoded in Unicode. It is, therefore, beneficial to describe a possible roadmap for handling the encoding of these scripts. The available information suggests that these scripts may be divided across three encodings:

- 'Old Sogdian': to encompass scripts #2 and #3, and possibly #1
- 'Cursive Sogdian': for script #4
- 'Sogdian': a unified block for scripts #5 and #6

These divisions are based upon the orthographic structure, letterforms, and character repertoires of the scripts. Scripts #1–#3 are structurally non-conjoining, while #4–#6 are inherently conjoining. The letters of #1–#3 retain their basic shapes in words, while letters of #4–#6 have contextual shapes based upon their position in a word. The letterforms of these six scripts may be classified into two groups based upon similarity of graphical structures: letters of #1–#4 exhibit common features, but differ from letter shapes of #5–#6. Furthermore, scripts #5–#6 use diacritics to indicate different phonetic values of a particular letter. Although these scripts are typologically related, their differing structures require different technical implementations, and therefore separate encodings. A comparison of scripts #2 and #4–#6 is shown in figure 10.

4 Script Details

4.1 Script name

The name ‘Old Sogdian’ is assigned for the script in Unicode in order to distinguish it from related scripts, but also to emphasize their genetic relationship. There is no standard or conventional name for the script of the ‘Ancient Letters’ or the ‘Upper Indus inscriptions’. The catalogue of the International Dunhuang Project at the British Library refers to the script in the ‘Ancient Letters’ generically as “Sogdian” and does not differentiate it from other related scripts, which are labelled using the same name. Skjærvø (1996) refers to the script as “Sogdian Aramaic”.

4.2 Structure

Old Sogdian is an *abjad* that is written from right to left. Structurally, it is a non-joining *abjad*, similar to Hebrew. Letters retain their shapes within a word; however, some letters have distinctive word-final shapes (see section 5.1.1). The available sources show instances where adjacent letters are connected by joined or overlapping strokes, but such conjunctions result from the regular flow of writing rather than any intrinsic conjoining behavior of the letters, as is the case for Arabic, Mongolian, and the later ‘cursive, ‘sūtra’ and ‘Uyghur’ Sogdian scripts. Similar to other *abjad* writing systems, vowels in Old Sogdian are represented using *aleph*, *yodh*, and *waw*; however, these are also used for short vowels, not only long forms.

4.3 Encoding model

The arrangement of core letters in the code block proposed for Old Sogdian is based upon a one-to-one correspondence with the Aramaic block. However, Old Sogdian analogues do not exist for every Aramaic letter. Only 20 Old Sogdian letters corresponding to the original 22 Aramaic are attested in the available sources. The repertoire lacks analogues for Aramaic *teth* and *qoph*. The number of distinctive letters is further reduced on account of the merging of glyphs for two or more letters into forms that are nearly identical or indistinguishable. Such is the case for *zayin* and *nun*; as well as for *ayin* and *resh*, and quite possibly *daleth*, which could be assimilated with these two. The result is a repertoire that contains 17 or 18 distinctive letters.

The encoding model for Old Sogdian differs from the Unicode models for other Iranian scripts derived from Aramaic. For the Inscriptional Parthian and Inscriptional Pahlavi encodings letters with nearly identical glyphic representations are merged into single letters. For instance, in Inscriptional Pahlavi the letters *waw*, *ayin*, and *resh* are represented using 𐭥 U+10B65 INSCRIPTIONAL PAHLAVI LETTER WAW-AYIN-RESH. A similar case occurs with Inscriptional Pahlavi *mem* and *qoph*, which are merged into the character 𐭬 U+10B6C INSCRIPTIONAL PAHLAVI LETTER MEM-QOPH.

The proposed arrangement for the Old Sogdian block departs from the pattern for Inscriptional Parthian and Inscriptional Pahlavi. These latter script blocks are not arranged in a one-to-one correspondence with Aramaic and do not have spaces reserved within them for potential disunification of merged letters. For Old Sogdian, spaces have been reserved for *teth* and *qoph* in the event that these characters are discovered in the future. The *nun* may be perceived to be identical to *zayin*, and therefore it is not encoded as separate character. But a space is reserved for it in the case that a distinctive form is identified later. Until that time, *zayin* is to be also used for *nun*. Similarly, there is no distinctive character for *ayin*, but a space is reserved for it. The letter *resh* is to be also used for *ayin*. This approach is preferred over merging these letters into single characters such as *OLD SOGDIAN LETTER ZAYIN-NUN and *OLD SOGDIAN LETTER AYIN-RESH.

The rationale for aligning the Old Sogdian code block with the full Aramaic repertoire is based upon a source from the late 7th century or early 8th century CE. An ostracon found at Panjikant, Tajikistan bears an inscription that enumerates twenty-three letters of the ‘cursive’ Sogdian script; the last letter is a repetition of *lamedh* (Livshits 2015: 227). This source suggests that knowledge and memory of the twenty-two letter repertoire persisted among Sogdian scribes well after the period of the ‘Ancient Letters’ and the ‘Upper Indus inscriptions’. Even if the repertoire etched upon the potsherd contains vestigial letters, such as *qoph*, a character-encoding standard for that script should provide a means for full representation. If a complete alphabet exists for ‘cursive’ Sogdian, then it is quite possible that something similar exists for Old Sogdian. For that reason it is practical to reserve spaces in the proposed code block for letters that appear to have disappeared in case that they are identified in future discoveries.

Word-final forms of several letters occur in the available sources. A complete list of distinctive word-final forms has yet to be produced.

4.4 Representative glyphs

Old Sogdian letters exhibit a high degree of glyphic uniformity across the available sources, as observed in the specimens of the ‘Ancient Letters’ and ‘Upper Indus inscriptions’ consulted by the proposal author. The representative glyphs for Old Sogdian characters are normalized forms of characters found in the ‘Ancient Letters’. The glyphs for letters are derived from Skjærvø (1996), shown here in figure 9, and have been modified by the proposal author. Glyphs for numbers have been drawn by the proposal author.

5 Proposed characters

The proposed repertoire for Old Sogdian contains 30 characters: 18 basic letters, 2 alternate letters, 9 numbers, and 1 fraction. Word-final forms of letters are not included at present. Names for basic letters align with corresponding characters of the ‘Imperial Aramaic’ block.

5.1 Letters

The following 18 basic letters are included in the proposed repertoire:

	Character name	Phonetic values
𐼀	OLD SOGDIAN LETTER ALEPH	a, ā
𐼂	OLD SOGDIAN LETTER BETH	β
𐼄	OLD SOGDIAN LETTER GIMEL	γ
𐼘	OLD SOGDIAN LETTER DALETH	d
𐼅	OLD SOGDIAN LETTER HE	Ø, a
𐼇	OLD SOGDIAN LETTER WAW	w, u, ū, o, ō
𐼈	OLD SOGDIAN LETTER ZAYIN	z, ž ; n
𐼉	OLD SOGDIAN LETTER HETH	x
𐼊	OLD SOGDIAN LETTER YODH	y, i, ī, e, ē

𐼍	OLD SOGDIAN LETTER KAPH	k, g
𐼎	OLD SOGDIAN LETTER LAMEDH	δ, l
𐼏	OLD SOGDIAN LETTER MEM	m
𐼓	OLD SOGDIAN LETTER SAMEKH	s
𐼖	OLD SOGDIAN LETTER PE	p, b, f
𐼗	OLD SOGDIAN LETTER SADHE	č, ǰ
𐼚	OLD SOGDIAN LETTER RESH	r ; Ø, a
𐼛	OLD SOGDIAN LETTER SHIN	š
𐼜	OLD SOGDIAN LETTER TAW	t, d

The following 2 alternate letters are included:

	Character name	Phonetic value
𐼇	OLD SOGDIAN LETTER ALTERNATE HE	a
𐼕	OLD SOGDIAN LETTER ALTERNATE AYIN	Ø

Notes on the repertoire

1. The letter 𐼚 DALETH is used only in Aramaic heterograms.
2. The letter 𐼆 HE is used only at the end of words. It has a variant form 𐼇 that occurs concurrently in various documents, which is proposed for encoding as the separate character 𐼇 ALTERNATE HE in order to enable the representation of both.
3. The letters *zayin* and *nun* have nearly identical, if not indistinguishable shapes. For this reason, a *nun* is not proposed for encoding as a separate letter. Instead, the letter ZAYIN is to be used for both *zayin* and *nun*.
4. The letters *ayin* and *resh* also have indistinguishable forms. A separate character for *ayin* is not proposed. Instead, the letter RESH is to be used for both *ayin* and *resh*.
5. In the ‘Ancient Letters’, the letter 𐼕 ALTERNATE LETTER AYIN occurs only in Aramaic heterograms, such as 𐼚𐼕 *OD* <𐼕 ALTERNATE LETTER AYIN, 𐼚 DALETH>, a preposition meaning “to” that is used in terms of address in correspondence.
6. Some scholars proposed that *qoph* was retained and reassigned for the number 100 (Sims-Williams 1985); however, the common position is that *qoph* was never used in the Old Sogdian script.

