L2/16-371R2
WG2 N4815

2017-01-25

Revised proposal to encode the Sogdian script in Unicode

Anshuman Pandey
pandey@umich.edu

January 25, 2017

1 Introduction

This is a proposal to encode the Sogdian script in Unicode. It supersedes the following document:

- L2/16-158 “Proposal to encode Sogdian in Unicode”

A proposal summary form is attached. In addition to substantial modifications, it addresses comments regarding L2/15-158 and previous drafts of the present proposal that have been made in:

- L2/16-037 “Recommendations to UTC #146 January 2016 on Script Proposals”
- L2/16-216 “Recommendations to UTC #148 August 2016 on Script Proposals”
- L2/17-037 “Recommendations to UTC #150 January 2017 on Script Proposals”

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

- L2/16-312R “Proposal to encode the Old 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 Sogdian script was used primarily for representing Sogdian (ISO 639: sog), an ancient Eastern Iranian language, but also for recording texts in Chinese, Sanskrit, and Uyghur. The script was used primarily for manuscripts written on paper (fig. 60–79), but also for inscriptions on coins (fig. 80), stone (fig. 81–82), pottery (fig. 83), and other media. The script is derived from ‘Old Sogdian’ (see L2/16-312R) and is related to the Syriac and Manichaean scripts (see fig. 1). It is the ancestor of the Uyghur script (see § 2.1) and, in turn, of the Mongolian writing system.

The proposed Unicode encoding for ‘Sogdian’ encompasses a group of related script styles that possess the same character repertoire and structural features, as described below:

- *Styles* The script may be differentiated into two major styles: ‘formal’ and ‘cursive’. These developed separately from Old Sogdian, the script used in the Ancient Letters. The ‘cursive’ style likely emerged first. The ‘formal’ script used in Buddhist manuscripts is often referred to as the ‘sūtra’ script. It is difficult to specify a clear division between the ‘formal’ and ‘cursive’ varieties. Rather, it is appropriate to consider them as belonging to a script continuum, with the most clear, formal styles at one terminal and the simplified, cursive styles at the other. Manuscript folios illustrating these styles are shown in fig. 60–76.

- *Repertoire* The alphabet is attested on an ostracon found at Panjakent, modern Tajikistan, which has been dated to the 7–8th century. The inscription shows 22 letters that correspond to the full Aramaic repertoire, as well as a 23rd, which is a redundant *lamedh* placed at the end of the order (see fig. 3). The glyphs used for representing *daleth*, *ayin*, *qoph*, *teth* are non-letter signs (see fig. 3 for description). The inscription suggests that while only 19 letters were used conventionally in the Sogdian script of this era, some scribes were aware of the original Aramaic template for the alphabet. This repertoire of 19 letters aligns with the Old Sogdian alphabet, although with differences in glyphic representations (see table 1). The repertoire is also attested in a manuscript fragment in the Otani collection (see fig. 4). However, the order of letters in this fragment differs from the Aramaic template. Apart from their enumeration in abecedaries, the letters *daleth*, *qoph*, *teth* are not used in the Sogdian script. The letter *ayin* took on a shape nearly identical to *resh*, and a special form of *ayin* emerged for writing an Aramaic heterogram. New letters were introduced, such as *feth* [f] and *lesh* or ‘hooked *resh*’ [l]. Diacritic marks were introduced for disambiguation and for transcription. Numerical signs similar to those used in Old Sogdian are attested. Various marks of punctuation were also used.

- *Structure* The script is a conjoining *abjad*, similar to Arabic. Letters connect and change shape based upon their position within a word. In later styles, some letters (ie. *zayin*, *heth*, *yodh*) remain unconnected from a following letter in order to distinguish letters with similar shapes (ie. *nun*, *gimel*, *beth*). Words are separated using spaces. The conjoining behavior of Sogdian contrasts with the non-joining Old Sogdian. The joining behavior is an evolution of the natural writing style found in the Ancient Letters, where strokes between adjacent letters are joined on account of rapid writing.

		Old Sogdian	Sogdian
swγδyk	‘Sogdian’	𐼌𐼄𐼂𐼉𐼇𐼈	𐼼𐼴𐼲𐼹𐼷𐼸
smʾrknδc	‘of Samarkand’	𐼌𐼊𐼀𐼐𐼈𐼋𐼉𐼏	𐼼𐼺𐼰𐽀𐼸𐼻𐼹𐼿

- *Directionality* The script is written both horizontally and vertically. In horizontal mode, the writing direction is right to left and lines proceed from top to bottom. When vertical, glyphs are rotated 90° counter-clockwise and are written from top to bottom in lines that advance from the left edge of the writing surface towards the right.

The varieties along the continuum between the ‘formal’ and ‘cursive’ styles are to be considered typologically identical on the basis of their repertoires, and 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 for each script variety.

2.1 Considerations for the Uyghur script

Sogdian is the ancestor of the script known as ‘Uyghur’. This ‘Uyghur’ script is also referred to as ‘Old Uyghur’, a term that is also used for the ‘Old Turkic’ script (U+10C00 .. U+10C4F). The ‘Uyghur’ and ‘Old Turkic’ scripts are separate writing systems.

The ‘Uyghur’ script is believed to have developed from the ‘cursive’ style during the 8th–9th century (Kara 1996: 539). To be sure, there is much similarity between ‘late cursive’ Sogdian and the ‘early’ Uyghur script, such that they may be considered to be the same style. However, there emerged a ‘normative’ representation of Uyghur with graphical characteristics and scribal peculiarities that differ from those of Sogdian. This script may be considered ‘formal’ Uyghur and a distinctive script in its own right (see fig. 84).

In terms of character encoding, it may be possible to unify Uyghur with the proposed Sogdian block. It may also be possible to consider it as a separate script, with its own stylistic variants. There is, however, an outstanding request to encode Uyghur separately in Unicode, which must be evaluated. In “Proposal to Encode the Uyghur Script in ISO/IEC 10646” (L2/13-071), Omarjan Osman illustrates digits, diacritics, and other characters that appear to be specific to the Uyghur script. Osman also describes requirements for managing different styles and directional orientations of the script. Determining the most appropriate method of handling Uyghur in Unicode requires additional research, especially if there are requirements for representing it in plain text. However, such an effort is out of scope for the present project.

3 Character Repertoire

The proposed repertoire for Sogdian contains 42 characters: 21 letters, 1 phonogram, 11 diacritic signs, 4 numbers, and 5 punctuation signs. Names for letters correspond to those of the proposed ‘Old Sogdian’ block, which are derived from character names of ‘Imperial Aramaic’. In general, representative glyphs are based upon the ‘formal’ variety. An attempt has been made to adapt glyphs of other styles to the ‘formal’ style for sake of normalization. However, it is not an easy task to normalize diverse handwritten styles used over the course of nine centuries.

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. 𐼰 is *aleph* and SOGDIAN LETTER ALEPH. For sake of brevity, the descriptor ‘SOGDIAN’ is dropped when refering to Sogdian characters, eg. SOGDIAN LETTER ALEPH is referred to as ALEPH. Characters of other scripts are designated by their full Unicode names.

Latin transliteration of 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

Letters included in the proposed repertoire are shown below in their isolated and positional forms (see fig. 5–25 for attestations):

Nominal	Character name	Latin	Final	Medial	Initial	Joining
𐼰	SOGDIAN LETTER ALEPH	ʾ	‍𐼰	‍𐼰‍	𐼰‍	dual
𐼱	SOGDIAN LETTER BETH	β ; B	‍𐼱	‍𐼱‍	𐼱‍	dual
𐼲	SOGDIAN LETTER GIMEL	γ ; G	‍𐼲	‍𐼲‍	𐼲‍	dual*
𐼳	SOGDIAN LETTER HE	h	‍𐼳	—	—	right
𐼴	SOGDIAN LETTER WAW	w	‍𐼴	‍𐼴‍	𐼴‍	dual
𐼵	SOGDIAN LETTER ZAYIN	z	‍𐼵	‍𐼵‍	𐼵‍	dual*
𐼶	SOGDIAN LETTER HETH	x ; Ḥ	‍𐼶	‍𐼶‍	𐼶‍	dual
𐼷	SOGDIAN LETTER YODH	y	‍𐼷	‍𐼷‍	𐼷‍	dual
𐼸	SOGDIAN LETTER KAPH	k	‍𐼸	‍𐼸‍	𐼸‍	dual
𐼹	SOGDIAN LETTER LAMEDH	δ ; L	‍𐼹	‍𐼹‍	𐼹‍	dual
𐼺	SOGDIAN LETTER MEM	m	‍𐼺	‍𐼺‍	𐼺‍	dual
𐼻	SOGDIAN LETTER NUN	n	‍𐼻	‍𐼻‍	𐼻‍	dual
𐼼	SOGDIAN LETTER SAMEKH	s	‍𐼼	‍𐼼‍	𐼼‍	dual
𐼽	SOGDIAN LETTER AYIN	ʿ	‍𐼽	‍𐼽‍	𐼽‍	*
𐼾	SOGDIAN LETTER PE	p	‍𐼾	‍𐼾‍	𐼾‍	dual
𐼿	SOGDIAN LETTER SADHE	c ; Ṣ	‍𐼿	‍𐼿‍	𐼿‍	dual
𐽀	SOGDIAN LETTER RESH-AYIN	r , ʿ	‍𐽀	‍𐽀‍	𐽀‍	dual
𐽁	SOGDIAN LETTER SHIN	š	‍𐽁	‍𐽁‍	𐽁‍	dual
𐽂	SOGDIAN LETTER TAW	t	‍𐽂	‍𐽂‍	𐽂‍	dual

𐽃	SOGDIAN LETTER FETH	f	‍𐽃	‍𐽃‍	𐽃‍	dual
𐽄	SOGDIAN LETTER LESH	l	‍𐽄	‍𐽄‍	𐽄‍	dual

3.1.1 Note on representive glyphs

The far left column labeled ‘Nominal’ contains the representative form for each letter. This form is identical to the isolated or independent form of a letter. It is based upon the form of a letter that would occur in word-final position and unjoined to the preceding letter on account of a break in cursive joining (see § 4.2).