5.1.1 Word-final shapes

In the ‘Ancient Letters’ certain letters are written using glyphically distinctive forms when they occur in word-final position. These final forms are distinguished from regular forms by the shape of the terminal stroke. A curved stroke may be modified into a vertical line or the curve may be rounded upwards; a horizontal stroke may be elongated.

For some letters the shape of the terminal stroke produces a discernable distinction between the regular and final forms. For instance, 𐼑 SAMEKH may be written as in final position; 𐼎 *nun* (ZAYIN) may be written as 𐼏 in final position (*zayin* retains the form 𐼈 when final). Final forms of other letters are not as distinctive. For example, 𐼀 ALEPH may occur as 𐼁; 𐼚 TAW may occur as 𐼛; For such letters, the terminal stroke may be interpreted as a stylistic flourish rather than a specific ‘final’ form. Moreover, there is no regularity in usage. For instance, regular and ‘final’ forms of SAMEKH are observed in final position in adjacent words.

At present, it is unclear if the differences between regular and word-final forms of letters should be distinguished at the character level, as is the case for the Hebrew and Nabatean encodings in Unicode.

5.2 Numbers

Distinctive signs are attested for ‘one’, ‘ten’, ‘one hundred’, and the fraction ‘one-half’. Other numbers are produced using repetitions and compounds of these signs, and using letters and words. The following characters are proposed for representing numbers:

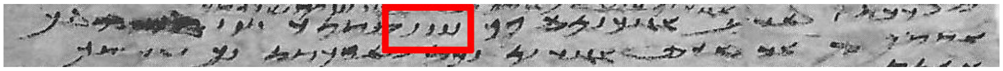
	Character name	Numeric value
𐼝	OLD SOGDIAN NUMBER ONE	1
𐼞	OLD SOGDIAN NUMBER TWO	2
𐼟	OLD SOGDIAN NUMBER THREE	3
𐼠	OLD SOGDIAN NUMBER FOUR	4
𐼢	OLD SOGDIAN NUMBER TEN	10
𐼣	OLD SOGDIAN NUMBER TWENTY	20
𐼤	OLD SOGDIAN NUMBER THIRTY	30
𐼥	OLD SOGDIAN NUMBER ONE HUNDRED	100
	OLD SOGDIAN NUMBER ONE THOUSAND	1000

The ordering of numbers follows the right-to-left directionality of the script. The expression of numbers is additive. Compound numbers of different decimal orders are written by placing the larger numbers first.

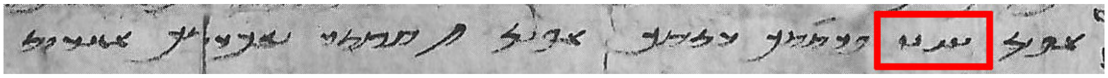
5.2.1 Primary units

The numbers 2–9 are produced using repetitions of the number 𐼝 ONE. These sequences are generally joined together at the left edge of ONE, eg. ‘3’ 𐼟. The numbers 4–9 are written using sequences of ONE arranged in groups of three or four separated by a space, eg. '6' 𐼟 𐼟; '7' 𐼠 𐼟. The representation of the primary units is modeled upon the encoding for Inscriptional Parthian, which has separate characters for the numbers 1–4, eg. 𐭘 U+10B58 INSCRIPTIONAL PARTHIAN NUMBER ONE … 𐭛 U+10B5B INSCRIPTIONAL PARTHIAN NUMBER FOUR. Similarly, the following characters are proposed for Old Sogdian: 𐼝 ONE, 𐼞 TWO, 𐼟 THREE, 𐼠 FOUR.

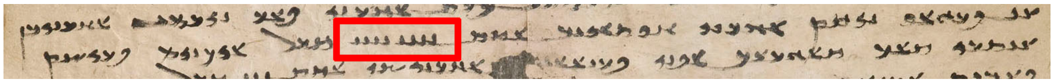
The number 4 written 𐼟 𐼝 (from 'Ancient Letter 5', line 26).



The number 5 written 𐼟 𐼞 (from 'Ancient Letter 5', line 10).



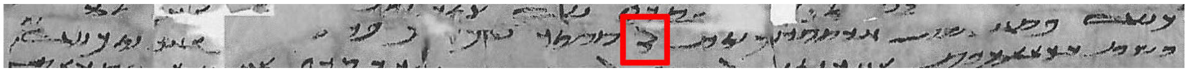
The number 8 written 𐼠 𐼠 (from 'Ancient Letter 2', line 26).



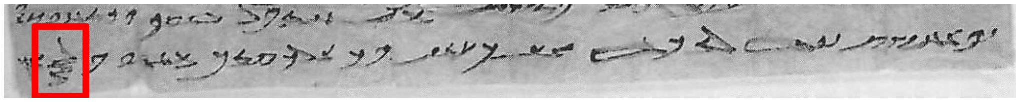
5.2.2 Tens

The number 𐼢 TEN resembles a vertically compressed form of the letter 𐼌 LAMEDH. Multiples of 10 are produced using horizontal sequences or vertical stacks of 𐼢. The only attested stacked forms are 20 𐼣 and 30 𐼤. These are encoded as the atomic characters 𐼣 TWENTY and 𐼤 THIRTY.

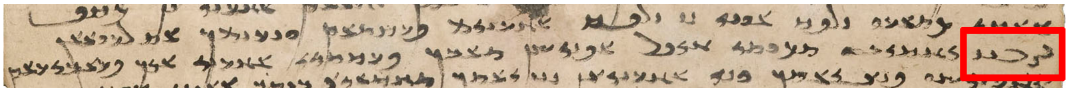
The number 20 𐼣 (from 'Ancient Letter 5', line 21):



The number 30 has two different representations in the available sources. In one instance it occurs as 𐼤 (from 'Ancient Letter 5', line 32):



In another source, 30 is represented as 𐼣𐼢, which is a compound of 𐼣 TWENTY and 𐼢 TEN. This form occurs in the number 𐼣𐼢𐼞 '32' (from 'Ancient Letter 2', line 62):

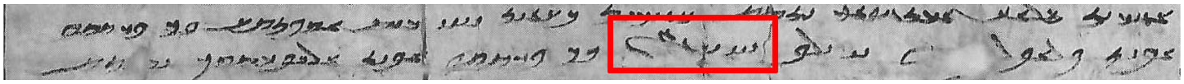


The proposed repertoire offers a means for representing both forms of 30.

5.2.3 Hundreds

The number 100 is written using 𐼥 ONE HUNDRED. The hundreds unit is represented using primary numbers followed by 𐼥 ONE HUNDRED, eg. 500 is expressed as 𐼟 𐼞𐼥 <𐼟 THREE, 𐼞 TWO, 𐼥 ONE HUNDRED>.

The number 500 written 𐼟 𐼞𐼥 (from ‘Ancient Letter 5’, line 9).

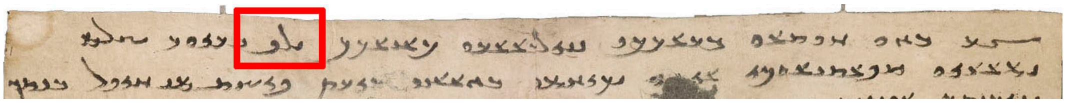


5.2.4 Thousands

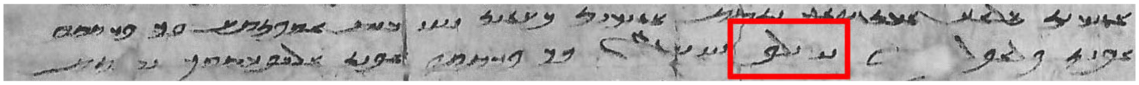
The number 1000 is written using 𐼝𐼌𐼔 ONE THOUSAND. This numerical sign is a ligature consisting of the number 𐼝 ONE joined to the sequence 𐼌𐼔 <𐼌 LAMEDH, 𐼔 PE>. The word 𐼌𐼔 represents the Aramaic heterogram *LP* “thousand” (= 𐡚 U+1085A IMPERIAL ARAMAIC NUMBER ONE THOUSAND). Although 𐼝𐼌𐼔 can be represented using the sequence <𐼝 ONE, 𐼌 LAMEDH, 𐼔 PE>, it is encoded as an atomic character on account of its conceptualization as a distinctive unit and the numerical value it possesses.

The thousands are represented using the appropriate repetitions of 𐼝 ONE followed by 𐼝𐼌𐼔 ONE THOUSAND, eg. 𐼞 𐼝𐼌𐼔 ‘2,000’. The presence of 𐼝 ONE in the glyph for ONE THOUSAND does not have a separate value.

The number 1,000 written 𐼝𐼌𐼔 (from ‘Ancient Letter 2’, line 1):

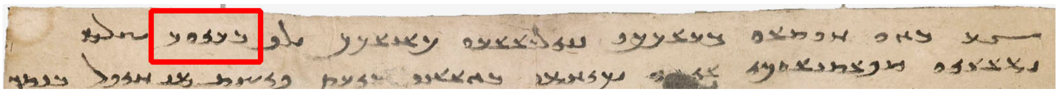


The number 2,000 written 𐼞 𐼝𐼌𐼔 (from ‘Ancient Letter 5’, line 9):



5.2.5 Ten thousand

There is no distinctive sign for the number 10,000. It is represented using the word 𐼂𐼘𐼊𐼇𐼀𐼘 *βryw’r* (from ‘Ancient Letter 2”, line 1):



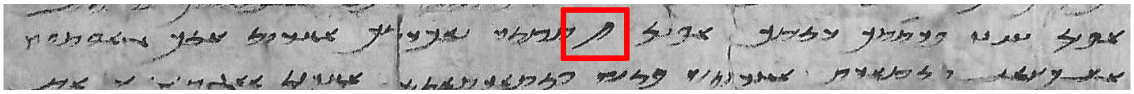
5.2.6 Fractions

The following fraction is attested:

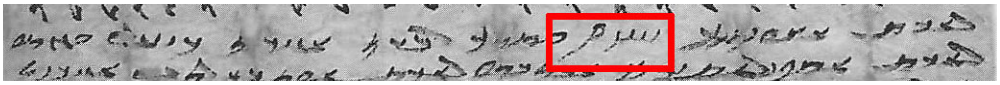
	Character name	Numeric value
𐼦	OLD SOGDIAN FRACTION ONE HALF	½

The 𐼦 FRACTION ONE HALF occurs in ‘Ancient Letter 5’. This character was identified as an alternate form of 100 by early scholars of Sogdian, but Grenet, et al (1998) have interpreted it as a sign for the fraction ½.

The fraction ½ written 𐼦 (from ‘Ancient Letter 5’, line 10):



The fraction 4½ written 𐼠𐼦 (from ‘Ancient Letter 5’, line 24):



5.3 Punctuation

There are no script-specific marks of punctuation. The available records show the usage of spaces for separating words. There are no punctuation signs for indicating the end of a sentence or larger sections of text.

5.4 Linebreaking

Linebreaks appear to occur after the end of a word. Words are not broken across lines and there are no continuation marks, such a hyphen. A sequence of numbers should be treated similarly, without being broken across lines.

5.5 Collation

The sort order for Old Sogdian is as follows:

𐼀 ALEPH < 𐼁 BETH < 𐼂 GIMEL < 𐼓 DALETH < 𐼃 HE ≪ 𐼃 ALTERNATE HE < 𐼄 WAW < 𐼅 ZAYIN < 𐼆 HETH < 𐼇 YODH < 𐼈 KAPH < 𐼉 LAMEDH < 𐼊 MEM < 𐼎 SAMEKH < 𐼐 ALTERNATE AYIN < 𐼑 PE < 𐼒 SADHE < 𐼓 RESH < 𐼔 SHIN < 𐼕 TAW

6 Character Data

Character Properties Properties in the format of `UnicodeData.txt`:

```
xx00;OLD SOGDIAN LETTER ALEPH;Lo;0;R;;;;;N;;;;;
xx01;OLD SOGDIAN LETTER BETH;Lo;0;R;;;;;N;;;;;
xx02;OLD SOGDIAN LETTER GIMEL;Lo;0;R;;;;;N;;;;;
xx03;OLD SOGDIAN LETTER DALETH;Lo;0;R;;;;;N;;;;;
xx04;OLD SOGDIAN LETTER HE;Lo;0;R;;;;;N;;;;;
xx05;OLD SOGDIAN LETTER WAW;Lo;0;R;;;;;N;;;;;
xx06;OLD SOGDIAN LETTER ZAYIN;Lo;0;R;;;;;N;;;;;
xx07;OLD SOGDIAN LETTER HETH;Lo;0;R;;;;;N;;;;;
xx08;<reserved>
```

```
xx09;OLD SOGDIAN LETTER YODH;Lo;0;R;;;;;N;;;;;
xx0A;OLD SOGDIAN LETTER KAPH;Lo;0;R;;;;;N;;;;;
xx0B;OLD SOGDIAN LETTER LAMEDH;Lo;0;R;;;;;N;;;;;
xx0C;OLD SOGDIAN LETTER MEM;Lo;0;R;;;;;N;;;;;
xx0D;<reserved>
xx0E;OLD SOGDIAN LETTER SAMEKH;Lo;0;R;;;;;N;;;;;
xx0F;<reserved>
xx10;OLD SOGDIAN LETTER PE;Lo;0;R;;;;;N;;;;;
xx11;OLD SOGDIAN LETTER SADHE;Lo;0;R;;;;;N;;;;;
xx12;<reserved>
xx13;OLD SOGDIAN LETTER RESH;Lo;0;R;;;;;N;;;;;
xx14;OLD SOGDIAN LETTER SHIN;Lo;0;R;;;;;N;;;;;
xx15;OLD SOGDIAN LETTER TAW;Lo;0;R;;;;;N;;;;;
xx16;OLD SOGDIAN LETTER ALTERNATE HE;Lo;0;R;;;;;N;;;;;
xx17;OLD SOGDIAN LETTER ALTERNATE AYIN;Lo;0;R;;;;;N;;;;;
xx18;OLD SOGDIAN NUMBER ONE;No;0;R;;;;1;N;;;;;
xx19;OLD SOGDIAN NUMBER TWO;No;0;R;;;;2;N;;;;;
xx1A;OLD SOGDIAN NUMBER THREE;No;0;R;;;;3;N;;;;;
xx1B;OLD SOGDIAN NUMBER FOUR;No;0;R;;;;4;N;;;;;
xx1C;OLD SOGDIAN NUMBER TEN;No;0;R;;;;10;N;;;;;
xx1D;OLD SOGDIAN NUMBER TWENTY;No;0;R;;;;20;N;;;;;
xx1E;OLD SOGDIAN NUMBER THIRTY;No;0;R;;;;30;N;;;;;
xx1F;OLD SOGDIAN NUMBER ONE HUNDRED;No;0;R;;;;100;N;;;;;
xx20;OLD SOGDIAN NUMBER ONE THOUSAND;No;0;R;;;;1000;N;;;;;
xx21;OLD SOGDIAN FRACTION ONE HALF;;No;0;R;;;;1/2;N;;;;;
```

Linebreaking Linebreaking properties in the format of `LineBreak.txt`:

```
xx00..xx07;AL      # Lo      [8] OLD SOGDIAN LETTER ALEPH..OLD SOGDIAN LETTER HETH
xx09..xx0C;AL      # Lo      [4] OLD SOGDIAN LETTER YODH..OLD SOGDIAN LETTER MEM
xx0E;AL            # Lo          OLD SOGDIAN LETTER SAMEKH
xx10..xx11;AL      # Lo      [2] OLD SOGDIAN LETTER PE..OLD SOGDIAN LETTER SADHE
xx13..xx15;AL      # Lo      [3] OLD SOGDIAN LETTER RESH..OLD SOGDIAN LETTER TAW
xx16;AL            # Lo          OLD SOGDIAN LETTER ALTERNATE HE
xx17;AL            # Lo          OLD SOGDIAN LETTER ALTERNATE AYIN
xx18..xx20;AL      # No      [9] OLD SOGDIAN NUMBER ONE..OLD SOGDIAN NUMBER ONE THOUSAND
xx21;AL            # No          OLD SOGDIAN FRACTION ONE HALF
```

7 Questions

1. *Glyphic representations* Are the glyphic representations for the proposed characters acceptable?

2. *Any abecedary?* Is there any primary source containing the letter repertoire of Old Sogdian, similar to the Panjikant ostracon for ‘cursive’ Sogdian?

3. *Merged letters* Is there any evidence to suggest that scribes perceived a distinction between letters such as *zayin* and *nun*, or *ayin* and *resh* when writing Old Sogdian? Did they knowingly use the same glyph for multiple letters?