Only the isolated form of each letter is included in the proposed repertoire. Positional forms are to be maintained in a font and substituted by the shaping engine (see § 4.1). Some positional forms may not be palaeographically distinctive, but are differentiated typographically in order to illustrate the joining features of glyphs.

3.1.2 Notes on letters

aleph Specimens of the letter 𐼰 ALEPH are given in fig. 5.

beth and ***yodh*** The shapes for 𐼱 BETH and 𐼷 YODH converged graphically in later ‘cursive’ styles, ie. 𐼷. See fig. 6 and fig. 12 for specimens of BETH and YODH, respectively. In these cursive styles YODH may be typically left unconnected to the following letter when initial and medial in order to differentiate it from BETH, which maintains its joining behavior (see § 4.2). In some cases BETH is distinguished from YODH using a diacritic, eg. 𐼷 *yodh* and 𐼷𐽈 *beth*. The 𐼷 form of BETH is a glyphic variant and can be used in place of 𐼱 in a font designed for a specific variety of the ‘cursive’ script.

gimel and ***heth*** Initial and medial forms of 𐼲 GIMEL and 𐼶 HETH are identical. In these positions the connection between GIMEL and a following letter may be broken in order to distinguish it from HETH, which maintains its regular joining behavior (see § 4.2). The letter HETH has a variant final form ‍𐼶. See fig. 7 and fig. 11 for specimens of GIMEL and HETH, respectively.

daleth The letter *daleth* is not used in the Sogdian script. In the Panjakant ostracon, the character that appears in the position for *daleth* in the Aramaic order is the number 𐽓 20 (see fig. 3).

he The letter 𐼳 HE is used for marking a long vowel. It occurs only in final position. The standalone glyph 𐼳 and variations of it are also used for punctuation (see § 3.5).

teth The letter *teth* is not used in Sogdian, and is unattested in Old Sogdian. However, the Panjakant ostracon shows the sign in the position for *teth* in the original Aramaic order (see fig. 3). This sign is similar to the shape of *ayin* as found in the Aramaic heterogram for “said”, ie. (see *ayin* below). The sign is not proposed for encoding.

waw Specimens of the letter 𐼴 WAW are given in fig. 9.

zayin and ***nun*** The initial and medial forms of 𐼵 ZAYIN and 𐼻 NUN are identical. In these positions the connection between ZAYIN and a following letter may be broken in order to distinguish it from NUN (see §

4.2). In various texts ZAYIN is marked explicitly using a diacritic, eg. 𐼵𐽆 or 𐼵𐽇 (see § 3.3). Final ZAYIN has the variant shape 𐼵. See fig. 10 and fig. 16 for specimens of ZAYIN and NUN, respectively.

kaph Specimens of the letter 𐼸 KAPH are given in fig. 13. See also § 4.6 for details on the shaping of KAPH. The letter 𐼸 KAPH has the variant final shape 𐼸 (see § 3.1.3).

lamedh The letter 𐼹 LAMEDH appears in several sources as the 'hooked' form 𐼹. This form is not a distinct letter, but a glyphic variant. See fig. 14 for specimens.

mem Specimens of the letter 𐼺 MEM are given in fig. 15.

nun The letter 𐼻 NUN has the variant final shape 𐼻 (see § 3.1.3).

samekh Specimens of the letter 𐼼 SAMEKH are given in fig. 17.

ayin The letter *ayin* has two shapes: 𐼽 and 𐽀. Both appear only in Aramaic heterograms (see fig. 18 for attestations). The first form occurs in the heterogram "said", eg. 𐽀𐼻𐼷𐼽𐼴 *'NY'W*; 𐽀𐼽 *''* (see § 4.5). Initial and final forms are attested, but there is no dual-joining medial form, eg. * ‍𐼽‍. A letter that precedes word-medial 𐼽 will be shaped using its final form. The 𐼽 is likely an evolution of 𐼒 *OLD SOGDIAN LETTER AYIN, which was also used solely in an Aramaic heterogram. The 𐽀 may be considered the 'regular' shape of *ayin*. It also occurs in the heterogram "said", and in the heterogram 𐽀𐼺 *'M* 'with'. The shape 𐽀 of *ayin* is identical to that of 𐽀 *resh*; a feature already present in Old Sogdian. Therefore, a separate character for 'regular' 𐽀 *ayin* is not proposed. It is to be represented using 𐽀 RESH-AYIN. The 𐼽 is encoded as AYIN. The Panjakant ostracon shows the sign 𐽔 in the position for *ayin* in the original Aramaic order. It is similar to the sign used for the number 100.

pe Specimens of the letter 𐼾 PE are given in fig. 19. See also § 4.6 for details on the shaping of PE.

sadhe Specimens of the letter 𐼿 SADHE appear in fig. 20. It has variant final forms: 𐼿 and 𐼿 (see § 3.1.3).

resh The letter 𐽀 is used for writing both *resh* and *ayin* (see fig. 21). According to the Unicode character-glyph model, letters with identical graphical representations are considered glyphic variants and are unified as a single character. Accordingly, *ayin* is unified with *resh* as the character 𐽀 RESH-AYIN. Despite occurring after *ayin* in the alphabetical order, *resh* is ordered first in the name RESH-AYIN because it occurs more frequently in the sources.

shin Specimens of the letter 𐽁 SHIN are given in fig. 22. The alternate form 𐽅 of isolated *shin* is used in So 14830 for transcribing Chinese 所 *suǒ*. This 𐽅 has been included in the repertoire as the character SOGDIAN PHONOGRAM SHIN (see § 3.2).

taw Specimens of the letter 𐽂 TAW are given in fig. 23. It has the variant final forms 𐽂, 𐽂, and 𐽂 (see § 3.1.3). The usage of combining signs with TAW for representing the Sanskrit retroflex consonant *ṭa* [ʈ], eg. 𐽂𐽊 and 𐽂𐽍, is described in § 3.3.

feth The letter 𐽃 FETH is an extension of 𐼱 BETH that contains an extra hook at the left edge of the head. It is used for representing [f]. The character name is not historical and has been suggested by modern scholars. See fig. 24 for attestations.

lesh The 𐽄 LESH or 'hooked *resh*' is an extension of 𐽀 RESH-AYIN with a below-base hook (see fig. 25). The hook is an intrinsic part of the letter and is not a combining sign, ie. ◌̡. The 𐽄 is an atomic letter. The letter likely evolved from the practice of indicating [l] by placing a subscript *resh* below a regular *resh*, eg. 𐽄 (Yoshida, personal communication, 2016). The name 'LESH' is not historical and has been suggested by modern scholars. The name 'hooked *resh*' has been specified as an alias in the names list.

3.1.3 Note on final forms

Letters with final forms that contain a vertical terminal exhibit stylistic variation across script styles. The terminal strokes may be oriented in different directions, even within the same line. The orientation of terminals vary according to the whim of the scribe or the space available on a page. Terminal variation occurs most often at the end of a line for filling space or for compensating for lack of space at a margin. These stroke variations are stylistic and there is no semantic difference between final forms with different terminals. In the above table of letters, final forms with vertical strokes have been selected as representative forms. The table below shows the glyphic variants of final letters as they occur in the available sources. Alternate final forms are considered glyphic variants and may be controlled through fonts.

	Normative final	Alternate final(s)
ZAYIN	𐼵	𐼵
HETH	𐼶	𐼶
KAPH	𐼸	𐼸
NUN	𐼻	𐼻 𐼻
SADHE	𐼿	𐼿 𐼿 𐼿 𐼿
TAW	𐽂	𐽂 𐽂 𐽂 𐽂

3.2 Phonogram

Nominal	Character name	Latin	Final	Medial	Initial	Joining
𐽅	SOGDIAN PHONOGRAM SHIN	š	—	—	—	non

The 𐽅 PHONOGRAM SHIN is an alternate form of isolated 𐽁 *shin* that is used in So 14830 for transcribing Chinese 所 *suǒ* (=U+6240 CJK UNIFIED IDEOGRAPH-6240). Usage of this letter is attested solely in So 14830. See fig. 22 for specimens of PHONOGRAM SHIN and fig. 65 for the full folio of So 14830.