4. *Missing characters?* Are there any characters in the ‘Ancient Letters’ that are not identified here? Alternate forms of other letters?

5. *he* Are there rules that specify the usage of 𐼅 alternate he in place of 𐼄 he? Or, are the two forms interchangable?

6. *teth* Is an Old Sogdian analogue for Aramaic *teth* attested in heterograms?

7. *qoph* Is an Old Sogdian analogue for Aramaic *qoph* attested in heterograms? Is there any relation between the 𐼦 FRACTION ONE HALF and Aramaic 𐡒 *qoph*?

8. *Final forms* Is there is list of letters that have distinctive final forms? Secondly, should final forms be represented using regular letters, or should they be encoded separately, as has been done in Unicode for Hebrew and Nabatean?

8 References

Everson, Michael. 2001. “Roadmapping early Semitic scripts” (L2/01-024). `http://www.unicode.org/L2/L2001/01024-n2311.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, edited by Osmund Bopearachchi, Carol Altman Bromberg, and Frantz Grenet, pp. 91–104.

Livshits, V. A. 2015. *Sogdian epigraphy of Central Asia and Semirech’e*. Corpus Inscriptonum Iranicarum, pt. II (Inscriptions of the Seleucid and Parthian periods of Eastern Iran and Central Asia), v. III (Sogdian), no. II. Translated by Tom Stableford, edited by Nicholas Sims-Williams. London: Published on behalf of Corpus Inscriptionum Iranicarum by School of Oriental and African Studies.

———. 2015. “A Sogdian alphabet from Penjikent”, in Livshits (2015), pp. 227–232.

Sims-Williams, Nicholas. 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; and Franz 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.

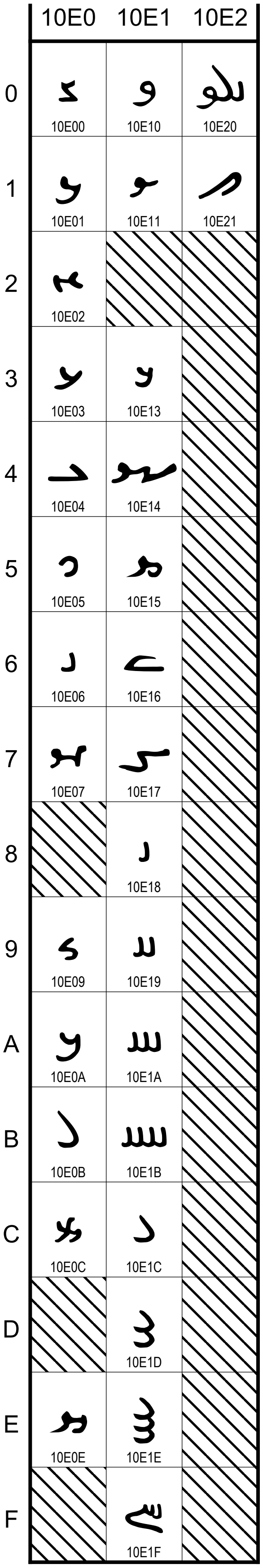
———. 2006. “Iran. VI. Iranian Languages and Scripts. (3) Writing Systems.” *Encyclopædia Iranica*, vol. XIII, fasc. 4, pp. 366–370. `http://www.iranicaonline.org/articles/iran-vi3-writing-systems`

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

9 Acknowledgments

I express my gratitude to Nicholas Sis-Williams (SOAS, University of London) for providing detailed comments on an earlier version of this proposal. I am also thankful to Yutaka Yoshida (University of Kyoto) for reviewing this proposal and for providing feedback. I am grateful to Roozbeh Pournader (Google, San Francisco) for discussing the encodings of Aramaic-based Iranian scripts in Unicode and also for his feedback on an earlier draft of this proposal.

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


10E0
10E1
10E2
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F
10E00
10E10
10E20
10E01
10E11
10E21
10E02
10E03
10E13
10E04
10E14
10E05
10E15
10E06
10E16
10E07
10E17
10E18
10E09
10E19
10E0A
10E1A
10E0B
10E1B
10E0C
10E1C
10E1D
10E0E
10E1E
10E1F

This block unifies the scripts used in the Ancient Letters and Upper Indus inscriptions. Spaces are reserved for Aramaic analogues, which are unattested at present.

Letters

10E00 OLD SOGDIAN LETTER ALEPH
10E01 OLD SOGDIAN LETTER BETH
10E02 OLD SOGDIAN LETTER GIMEL
10E03 OLD SOGDIAN LETTER DALETH
• used only in Aramaic heterograms
10E04 OLD SOGDIAN LETTER HE
• only used in word-final position
10E05 OLD SOGDIAN LETTER WAW
10E06 OLD SOGDIAN LETTER ZAYIN
= zayin, nun
10E07 OLD SOGDIAN LETTER HETH
10E08 ▨ <reserved>
10E09 OLD SOGDIAN LETTER YODH
10E0A OLD SOGDIAN LETTER KAPH
10E0B OLD SOGDIAN LETTER LAMEDH
10E0C OLD SOGDIAN LETTER MEM
10E0D ▨ <reserved>
10E0E OLD SOGDIAN LETTER SAMEKH
10E0F ▨ <reserved>
10E10 OLD SOGDIAN LETTER PE
10E11 OLD SOGDIAN LETTER SADHE
10E12 ▨ <reserved>
10E13 OLD SOGDIAN LETTER RESH
= daleth, ayin, resh
10E14 OLD SOGDIAN LETTER SHIN
10E15 OLD SOGDIAN LETTER TAW

Alternate letters

10E16 OLD SOGDIAN LETTER ALTERNATE HE
• used only in word-final position
10E17 OLD SOGDIAN LETTER ALTERNATE AYIN
• used only in Aramaic heterograms

Numbers

10E18 OLD SOGDIAN NUMBER ONE
10E19 OLD SOGDIAN NUMBER TWO
10E1A OLD SOGDIAN NUMBER THREE
10E1B OLD SOGDIAN NUMBER FOUR
10E1C OLD SOGDIAN NUMBER TEN
10E1D OLD SOGDIAN NUMBER TWENTY
10E1E OLD SOGDIAN NUMBER THIRTY
10E1F OLD SOGDIAN NUMBER ONE HUNDRED
10E20 OLD SOGDIAN NUMBER ONE THOUSAND

Fraction

10E21 OLD SOGDIAN FRACTION ONE HALF

	Old Sogdian	Inscriptional Pahlavi	Inscriptional Parthian	Imperial Aramaic
ALEPH	𐼀	𐭠	𐭀	𐡀
BETH	𐼂	𐭡	𐭁	𐡁
GIMEL	𐼄	𐭢	𐭂	𐡂
DALETH	𐼘	𐭣	𐭃	𐡃
HE	𐼅	𐭤	𐭄	𐡄
WAW	𐼇	𐭥	𐭅	𐡅
ZAYIN	𐼈	𐭦	𐭆	𐡆
HETH	𐼉	𐭧	𐭇	𐡇
TETH	—	𐭨	𐭈	𐡈
YODH	𐼊	𐭩	𐭉	𐡉
KAPH	𐼋	𐭪	𐭊	𐡊
LAMEDH	𐼌	𐭫	𐭋	𐡋
MEM	𐼍	𐭬	𐭌	𐡌
NUN	(𐼈)	𐭭	𐭍	𐡍
SAMEKH	𐼑	𐭮	𐭎	𐡎
AYIN	(𐼘)	(𐭥)	𐭏	𐡏
PE	𐼔	𐭯	𐭐	𐡐
SADHE	𐼕	𐭰	𐭑	𐡑
QOPH	—	(𐭬)	𐭒	𐡒
RESH	𐼘	(𐭥)	𐭓	𐡓
SHIN	𐼙	𐭱	𐭔	𐡔
TAW	𐼚	𐭲	𐭕	𐡕

Table 5: Comparison of Old Sogdian and other Iranian scripts derived from Aramaic. Parenthesis indicate that the respective letter is not encoded as a distinct character and that it is represented in the encoded repertoire using another character. In the encoding for Inscriptional Pahlavi, *ayin* and *resh* are represented using WAW, and *qoph* is represented using MEM. Dashes indicate the absence of a given letter in the respective script.

	Old Sogdian	Inscriptional Pahlavi	Inscriptional Parthian	Imperial Aramaic
ONE	𐼝	𐭸	𐭘	𐡘
TWO	𐼞	𐭹	𐭙	𐡙
THREE	𐼟	𐭺	𐭚	𐡚
FOUR	𐼠	𐭻	𐭛	—
TEN	𐼢	𐭼	𐭜	𐡛
TWENTY	𐼣	𐭽	𐭝	𐡜
THIRTY	𐼤	—	—	—
ONE HUNDRED	𐼥	𐭾	𐭞	𐡝
ONE THOUSAND		𐭿	𐭟	𐡞
TEN THOUSAND	—	—	—	𐡟
ONE HALF	𐼦	—	—	—

Table 6: Comparison of numerical characters proposed for Old Sogdian and those encoded in Unicode blocks for Aramaic and Iranian scripts derived from it. Note: There is no distinctive numerical sign for 10,000 in Old Sogdian, instead this number is represented using the word 𐼂𐼘𐼊𐼇𐼘 *βrewar*.

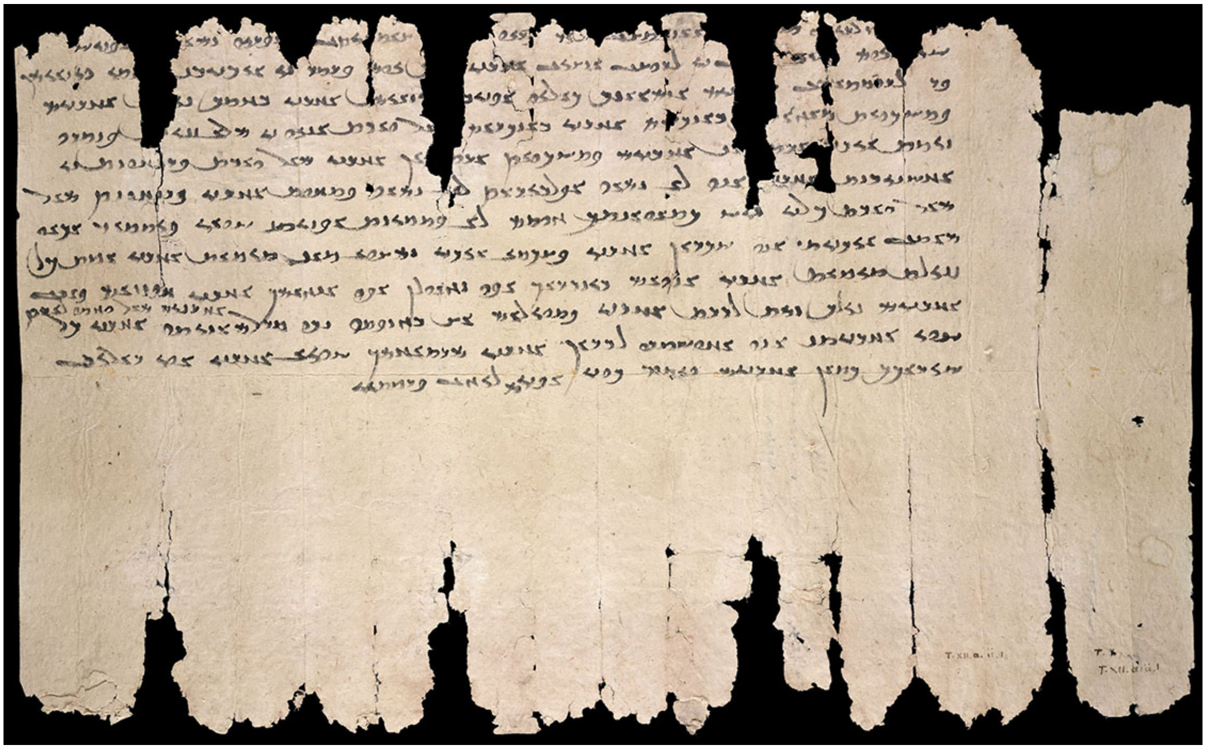


Figure 1: Excerpt of ‘Ancient Letter 1’ (British Library, International Dunhuang Project: Or. 8212/92.1 recto 1). “From her daughter, the free-woman Miwnay, to her d[ear] mother [Chatis].” (translation by Sims-Williams in Waugh 2004).

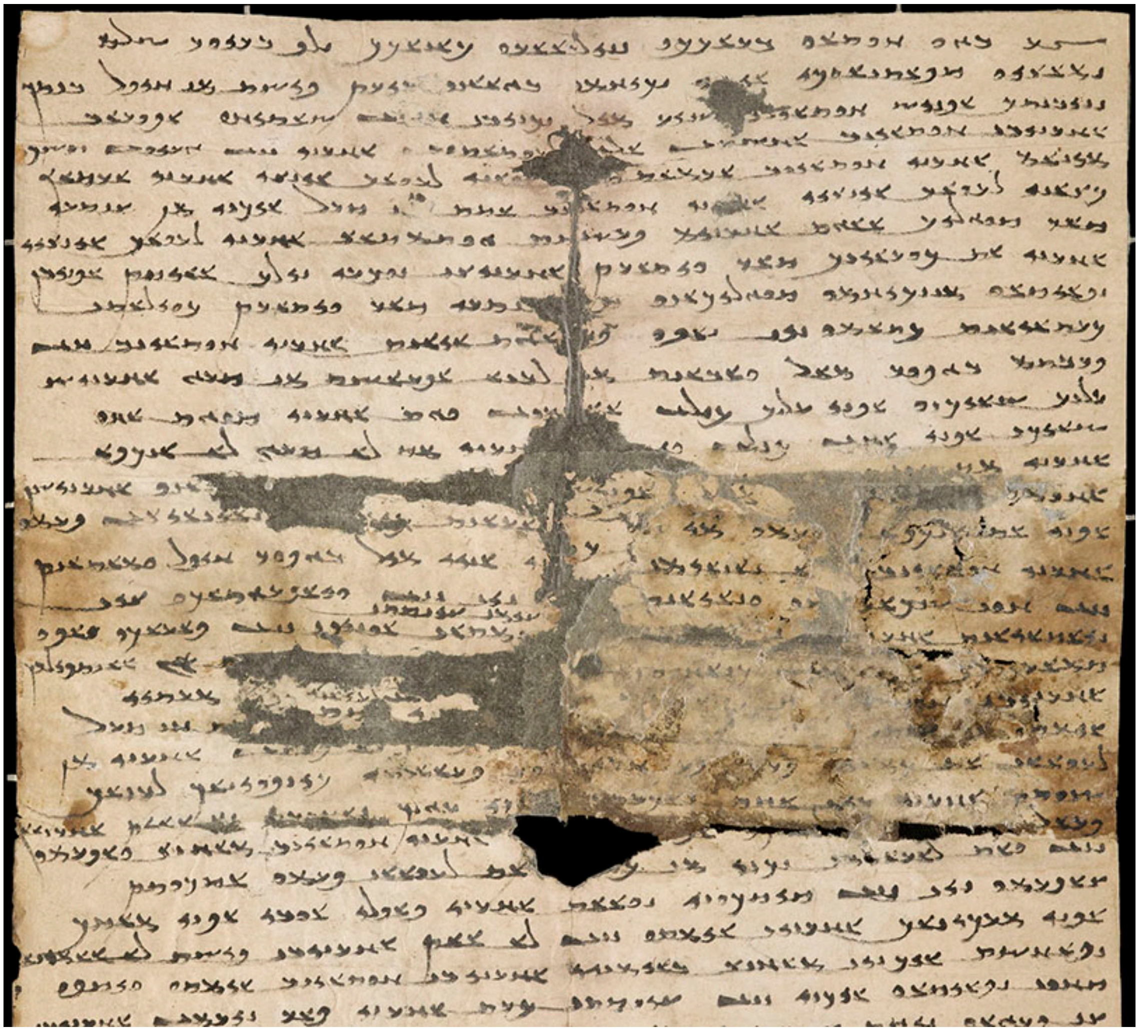


Figure 2: Excerpt of ‘Ancient Letter 2’ (British Library, International Dunhuang Project: Or. 8212/95 side a). “To the noble lord Varzakk (son of) Nanai-thvar (of the family) Kanakk. Sent [by] his servant Nanai-vandak.” (translation by Sims-Williams in Waugh 2004).