3.3 Combining signs

Eleven combining signs are used for disambiguation and transcription (see attestations in fig. 26–36):

	Character name	Example
◌𐽆	SOGDIAN COMBINING DOT BELOW	𐼵𐽆 ẓ
◌𐽇	SOGDIAN COMBINING TWO DOTS BELOW	𐼵𐽇 z̤
◌𐽈	SOGDIAN COMBINING DOT ABOVE	𐼱𐽈 ḃ
◌𐽉	SOGDIAN COMBINING TWO DOTS ABOVE	𐼲𐽉𐼰 ẍ, 𐼱𐽉 b̈
◌𐽊	SOGDIAN COMBINING CURVE ABOVE	𐽂𐽊𐼰 t̂
◌𐽋	SOGDIAN COMBINING CURVE BELOW	𐼲𐽋𐼰 x̮, 𐽀𐽋 r̮
◌𐽌	SOGDIAN COMBINING HOOK ABOVE	𐼸𐽌𐼰 k̇, 𐼾𐽌𐼰 ṗ,
◌𐽍	SOGDIAN COMBINING HOOK BELOW	𐼾𐽍𐼰 p̣ , 𐼲𐽍𐼰 x̣, 𐼸𐽍𐼰 ḳ, 𐼷𐽍 ỵ , 𐼵𐽍 ẓ
◌𐽎	SOGDIAN COMBINING LONG HOOK BELOW	𐼱𐽎 β̣
◌𐽏	SOGDIAN COMBINING RESH BELOW	𐽀𐽏 ṛ
◌𐽐	SOGDIAN COMBINING STROKE BELOW	𐽂𐽐 *t*

Usage of several of these signs is described in Yoshida 1994. The shapes of dots in these signs may vary according to script style or scribal whim: ◌𐽇 ◌𐽉 round; ◌𐽇 ◌𐽉 square; ◌𐽇 ◌𐽉 oblong. In some cases, a combining sign may be attached to the base letter by a line of ink, eg. 𐼵𐽆 is rendered as 𐼵𐽆.

3.4 Numbers

The following 4 numerical characters are included in the repertoire (see § 4.4 for a description of the notation system, and fig. 37–42 for attestations):

	Character name	Value	Final	Medial	Initial	Joining
𐽑	SOGDIAN NUMBER ONE	1	𐽑	𐽑	𐽑	dual
𐽒	SOGDIAN NUMBER TEN	10	𐽒	𐽒	𐽒	dual
𐽓	SOGDIAN NUMBER TWENTY	20	𐽓	𐽓	𐽓	dual

𐽔	SOGDIAN NUMBER ONE HUNDRED	100	𐽔	—	—	right

ONE Single and multiple instances of the number 𐽑 ONE are used for representing 1–9. The primary units are expressed using groups of three or four instances of ONE separated by spaces, eg. 𐽑𐽑 for 2, 𐽑𐽑 𐽑𐽑𐽑 for 5, 𐽑𐽑𐽑𐽑 𐽑𐽑𐽑𐽑 for 8, etc. The number 1 is generally not written using 𐽑 ONE, but as the word 𐼴𐼷𐼰 *ʾyw*. The grouping principle is derived from Old Sogdian, for which the characters 𐼝 *OLD SOGDIAN NUMBER ONE .. 𐼡 *OLD SOGDIAN NUMBER FIVE are proposed as atomic characters in order to facilitate grouping. The Old Sogdian encoding requires pre-composed groups because the script is non-conjoining and there is no simple method for ligating sequences of 𐼝 *OLD SOGDIAN NUMBER ONE. As Sogdian is an inherently conjoining, sequences of 𐽑 ONE will join to preceding and following characters by default.

TEN The 𐽒 TEN resembles 𐼹 LAMEDH, but is generally written with a smaller angle.

TWENTY The 𐽓 TWENTY is palaeographically composed of a vertical stack of two 𐽒 TEN-s. In some sources this origin is apparent in the glyphic representation, ie. 𐽓 (So 14680 v; see fig. 39).

ONE HUNDRED The 𐽔 ONE HUNDRED is a graphically complex character, which appears to be composed of a sequence of 𐼰 *aleph* and 𐼲 *gimel* or 𐼶 *heth* terminated by a loop. Nonetheless, it is interpreted as a single unit and is encoded as an atomic character. The sign represents the value 100, but also functions as a unit mark for multiples of hundred. The ONE HUNDRED is generally unconnected from preceding numbers or words when it is used as a unit mark, but in some sources it is joined to the right. This character is highly stylized and has glyphic variants, eg. 𐽔, 𐽔, etc.

3.5 Punctuation

Five punctuation signs are included in the proposed repertoire (see attestations in fig. 43–47):

	Character name
𐽕	SOGDIAN PUNCTUATION TWO VERTICAL BARS
𐽖	SOGDIAN PUNCTUATION TWO VERTICAL BARS WITH DOTS
𐽗	SOGDIAN PUNCTUATION CIRCLE WITH DOT
𐽘	SOGDIAN PUNCTUATION TWO CIRCLES WITH DOTS
𐽙	SOGDIAN PUNCTUATION HALF CIRCLE WITH DOT

The signs 𐽕, 𐽖, 𐽗, and 𐽘 are the common forms of punctuation. They are used for delimiting text segments of various length: word boundaries, the end of major sections, complete texts. The 𐽙 is generally used as a marker for indicating the completion of a text.

There is variation in the shapes of these signs. Strokes may be elongated, rounded, or truncated. For example, the shape of PUNCTUATION TWO VERTICAL BARS may vary from 𐽕 full-height lines to ıı mid-height lines to •• dots. The dots in 𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS may be rendered variously, eg. 𐽖 and 𐽖.

Such variation complicates a determination regarding the number of unique forms of punctuation in Sogdian records, especially for purposes of character encoding. Both •• and 𐽕 are used in similar contexts and may have identical functions. But, it is not certain if the scribe intentionally drew •• dots instead of 𐽕 strokes. Moreover, it is quite possible that PUNCTUATION TWO VERTICAL BARS has different renderings in particular styles of the script. It is possible that scribes viewed •• and 𐽕 as two separate signs of punctuation at different stages in the development of the script. Similarly, both ❖ and 𐽖 occur commonly. Yet, in some cases it is difficult to determine if ❖ is a unique sign or a form of 𐽖 in which the vertical bars are rendered as dots. In So 18242, a colophon is indicated using 𐽕❖𐽕❖𐽕, in which the proportions of the strokes of 𐽕 and ❖ makes it clear that ❖ is not a variant form of 𐽖 (see fig. 51).

Various other forms of punctuation are attested in Sogdian manuscripts (see fig. 48–55). In addition to the six proposed characters, the • dot, ● large dot, ⦂ two vertical dots, and ○ circle are used as punctuation in various contexts. Characters such as ⁜, ✛, ✥ are used for indicating the end of major sections of text. A sign |ı| is used in conjunction with 𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS for indicating end of a major section of text. The sign ıı=ıı is used at the end of text in Pelliot Sogdien 18.

A sign resembling ▬ is used for filling gaps at the end of line. It also has various shapes. It may occur as 𐼳, which resembles the letter 𐼳 HE; as the horizonally compressed ▬, which may be confused with the Aramaic heterogram 𐼵𐼷 *ZY* in some instances; and as a simpler stroke ▬ that may be notched at the beginning or ▬ elevated at the end. Even the shape of ▬ may be open to interpretation, as the triangular terminal may be a result of drawing the terminal of 𐼳 downward instead of parallel to the baseline.

These other signs are not proposed for encoding within the Sogdian block at present. The reason is that they resemble punctuation characters that are already encoded in Unicode. For example, it is apparent that ❖ is a commonly-used sign in Sogdian manuscripts, but the character ⁘ U+2058 FOUR DOT PUNCTUATION already exists in Unicode. It may be appropriate to use U+2058 FOUR DOT PUNCTUATION in Sogdian contexts in order to not encode another punctuation sign in Unicode with nearly identical graphical and semantic properties. Shown below are punctuation characters that appear in Sogdian documents, and corresponding Unicode characters with similar appearances, if such exist:

	Description	Existing character
•	dot	. U+002E FULL STOP · U+00B7 MIDDLE DOT • U+10AF4 MANICHAEAN PUNCTUATION DOT
●	large dot	● U+26AB MEDIUM BLACK CIRCLE
••	two dots	two instances of · U+00B7 MIDDLE DOT
⦂	two vertical dots	: U+003A COLON ⁚ U+205A TWO DOT PUNCTUATION ⦂ U+10AF5 MANICHAEAN PUNCTUATION TWO DOTS

❖	four dots	⁘ U+2058 FOUR DOT PUNCTUATION ⁛ U+205B FOUR DOT MARK ܀ U+0700 SYRIAC END OF PARAGRAPH
⸭	five dots	⸭ U+2E2D FIVE DOT MARK
✢	cross with four dots	
⁜	cross with eight dots	܍ U+070D SYRIAC HARKLEAN ASTERISCUS ⁜ U+205C DOTTED CROSS
⦀	three vertical bars	*no corresponding form currently encoded*
○	circle	○ U+25CB WHITE CIRCLE
▬	notched line	*no corresponding form currently encoded*

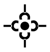
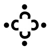
Of the five Sogdian punctuation signs proposed for encoding, three resemble characters already encoded in Unicode, as shown below. These three have been included in the Sogdian block because they are used very commonly within Sogdian contexts.