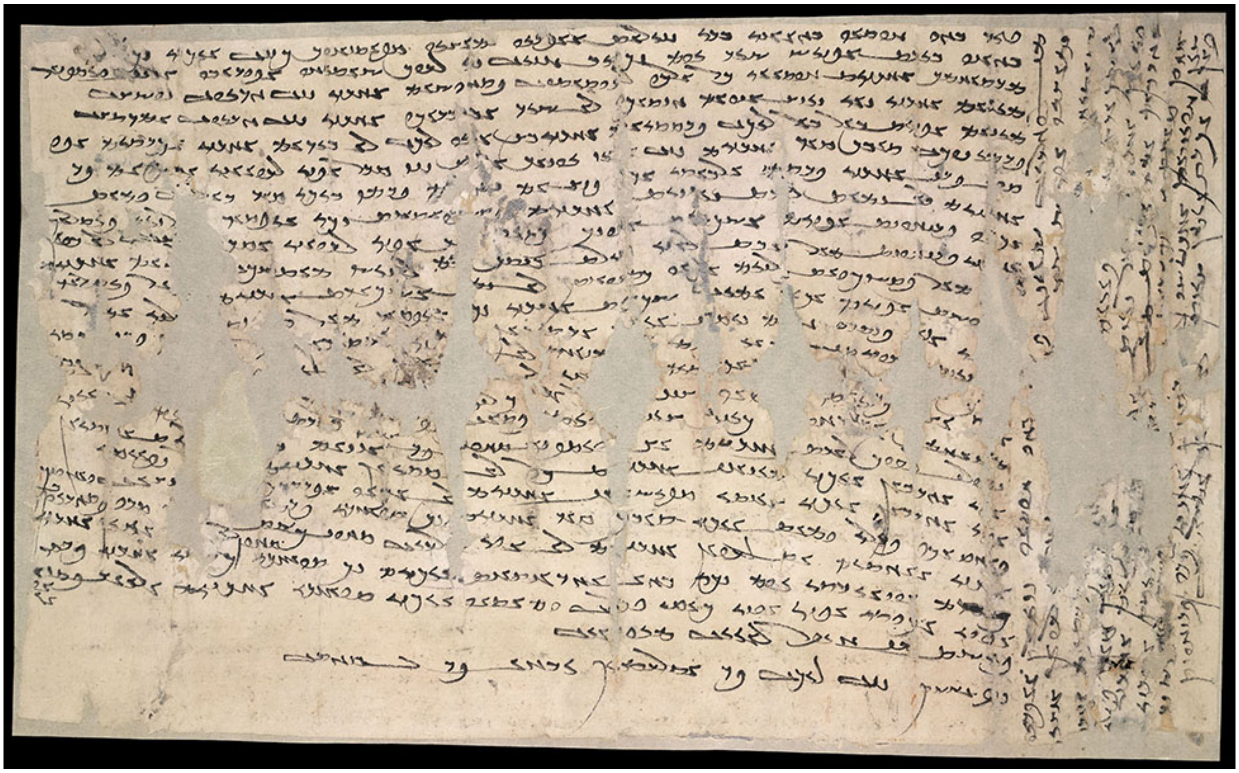


Figure 3: Excerpt of ‘Ancient Letter 3’ (British Library, International Dunhuang Project: Or. 8212/98 recto 1). “From (his) daughter Shayn to the noble lord Nanai-dhat.” (translation by Sims-Williams in Waugh 2004).

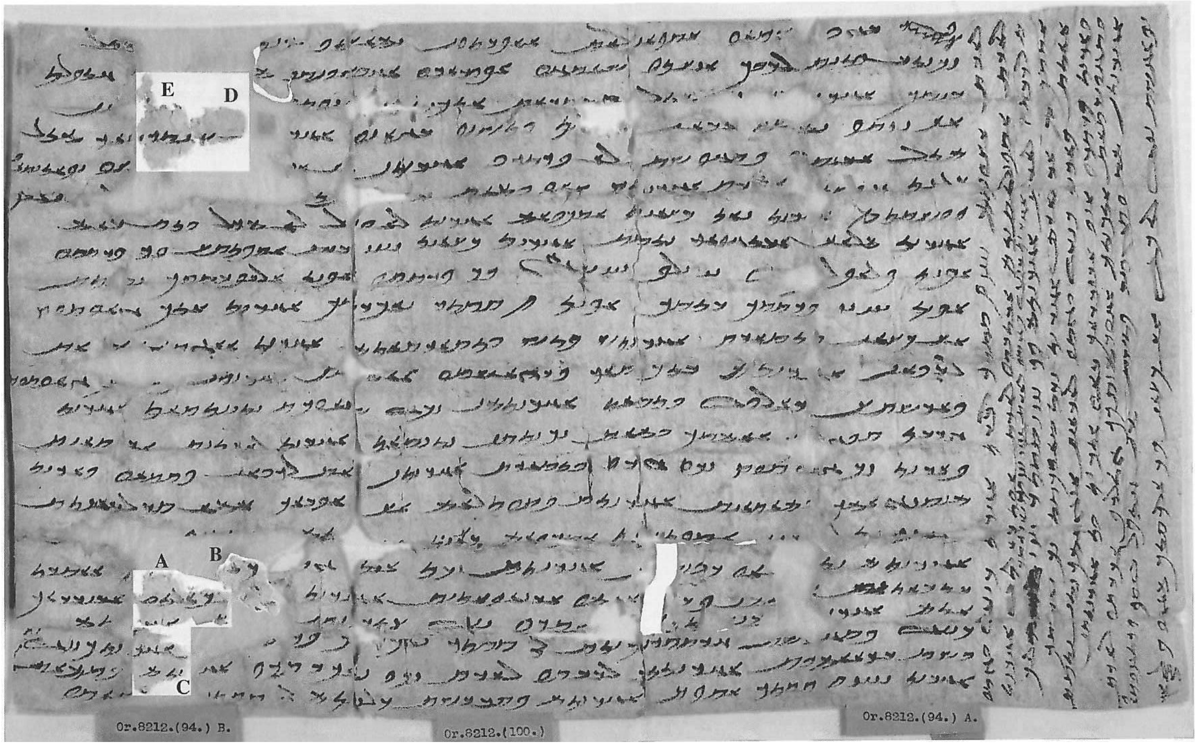

E
D
A
B
C
Or.8212.(94.) B.
Or.8212.(100.)
Or.8212.(94.) A.

Figure 4: Excerpt of ‘Ancient Letter 5’ (from Grenet et al. 1998: 94). “To the noble lord, the chief merchant Aspandhāt. [Sent] by your servant [Frī-khwatāw].”

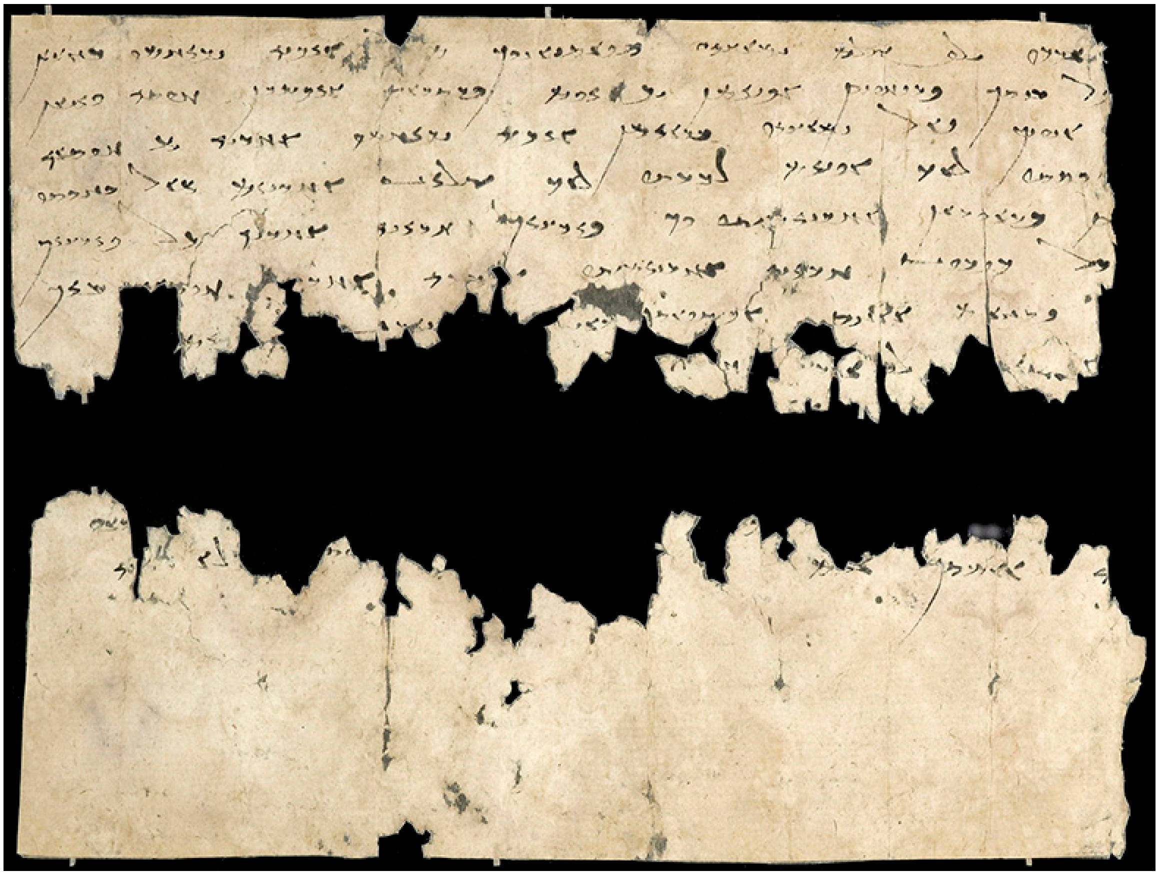


Figure 5: Excerpt of Ancient Letter 6 (British Library, International Dunhuang Project: Or. 8212/97).