Sogdian sign	Existing character
॥	॥ U+0965 DEVANAGARI DOUBLE DANDA
⊙	𐫳 U+10AF3 MANICHAEAN PUNCTUATION DOT WITHIN DOT
⊙⊙	𐫲 U+10AF2 MANICHAEAN PUNCTUATION DOUBLE DOT WITHIN DOT

At present, the five punctuation signs proposed for encoding are sufficient for representing the majority of Sogdian texts. Space exists within the Sogdian block for the encoding of additional punctuation signs in the future. The proposal author will continue discussing the matter with experts and may present requests for encoding additional punctuation if experts require such characters. Until that time, other punctuation signs are to be unified with existing Unicode characters specified in the tables above. For these, the code point for the unified characters should be replaced in Sogdian fonts with glyphs designed to reflect the script style appropriately.

3.5.1 Ornamental characters

Various types of characters are used for illumination or ornamentation, as shown below (see also fig. 56–57). They are not presently proposed for inclusion in the Sogdian block because of their ornamental nature, although some are similar to characters already encoded in Unicode:

	Description	Existing character
	three petal fleuron	*no corresponding form currently encoded*
	four petal fleuron	𐫱 U+10AF1 MANICHAEAN PUNCTUATION FLEURON
	four petal fleuron with rays	*no corresponding form currently encoded*
	four petal fleuron with outer rays	𐫰 U+10AF0 MANICHAEAN PUNCTUATION STAR
	four petal fleuron with dots	*no corresponding form currently encoded*
	three petal fleuron with rays	*no corresponding form currently encoded*
	right-facing three petal fleuron	*no corresponding form currently encoded*
	left-facing three petal fleuron	*no corresponding form currently encoded*

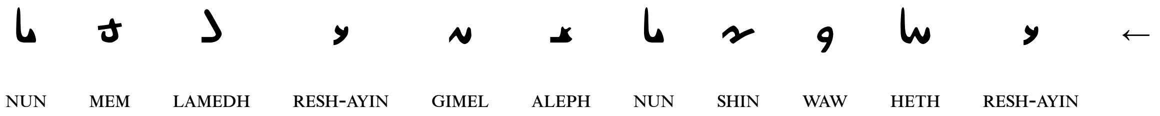
3.5.2 Editorial marks

Signs such as '+' are used in manuscripts for editorial insertions (see fig. 58). These characters are not included in the proposed repertoire at this time. It may be possible to unify them with signs used in other scripts for similar purposes.

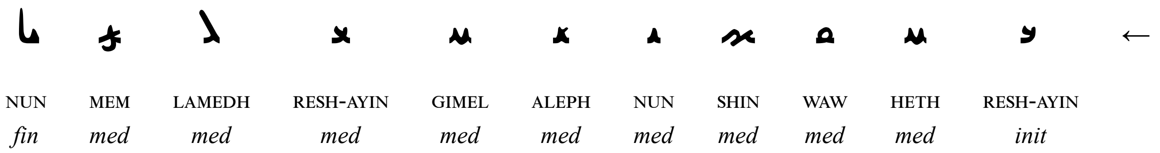
4 Script Details

4.1 Encoding model

Sogdian may be implemented using the Unicode Bidirectional Algorithm. The shaping requirements are similar to that of Arabic. Letters join to adjacent letters at the baseline. Letters have different forms when they occur in initial, medial, and final positions in a word. The shaping behavior for Sogdian is shown below using the input string for the word *rxwšnʾγrδmn* 'paradise':



The shaping engine substitutes the nominal glyph for each letter with the appropriate positional glyph:



The rendered output would be:

𐽀𐼶𐼴𐽁𐼻𐼰𐼷𐽀𐼹𐼺𐼻

rxwšnʾyrδmn

4.2 Modifying cursive joining

The letters 𐼲 GIMEL, 𐼵 ZAYIN, 𐼷 YODH may remain unjoined from a following letter when initial or medial. The breaking of regular joining behavior can be managed using the generic Unicode control character ‌ U+200C ZERO WIDTH NON-JOINER (abbreviated as ZWNJ). This character is placed after the letter that should remain unjoined. The letter before ZWNJ is shaped using its final form; the following letter takes its initial form if non-final in a word, or its isolated form if in word-final position.

Shown below are different spellings of *syγwδyk* ‘Sogdian’ (= *swγδyk*) that occur in correspondence from Đēwāštīc (*δywʾštyc*), a Sogdian ruler of Panjakant, which were found at a fortress on Mt. Mug. These spellings show unjoined forms of medial GIMEL and YODH, ie. *syγwδyk* (Mug A-2), *syγwδy-k* (Mug A-3), *syγ-wδyk* (Mug A-16). When a letter is unjoined, it is followed by a ‘-’ hyphen in transliteration.

	Output	Input string →
syγwδyk	𐼼𐼲𐼴𐼹𐼷𐼸	𐼼 SAMEKH, 𐼲 GIMEL, 𐼴 WAW, 𐼹 LAMEDH, 𐼷 YODH, 𐼸 KAPH
syγ-wδyk	𐼼𐼲‌𐼴𐼹𐼷𐼸	𐼼 SAMEKH, 𐼲 GIMEL, ‌ ZWNJ, 𐼴 WAW, 𐼹 LAMEDH, 𐼷 YODH, 𐼸 KAPH
syγwδy-k	𐼼𐼲𐼴𐼹𐼷‌𐼸	𐼼 SAMEKH, 𐼲 GIMEL, 𐼴 WAW, 𐼹 LAMEDH, 𐼷 YODH, ‌ ZWNJ, 𐼸 KAPH

4.3 Combining signs

Combining signs are placed after the base letter in the input sequence. If ‌ ZWNJ is used after a letter for breaking cursive joining and a combining sign is also applied to the base letter, then the sign is placed after ZWNJ.

	Output	Input string →
cr̮	𐼿𐽀𐽋	𐼿 SADHE, 𐽀 RESH-AYIN, ◌𐽋 COMBINING CURVE BELOW
rẓʾy	𐽀𐼵𐽆𐼷	𐽀 RESH-AYIN, 𐼵 ZAYIN, ◌𐽆 COMBINING DOT BELOW, 𐼷 YODH
ṛy	𐽀𐽏𐼷	𐽀 RESH-AYIN, ◌𐽏 COMBINING RESH BELOW, 𐼷 YODH
kwṭy	𐼸𐼴𐽂𐽐𐼷	𐼸 KAPH, 𐼴 WAW, 𐽂 TAW, ◌𐽐 COMBINING STROKE BELOW, 𐼷 YODH

z-wrpw	𐼵𐽇‌𐼴𐽀𐼾𐼴	𐼵 ZAYIN, ◌𐽇 COMBINING TWO DOTS BELOW, ‌ ZWNJ, 𐼴 WAW, 𐽀 RESH-AYIN, 𐼾 PE, 𐼴 WAW

4.4 Numerical notation

Numerical notation follows that of ‘Old Sogdian’. Attestations are given in fig. 37–42. The numbers 1–9 are expressed using repetitions of 𐽑 ONE. The numbers 5–9 are arranged in groups of three or four instances of ONE separated by spaces. The arrangements are shown below:

	Output	Input string →
1	𐽑	𐽑 ONE
2	𐽑𐽑	𐽑 ONE, 𐽑 ONE
3	𐽑𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE
4	𐽑𐽑𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE
5	𐽑𐽑𐽑 𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽑 ONE, 𐽑 ONE
6	𐽑𐽑𐽑 𐽑𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE
7	𐽑𐽑𐽑𐽑 𐽑𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE
8	𐽑𐽑𐽑𐽑 𐽑𐽑𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE
9	𐽑𐽑𐽑 𐽑𐽑𐽑 𐽑𐽑𐽑	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE

The tens are written using sequences of 𐽒 TEN and 𐽓 TWENTY. Even multiples are expressed with repetitions of TWENTY, not TEN. Odd multiples are produced by attaching TEN at the end.

	Output	Input string →
10	𐽒	𐽒 TEN
20	𐽓	𐽓 TWENTY
30	𐽓𐽒	𐽓 TWENTY, 𐽒 TEN

40	𐽓𐽓	𐽓 TWENTY, 𐽓 TWENTY
50	𐽓𐽓𐽒	𐽓 TWENTY, 𐽓 TWENTY, 𐽒 TEN
60	𐽓𐽓𐽓	𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY
70	𐽓𐽓𐽓𐽒	𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY, 𐽒 TEN
80	𐽓𐽓𐽓𐽓	𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY
90	𐽓𐽓𐽓𐽓𐽒	𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY, 𐽒 TEN

The hundreds are represented using 𐽔 ONE HUNDRED. Multiples are generally expressed using Sogdian names for cardinal numbers followed by 𐽔, eg. 200 is “*ʾδwy* 100”; 500 is “*pnc* 100”. The number word may be written separately or may ligate with ONE HUNDRED. However, multiples may also be expressed using iterations of 𐽑 ONE, eg. 300 may be written as “3 100”.

	Output	Input string →
100	𐽔	𐽔 ONE HUNDRED
200	𐼰𐼹𐼴𐼷 𐽔	𐼰 ALEPH, 𐼹 LAMEDH, 𐼴 WAW, 𐼷 YODH, ␠ SPACE, 𐽔 ONE HUNDRED
300	𐽑𐽑𐽑 𐽔	𐽑 ONE, 𐽑 ONE, 𐽑 ONE, ␠ SPACE, 𐽔 ONE HUNDRED
500	𐼾𐼻𐼿 𐽔	𐼾 PE, 𐼻 NUN, 𐼿 SADHE, ␠ SPACE, 𐽔 ONE HUNDRED
500	𐼾𐼻𐼿𐽔	𐼾 PE, 𐼻 NUN, 𐼿 SADHE, 𐽔 ONE HUNDRED

One thousand is represented as 𐽑𐼹𐼾𐼴 *1LPw*, which consists of the number 𐽑 ONE prefixed to the Aramaic heterogram 𐼹𐼾𐼴 *LPw*. The sequence 𐽑𐼹𐼾𐼴 functions as a unit mark for the thousands. Multiples of this order are expressed using number words. The unit 𐽑𐼹𐼾𐼴 does not join to other words.

	Output	Input string →
1000	𐽑𐼹𐼾𐼴	𐽑 ONE, 𐼹 LAMEDH, 𐼾 PE, 𐼴 WAW
8000	𐼰𐽁𐼹𐽂 𐽑𐼹𐼾𐼴	𐼰 ALEPH, 𐽁 SHIN, 𐽂 TAW, ␠ SPACE, 𐽑 ONE, 𐼹 LAMEDH, 𐼾 PE, 𐼴 WAW

The primary units are generally written after the tens in compound numbers:

	Output	Input string →
11	𐽒𐽑	𐽒 TEN, 𐽑 ONE
32	𐽓𐽒𐽑𐽑	𐽓 TWENTY, 𐽒 TEN, 𐽑 ONE, 𐽑 ONE
81	𐽓𐽓𐽓𐽓𐽑	𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY, 𐽓 TWENTY, 𐽑 ONE
155	𐽔 𐽓𐽓𐽒𐽑𐽑𐽑 𐽑𐽑	𐽔 ONE HUNDRED, SP SPACE, 𐽓 TWENTY, 𐽓 TWENTY, 𐽒 TEN, 𐽑 ONE, 𐽑 ONE, 𐽑 ONE, SP SPACE, 𐽑 ONE, 𐽑 ONE

4.5 Aramaic heterograms

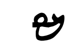
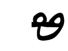
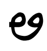
Aramaic heterograms are represented as words spelled using conventional letters, eg. 𐼸𐼵𐼻𐼳 *KZNH*, is written <𐼸 KAPH, 𐼵 ZAYIN, 𐼻 NUN, 𐼳 HE>. The heterogram for "said" presents a curious exception. It has several representations, all of which contain a special form of *ayin* (see fig. 18):

	Output	Input string →
ʿNYʿW	𐽀𐼻𐼷𐼽𐼴	𐽀 RESH-AYIN, 𐼻 NUN, 𐼷 YODH, 𐼽 AYIN, 𐼴 WAW
ʿʿ	𐽀𐼽	𐽀 RESH-AYIN, 𐼽 AYIN
ʿWYʿ	𐽀𐼴𐼷𐼽	𐽀 RESH-AYIN, 𐼴 WAW, 𐼷 YODH, 𐼽 AYIN
⋯*ʿW*	⋯𐼽𐼴	⋯ U+22EF MIDLINE HORIZONTAL ELLIPSIS, 𐼽 AYIN, 𐼴 WAW

Each of these spellings contains an extra *ayin* that is not present in the original Aramaic. The word-initial *ayin* is represented using 𐽀 RESH-AYIN, while the special *ayin* is represented using 𐼽 AYIN. In 𐽀𐼻𐼷𐼽𐼴 and 𐽀𐼴𐼷𐼽 the 𐼽 is disconnected from the preceding 𐼷 YODH. In 𐽀𐼽 the initial 𐽀 connects to 𐼽. The heterogram 𐽀𐼽 may be a contraction of 𐽀𐼴𐼷𐼽 in which the middle letters 𐼴𐼷 have been omitted. It is transliterated as *ʿNY~W* by MacKenzie (1976: 42) and Reck (2016: 300), and as ® by Yakubovich and Yoshida (2005: 245, 247). The 𐼽 in 𐽀𐼴𐼷𐼽 is ignored by Benveniste, who transliterates the word as *RWY* using the value of 𐽀 as *resh* (1940: 102). The form ⋯𐼽𐼴, which occurs in So 20241a, is interesting in that only the end is represented, while the beginning (presumably 𐽀𐼴𐼷 or 𐽀𐼻𐼷) is omitted and replaced by an ellipsis. The attestations for the heterogram "said" illustrate that it may adequately be represented as a conventional word, spelled using letters in the proposed repertoire. It is unnecessary to treat the heterogram as an atomic unit. Indeed, representing it using individual letters enables users to capture the text in an encoded string as it was originally written by scribes.

4.6 Glyph interactions

The letters 𐼸 KAPH and 𐼾 PE exhibit special joining behavior with a following 𐼴 WAW. They do not generally touch the body of WAW at the baseline, but connect to it with their terminal strokes at the baseline from below.

Input string →	incorrect	correct
𐼸 KAPH, 𐼴 WAW		
𐼾 PE, 𐼴 WAW		

By default, the glyph 𐼴 for WAW would be substituted by its medial form 𐼴. Instead, the initial form 𐼴 should be used.

4.7 Elongation

In late cursive styles the connection between letters within a word may be elongated at the baseline (see fig. 59). This also occurs with the stroke of a final letter. This technique is used for justification and line filling, and is similar to *kashida* in the Arabic script. It can be implemented for Sogdian through the usage of ـ U+0640 ARABIC TATWEEL, as has been done for other right-to-left scripts. This character may be placed after a letter in any position within a word. The elongation should not break across the end of a line.

	Output	Input string →
prtw	𐼾𐽀𐽂𐼴	𐼾 PE, 𐽀 RESH-AYIN, 𐽂 TAW, 𐼴 WAW
prtw	𐼾𐽀𐽂ـ𐼴	𐼾 PE, 𐽀 RESH-AYIN, 𐽂 TAW, ـ U+0640 ARABIC TATWEEL, 𐼴 WAW

There is no need to encode a Sogdian-specific character. The U+0640 ARABIC TATWEEL has been extended for usage with Sogdian in `ScriptExtensions.txt` (see § 5.4). Accordingly, elongation would be implemented for Sogdian by placing a Sogdian-specific glyph at the code point U+0640 in the Sogdian font. This glyph should be designed such that the stroke thickness matches that of other Sogdian glyphs.

4.8 Line-breaking

There are no conventions for line-breaking; consequently, hyphens or other continuation marks are not attested. The available sources show line-break occuring after the end of a word or a numerical sequence. Words are not usually split across lines. In the rare cases where words are broken across lines, such splits are not marked.

4.9 Collation

The sort order for Sogdian is as follows:

> 𐼰 ALEPH < 𐼱 BETH < 𐼲 GIMEL < 𐼳 HE < 𐼴 WAW < 𐼵 ZAYIN < 𐼶 HETH <
> 𐼷 YODH < 𐼸 KAPH < 𐼹 LAMEDH < 𐼺 MEM < 𐼻 NUN < 𐼼 SAMEKH < 𐼽 AYIN <
> 𐼾 PE < 𐼿 SADHE < 𐽀 RESH-AYIN < 𐽁 SHIN ≪ 𐽅 PHONOGRAM SHIN < 𐽂 TAW <
> 𐽃 FETH < 𐽄 LESH

4.10 Vertical text

While the Sogdian folios shown in this proposal are displayed horizontally, and modern scholars are accustomed to reading the script horizontally, Yoshida (2013) suggests that the script was often written vertically, and that the correct orientation of several manuscripts may inn fact be vertical.

Given the constraints of software and user interfaces, Sogdian may be displayed horizontally in plain text. However, support for vertical orientations of the script is required for accurately displaying Sogdian text that is natively vertical. In vertical environments, Sogdian text is oriented from top to bottom with lines that advance from left to right. Letters are rotated 90° counter-clockwise from their upright orientations.

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 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.

There are some exceptions to the orientation of Sogdian in manuscripts containing Indic scripts. For example, as shown in the folio of the *Nīlakaṇṭha-dhāraṇī* (fig. 63), Sogdian is written upside down beneath the left-to-right Siddham text.