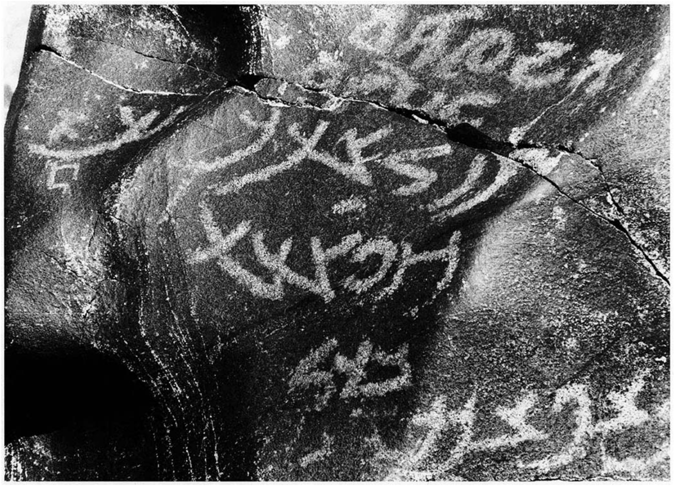


Figure 6: Sogdian rock inscription from Shatial (from Sims-Williams 1989: plate 10b) The inscription reads 𐼎𐼎𐼊𐼀𐼋𐼋 𐼈𐼋 *nny'kk ZK* (top line), 𐼑𐼇𐼀𐼂𐼘 *sw'βr* (middle), 𐼂𐼘𐼊 *BRY* (bottom). Latin transcription from *ibid*: 14. The inscription in the bottom right-hand corner is shown in detail in figure 7.

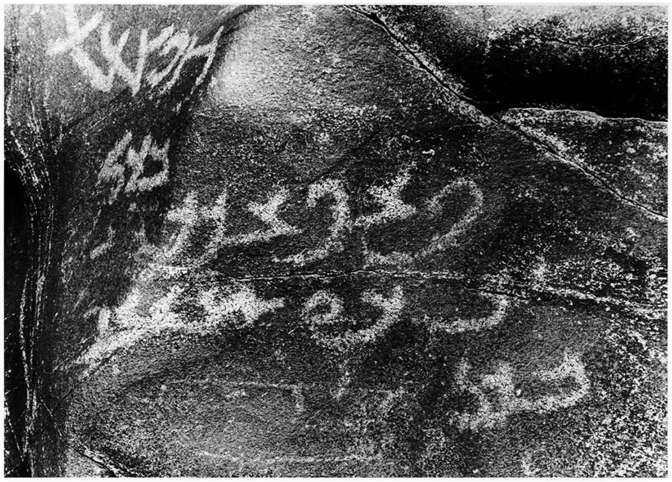
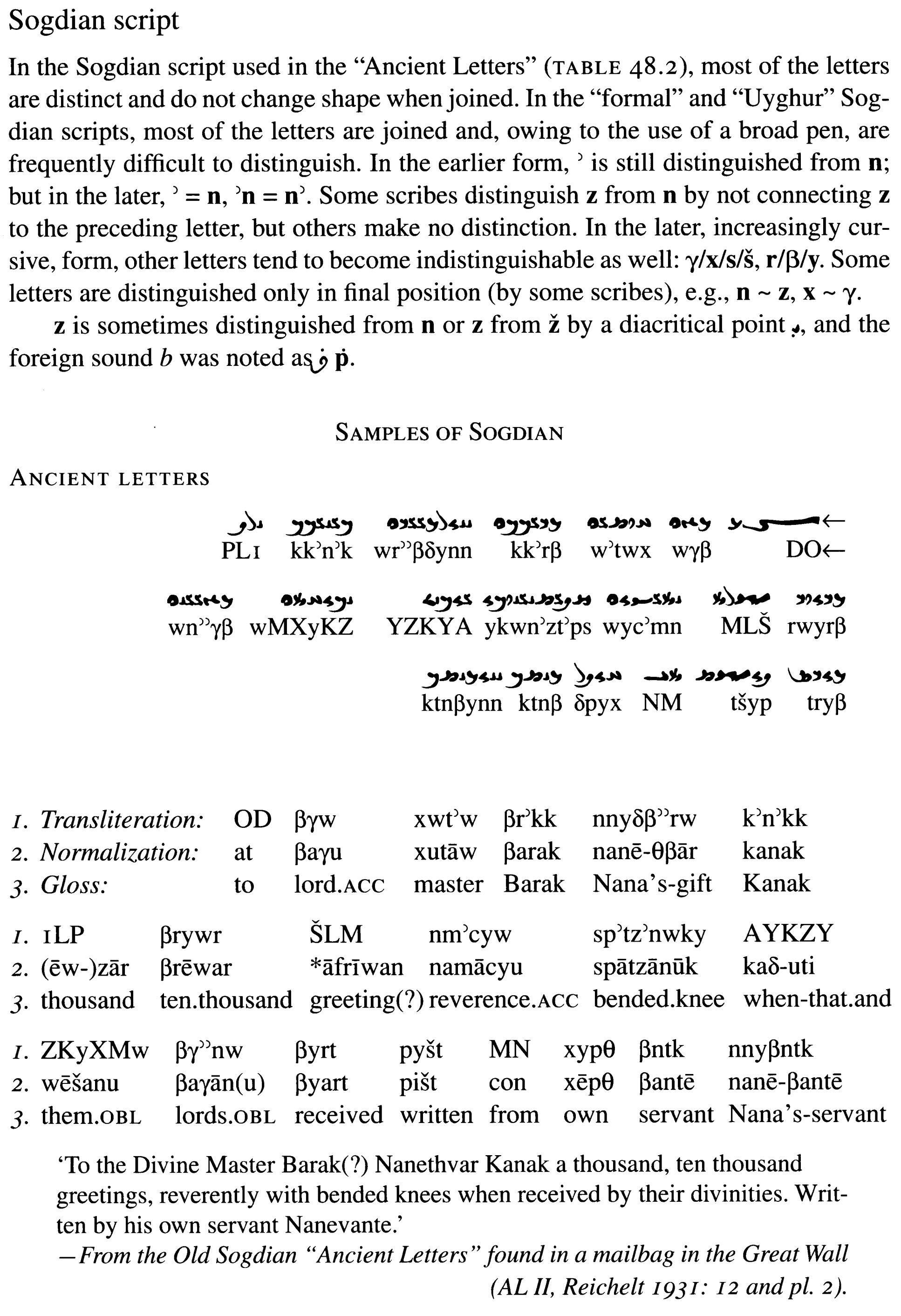


Figure 7: Sogdian rock inscription from Shatial (from Sims-Williams 1989: plate 10a). The central inscription reads 𐼔𐼀𐼔𐼀𐼋𐼋 *p'p'kk* (top line), 𐼈𐼋 𐼋𐼇𐼙𐼀𐼀𐼎 *ZK kwš''n* (middle), 𐼂𐼘𐼊 *BRY* (bottom). Latin transcription from *ibid*: 14. The inscription in the top left-hand corner is shown in detail in figure 6.