5 Character Data

5.1 Character Properties

In the format of `UnicodeData.txt`:

```
10F30;SOGDIAN LETTER ALEPH;Lo;0;AL;;;;;N;;;;;
10F31;SOGDIAN LETTER BETH;Lo;0;AL;;;;;N;;;;;
10F32;SOGDIAN LETTER GIMEL;Lo;0;AL;;;;;N;;;;;
10F33;SOGDIAN LETTER HE;Lo;0;AL;;;;;N;;;;;
10F34;SOGDIAN LETTER WAW;Lo;0;AL;;;;;N;;;;;
10F35;SOGDIAN LETTER ZAYIN;Lo;0;AL;;;;;N;;;;;
10F36;SOGDIAN LETTER HETH;Lo;0;AL;;;;;N;;;;;
10F37;SOGDIAN LETTER YODH;Lo;0;AL;;;;;N;;;;;
10F38;SOGDIAN LETTER KAPH;Lo;0;AL;;;;;N;;;;;
10F39;SOGDIAN LETTER LAMEDH;Lo;0;AL;;;;;N;;;;;
10F3A;SOGDIAN LETTER MEM;Lo;0;AL;;;;;N;;;;;
```

```
10F3B;SOGDIAN LETTER NUN;Lo;0;AL;;;;;N;;;;;
10F3C;SOGDIAN LETTER SAMEKH;Lo;0;AL;;;;;N;;;;;
10F3D;SOGDIAN LETTER AYIN;Lo;0;AL;;;;;N;;;;;
10F3E;SOGDIAN LETTER PE;Lo;0;AL;;;;;N;;;;;
10F3F;SOGDIAN LETTER SADHE;Lo;0;AL;;;;;N;;;;;
10F40;SOGDIAN LETTER RESH-AYIN;Lo;0;AL;;;;;N;;;;;
10F41;SOGDIAN LETTER SHIN;Lo;0;AL;;;;;N;;;;;
10F42;SOGDIAN LETTER TAW;Lo;0;AL;;;;;N;;;;;
10F43;SOGDIAN LETTER FETH;Lo;0;AL;;;;;N;;;;;
10F44;SOGDIAN LETTER LESH;Lo;0;AL;;;;;N;;;;;
10F45;SOGDIAN PHONOGRAM SHIN;Lo;0;AL;;;;;N;;;;;
10F46;SOGDIAN COMBINING DOT BELOW;Mn;220;NSM;;;;;N;;;;;
10F47;SOGDIAN COMBINING TWO DOTS BELOW;Mn;220;NSM;;;;;N;;;;;
10F48;SOGDIAN COMBINING DOT ABOVE;Mn;230;NSM;;;;;N;;;;;
10F48;SOGDIAN COMBINING TWO DOTS ABOVE;Mn;230;NSM;;;;;N;;;;;
10F4A;SOGDIAN COMBINING CURVE ABOVE;Mn;230;NSM;;;;;N;;;;;
10F4B;SOGDIAN COMBINING CURVE BELOW;Mn;220;NSM;;;;;N;;;;;
10F4C;SOGDIAN COMBINING HOOK ABOVE;Mn;230;NSM;;;;;N;;;;;
10F4D;SOGDIAN COMBINING HOOK BELOW;Mn;220;NSM;;;;;N;;;;;
10F4E;SOGDIAN COMBINING LONG HOOK BELOW;Mn;220;NSM;;;;;N;;;;;
10F4F;SOGDIAN COMBINING RESH BELOW;Mn;220;NSM;;;;;N;;;;;
10F50;SOGDIAN COMBINING STROKE BELOW;Mn;220;NSM;;;;;N;;;;;
10F51;SOGDIAN NUMBER ONE;No;0;AL;;;;1;N;;;;;
10F52;SOGDIAN NUMBER TEN;No;0;AL;;;;10;N;;;;;
10F53;SOGDIAN NUMBER TWENTY;No;0;AL;;;;20;N;;;;;
10F54;SOGDIAN NUMBER ONE HUNDRED;No;0;AL;;;;100;N;;;;;
10F55;SOGDIAN PUNCTUATION TWO VERTICAL BARS;Po;0;AL;;;;;N;;;;;
10F56;SOGDIAN PUNCTUATION TWO VERTICAL BARS WITH DOTS;Po;0;AL;;;;;N;;;;;
10F57;SOGDIAN PUNCTUATION CIRCLE WITH DOT;Po;0;AL;;;;;N;;;;;
10F58;SOGDIAN PUNCTUATION TWO CIRCLES WITH DOTS;Po;0;AL;;;;;N;;;;;
10F59;SOGDIAN PUNCTUATION HALF CIRCLE WITH DOT;Po;0;AL;;;;;N;;;;;
```

5.2 Linebreaking

In the format of `LineBreak.txt`:

```
10F30..10F44;AL   # Lo  [21] SOGDIAN LETTER ALEPH..SOGDIAN LETTER LESH
10F45;AL          # Lo       SOGDIAN PHONOGRAM SHIN
10F46..10F50;CM   # Mn  [11] SOGDIAN COMBINING DOT BELOW..
                             SOGDIAN COMBINING STROKE BELOW
10F51..10F54;AL   # No   [4] SOGDIAN NUMBER ONE..SOGDIAN NUMBER ONE HUNDRED
10F55..10F59;AL   # Po   [5] SOGDIAN PUNCTUATION TWO VERTICAL BARS..
                             SOGDIAN PUNCTUATION HALF CIRCLE WITH DOT
```

5.3 Shaping Properties

In the format of `ArabicShaping.txt`:

```
# Sogdian Characters

10F30; SOGDIAN ALEPH; D; SOGDIAN ALEPH
10F31; SOGDIAN BETH; D; SOGDIAN BETH
10F32; SOGDIAN GIMEL; D; SOGDIAN GIMEL
10F33; SOGDIAN HE; R; SOGDIAN HE
10F34; SOGDIAN WAW; D; SOGDIAN WAW
10F35; SOGDIAN ZAYIN; D; SOGDIAN ZAYIN
```

```
10F36; SOGDIAN HETH; D; SOGDIAN HETH
10F37; SOGDIAN YODH; D; SOGDIAN YODH
10F38; SOGDIAN KAPH; D; SOGDIAN KAPH
10F39; SOGDIAN LAMEDH; D; SOGDIAN LAMEDH
10F3A; SOGDIAN MEM; D; SOGDIAN MEM
10F3B; SOGDIAN NUN; D; SOGDIAN NUN
10F3C; SOGDIAN SAMEKH; D; SOGDIAN SAMEKH
10F3D; SOGDIAN AYIN; D; SOGDIAN AYIN
10F3E; SOGDIAN PE; D; SOGDIAN PE
10F3F; SOGDIAN SADHE; D; SOGDIAN SADHE
10F40; SOGDIAN RESH-AYIN; D; SOGDIAN RESH-AYIN
10F41; SOGDIAN SHIN; D; SOGDIAN SHIN
10F42; SOGDIAN TAW; D; SOGDIAN TAW
10F43; SOGDIAN FETH; D; SOGDIAN FETH
10F44; SOGDIAN LESH; D; SOGDIAN LESH
10F45; SOGDIAN PHONOGRAM SHIN; N; SOGDIAN PHONOGRAM SHIN
10F50; SOGDIAN ONE; D; SOGDIAN ONE
10F51; SOGDIAN TEN; D; SOGDIAN TEN
10F52; SOGDIAN TWENTY; D; SOGDIAN TWENTY
10F53; SOGDIAN ONE HUNDRED; R; SOGDIAN ONE HUNDRED
```

5.4 Script Extensions

The following character is to be extended for usage in Sogdian: in `ScriptExtensions.txt`:

```
0640 ; # Lm  ARABIC TATWEEL
```

6 References

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

———. 2016b. “Recommendations to UTC #148 August 2016 on Script Proposals”. L2/16-216. http://www.unicode.org/L2/L2016/16216-script-ad-hoc.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

Benveniste, Émile. 1940. *Textes Sogdiens: étités, traduits et commentés.* Mission Pelliot en Asie Centrale, série in-quarto, III. Paris: Librairie orientaliste Paul Geuthner.

———. 1946. *Vessantara Jātaka: texte Sogdien*. Mission Pelliot en Asie Centrale, série in-quarto, IV. Paris: Librairie orientaliste Paul Geuthner.

“Coins of Central Asia”. http://www.sogdcoins.narod.ru/english/sogdiana/e_coins2.html

Coulmas, Florian. 1996. *The Blackwell Encyclopedia of Writing Systems*. Oxford: Blackwell Publishers.

Kara, György. 1996. “Aramaic Scripts for Altaic Languages”. *The World's Writing Systems*, edited by Peter T. Daniels and W. Bright, pp. 536–558. New York and Oxford: Oxford University Press.

Livshits, V. A. 2015. “A Sogdian alphabet from Penjikent”. *Sogdian Epigraphy of Central Asia and Semirech'e*, pp. 227–232. Corpus inscriptonum iranicarum. Part II. Inscriptions of the Seleucid and Parthian periods of Eastern Iran and Central Asia. Vol. III. Sogdian. Translated by Tom Stableford, edited by Nicholas Sims-Williams. London: School of Oriental and African Studies.

———. 2015. “Sogdian documents from the Fortress of Chilkhujra”, pp. 217–226. *Ibid*.

———. 2015. “Sogdian epigraphy from Semirech'e”, pp. 269–296. *Ibid*.

MacKenzie, David N. [ed]. 1976. *The Buddhist Sogdian Texts of the British Library*. Acta Iranica 10. Troisième série, Textes et mémoires. Téhéran-Liège: Bibliothèque Pahlavi; Leiden: E. J. Brill.

Osman, Omarjan. 2013. “Proposal to Encode the Uyghur Script in ISO/IEC 10646”. L2/13-071. `http://www.unicode.org/L2/L2013/13071-uyghur.pdf`

Pandey, Anshuman. 2016. “Proposal to encode the Old Sogdian script in Unicode”. L2/16-312R. `http://www.unicode.org/L2/L2016/16312r-old-sogdian.pdf`

Reck, Christiane. 2016. *Mitteliranische Handschriften*. Teil 2. Berliner Turfanfragmente buddhistischen Inhalts in sogdischer Schrift. Stuttgart: Franz Steiner Verlag.

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.

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`

Yakubovich, Ilya; Yoshida, Yutaka. 2005. “The Sogdian Fragments of *Saṃghāṭasūtra* in the German Turfan Collection”. *Languages of Iran: Past and Present*, ed. Dieter Weber, pp. 239–268. Weisbaden: Harrassowitz Verlag.

Yoshida, Yutaka. 1994. 「ソグド文字で表記された漢字音」 [“Sogudomoji de hyōkisareta kanjion” = “Chinese in Sogdian script”]. 『東方学報　京都』 [*Tōhō Gakuhō: Journal of Oriental Studies*], 京都市 [Kyōto-shi] 66, 1994, pp. 380–271.

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

7 Acknowledgments

I express my sincere gratitude to Nicholas Sims-Williams (SOAS, University of London) and Yutaka Yoshida (University of Kyoto) for reviewing earlier versions of this proposal, for providing extensive and insightful comments, and for informative discussions. I thank them both for their patience and for indulging my ignorance of the script.

I also thank Gabriel McKee for granting me access to materials at the Institute for the Study of the Ancient World, New York University.

The present proposal was funded in part by the Adopt-A-Character Program of the Unicode Consortium, and supervised by Deborah Anderson and Rick McGowan.

	10F3	10F4	10F5	10F6
0	𐼰 10F30	𐽀 10F40	◌𐽐 10F50	
1	𐼱 10F31	𐽁 10F41	𐽑 10F51	
2	𐼲 10F32	𐽂 10F42	𐽒 10F52	
3	𐼳 10F33	𐽃 10F43	𐽓 10F53	
4	𐼴 10F34	𐽄 10F44	𐽔 10F54	
5	𐼵 10F35	𐽅 10F45	𐽕 10F55	
6	𐼶 10F36	◌𐽆 10F46	𐽖 10F56	
7	𐼷 10F37	◌𐽇 10F47	𐽗 10F57	
8	𐼸 10F38	◌𐽈 10F48	𐽘 10F58	
9	𐼹 10F39	◌𐽉 10F49	𐽙 10F59	
A	𐼺 10F3A	◌𐽊 10F4A		
B	𐼻 10F3B	◌𐽋 10F4B		
C	𐼼 10F3C	◌𐽌 10F4C		
D	𐼽 10F3D	◌𐽍 10F4D		
E	𐼾 10F3E	◌𐽎 10F4E		
F	𐼿 10F3F	◌𐽏 10F4F		

This block unifies the 'formal' and 'cursive' scripts. Representative glyphs are based upon the ‘formal’ style.

Letters

10F30 𐼰 SOGDIAN LETTER ALEPH
10F31 𐼱 SOGDIAN LETTER BETH
10F32 𐼲 SOGDIAN LETTER GIMEL
10F33 𐼳 SOGDIAN LETTER HE
10F34 𐼴 SOGDIAN LETTER WAW
10F35 𐼵 SOGDIAN LETTER ZAYIN
10F36 𐼶 SOGDIAN LETTER HETH
10F37 𐼷 SOGDIAN LETTER YODH
10F38 𐼸 SOGDIAN LETTER KAPH
10F39 𐼹 SOGDIAN LETTER LAMEDH
10F3A 𐼺 SOGDIAN LETTER MEM
10F3B 𐼻 SOGDIAN LETTER NUN
10F3C 𐼼 SOGDIAN LETTER SAMEKH
10F3D 𐼽 SOGDIAN LETTER AYIN
• used only in Aramaic heterograms
10F3E 𐼾 SOGDIAN LETTER PE
10F3F 𐼿 SOGDIAN LETTER SADHE
10F40 𐽀 SOGDIAN LETTER RESH-AYIN
10F41 𐽁 SOGDIAN LETTER SHIN
10F42 𐽂 SOGDIAN LETTER TAW
10F43 𐽃 SOGDIAN LETTER FETH
10F44 𐽄 SOGDIAN LETTER LESH
= hooked resh

Phonogram

10F45 𐽅 SOGDIAN PHONOGRAM SHIN
→ 6240 所 cjk unified ideograph-6240

Modifier signs

10F46 ◌𐽆 SOGDIAN COMBINING DOT BELOW
10F47 ◌𐽇 SOGDIAN COMBINING TWO DOTS BELOW
10F48 ◌𐽈 SOGDIAN COMBINING DOT ABOVE
10F49 ◌𐽉 SOGDIAN COMBINING TWO DOTS ABOVE
10F4A ◌𐽊 SOGDIAN COMBINING CURVE ABOVE
10F4B ◌𐽋 SOGDIAN COMBINING CURVE BELOW
10F4C ◌𐽌 SOGDIAN COMBINING HOOK ABOVE
10F4D ◌𐽍 SOGDIAN COMBINING HOOK BELOW
10F4E ◌𐽎 SOGDIAN COMBINING LONG HOOK BELOW
10F4F ◌𐽏 SOGDIAN COMBINING RESH BELOW
10F50 ◌𐽐 SOGDIAN COMBINING STROKE BELOW

Numbers

10F51 𐽑 SOGDIAN NUMBER ONE
10F52 𐽒 SOGDIAN NUMBER TEN
10F53 𐽓 SOGDIAN NUMBER TWENTY
10F54 𐽔 SOGDIAN NUMBER ONE HUNDRED

Punctuation

10F55 𐽕 SOGDIAN PUNCTUATION TWO VERTICAL BARS
10F56 𐽖 SOGDIAN PUNCTUATION TWO VERTICAL BARS WITH DOTS
10F57 𐽗 SOGDIAN PUNCTUATION CIRCLE WITH DOT
10F58 𐽘 SOGDIAN PUNCTUATION CIRCLES WITH DOTS
10F59 𐽙 SOGDIAN PUNCTUATION HALF CIRCLE WITH DOT