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

PLı kkʾnʾk wrʾʾβδynn kkʾrβ wʾtwx wγβ DO←

wnʾʾγβ wMXyKZ YZKYA ykwnʾztʾps wycʾmn MLŠ rwyrβ

ktnβynn ktnβ δpyx NM tšyp tryβ

1. Transliteration:	OD	βyw	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. ıLP	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

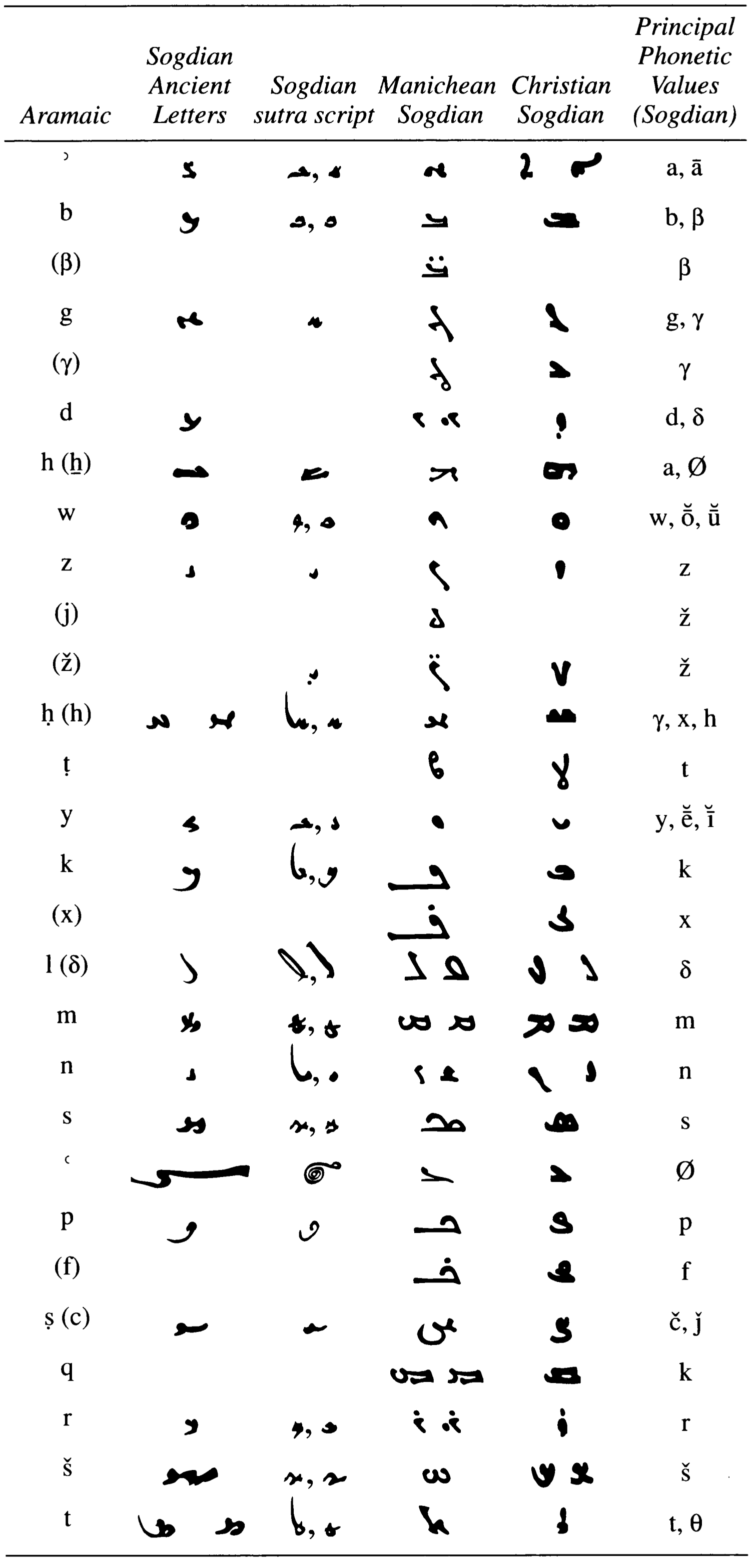
1. ZKyXMw	βγʾʾnw	βyrt	pyšt	MN	xypθ	βntk	nnyβntk
2. wēšanu	βaγān(u)	βyart	pišt	con	xēpθ	βantē	nanē-βantē
3. them.OBL	lords.OBL	received	written	from	own	servant	Nana's-servant

'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).

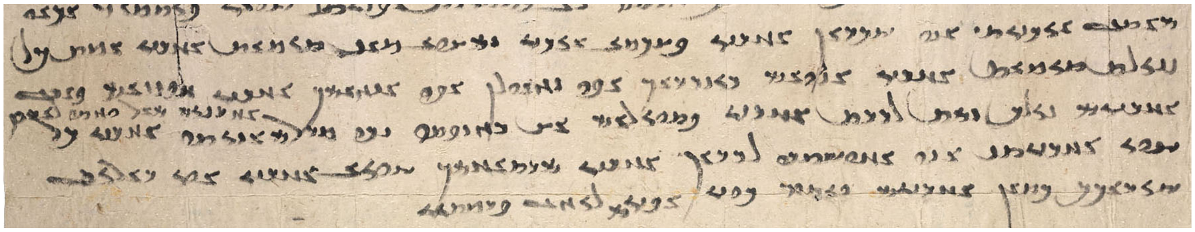
Figure 8: Description and specimen of the script in the 'Ancient Letters' (from Skjærvø 1996: 529).

TABLE 48.2: *Main East Iranian Scripts Developed from Aramaic*

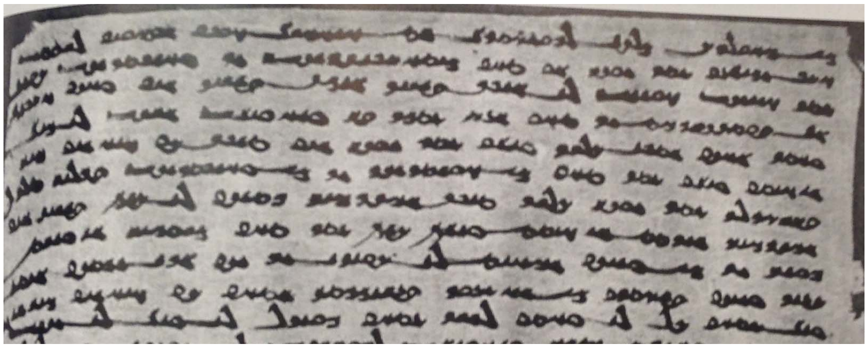


Aramaic	*Sogdian Ancient Letters*	*Sogdian sutra script*	*Manichean Sogdian*	*Christian Sogdian*	*Principal Phonetic Values (Sogdian)*
ʾ					a, ā
b					b, β
(β)					β
g					g, γ
(γ)					γ
d					d, δ
h (h̲)					a, Ø
w					w, ŏ, ŭ
z					z
(j)					ž
(ž)					ž
ḥ (h)					γ, x, h
ṭ					t
y					y, ĕ, ĭ
k					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

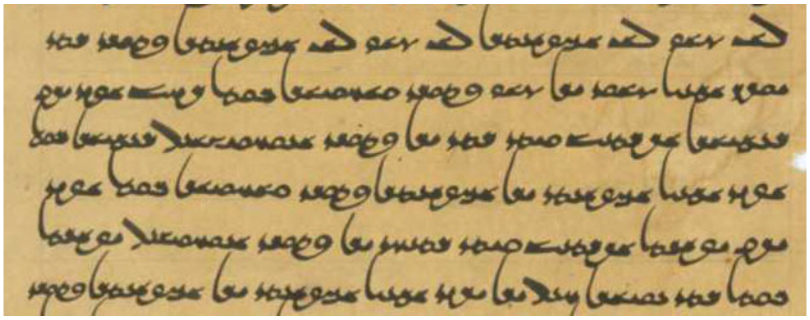
Figure 9: Table showing various scripts for writing Sogdian (from Skjærvø 1996: 519).



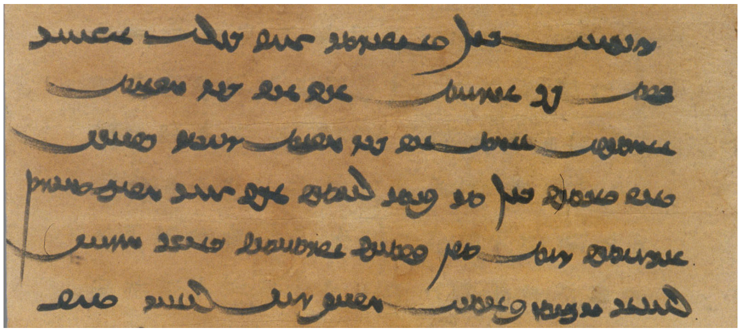
Sogdian script of the ‘Ancient Letters’ (excerpt from BL IDP: Or. 8212/92.1 recto 1)



‘Cursive’ Sogdian script of the Mount Mug documents (excerpt of B-18, from Livshits 2015: 99)



Sogdian ‘sūtra’ script (excerpt from Digitales Turfan-Archiv: So 14851, recto 1)



Sogdian ‘Uyghur’ script (excerpt from Digitales Turfan-Archiv: Ch/So 14744, recto 1)

Figure 10: Comparison of Old Sogdian (top) and later Sogdian scripts.