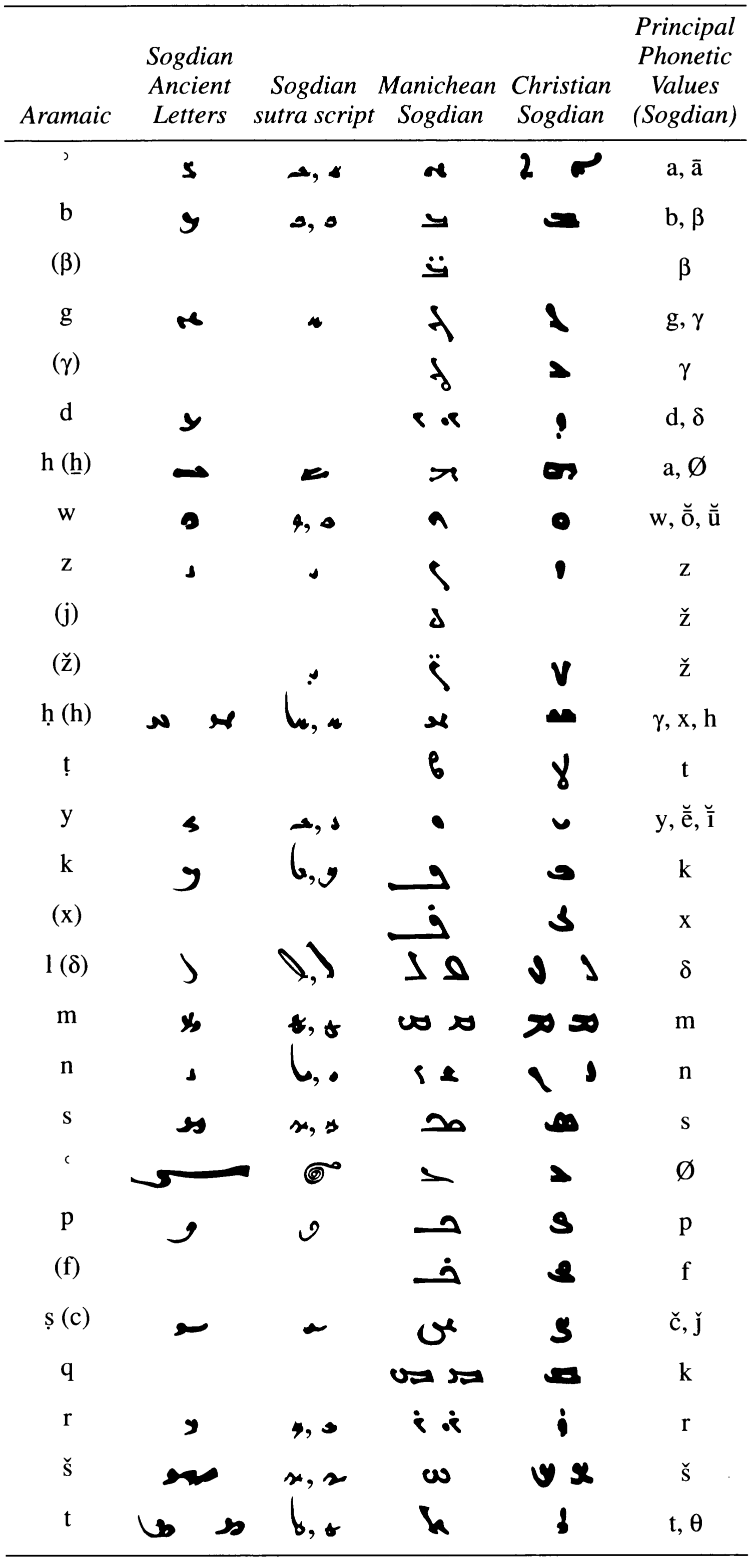
	Old Sogdian	Sogdian
aleph	𐼀, 𐼁	𐼰
beth	𐼂, 𐼃	𐼱
gimel	𐼄	𐼲
daleth	𐼘	—
he	𐼅, 𐼆	𐼳
waw	𐼇	𐼴
zayin	𐼈	𐼵
heth	𐼉	𐼶
teth	—	—
yodh	𐼊	𐼷
kaph	𐼋	𐼸
lamedh	𐼌	𐼹
mem	𐼍	𐼺
nun	𐼎, 𐼏, 𐼐	𐼻
samekh	𐼑	𐼼
ayin	𐼒, 𐼓, 𐼘	𐼽, 𐽀
pe	𐼔	𐼾
sadhe	𐼕, 𐼖, 𐼗	𐼿
qoph	—	—
resh	𐼘	𐽀
shin	𐼙	𐽅, 𐽁
taw	𐼚, 𐼛, 𐼜	𐽂

Table 1: Comparison of letters of ‘Old Sogdian’ and ‘Sogdian’.

	Sogdian	Syriac	Manichaean
aleph	𐼰	ܐ	𐫀
beth	𐼱	ܒ	𐫁
gimel	𐼲	ܓ	𐫃
daleth	—	ܕ	𐫅
he	𐼳	ܗ	𐫆
waw	𐼴	ܘ	𐫇
zayin	𐼵	ܙ	𐫉
heth	𐼶	ܚ	𐫍
teth	—	ܛ	𐫎
yodh	𐼷	ܝ	𐫏
kaph	𐼸	ܟ	𐫐
lamedh	𐼹	ܠ	𐫓
mem	𐼺	ܡ	𐫖
nun	𐼻	ܢ	𐫗
samekh	𐼼	ܣ	𐫘
ayin	𐼽, (𐽀)	ܥ	𐫙
pe	𐼾	ܦ	𐫛
sadhe	𐼿	ܨ	𐫝
qoph	—	ܩ	𐫞
resh	𐽀	ܪ	𐫡
shin	𐽁	ܫ	𐫢
taw	𐽂	ܬ	𐫤

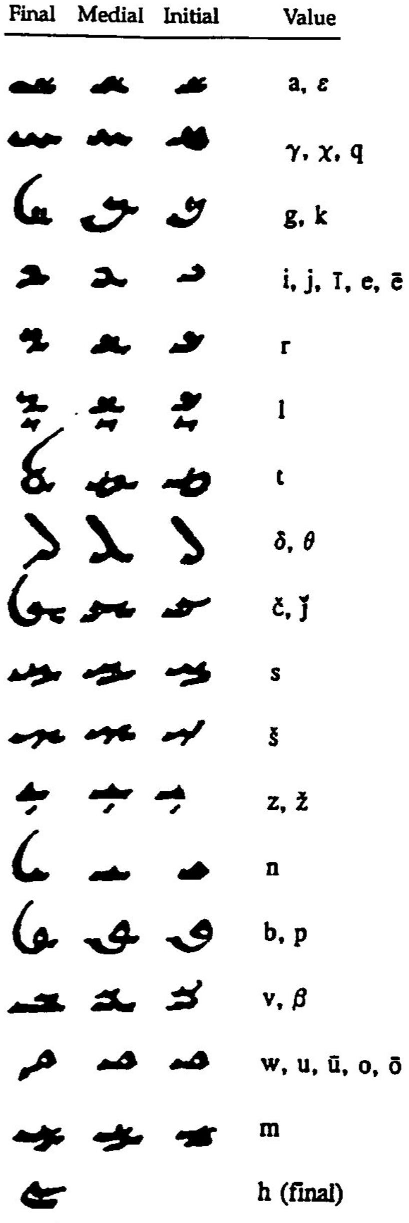
Table 1: Comparison of Sogdian letters with those in Unicode blocks for related scripts. For Sogdian, regular *ayin* is unified with *resh*.

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 (ẖ)					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 1: Table showing various scripts for writing Sogdian (from Skjærvø 1996: 519).



Final	Medial	Initial	Value
			a, ε
			γ, χ, q
			g, k
			i, j, ī, e, ē
			r
			l
			t
			δ, θ
			č, ǰ
			s
			š
			z, ž
			n
			b, p
			v, β
			w, u, ū, o, ō
			m
			h (final)

Table 9 The Sogdian alphabet

Figure 2: Table showing positional forms of Sogdian letters (from Coulmas 1996: 474).

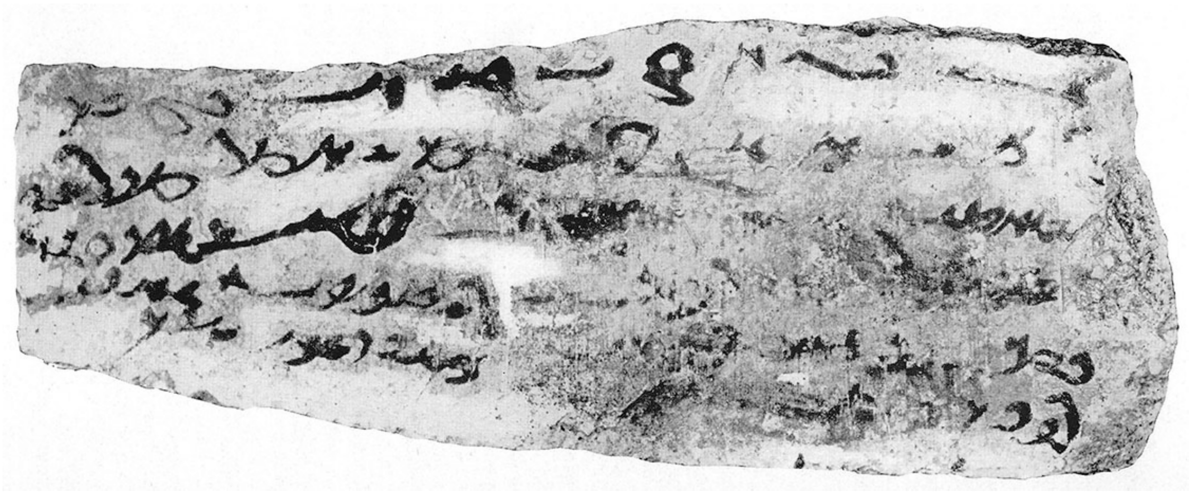


Figure 3: An ostracon from Panjakent, in modern Tajikistan, dated to the end of the 7th or first half of the 8th century bearing an inscription with the letters of the Sogdian script (Livshits 2015: 228). The alphabet appears in lines 1–2 and contains 23 letters: *aleph, beth, gimel, daleth, he, waw, zayin, heth, teth, yodh, kaph // lamedh, mem, nun, samekh, ayin, pe, sadhe, qoph, resh, shin, taw, lamedh*. Livshits notes that "[t]he shapes of the majority of letters in the Penjikent alphabet are the usual ones for 7th to 8th century Sogdian cursive" (*ibid*). The full Aramaic repertoire is given; the *lamedh* occurs twice, in its usual position within the order and at the end. The letters *daleth*, *teth*, *ayin*, *qoph* are represented using signs using various glyphs: number-like signs '20' for *daleth* and '100' for *ayin*; a sign for *teth*, which resembles a form of *ayin*; a digraph ligature '*kaph-resh*' for *qoph*.

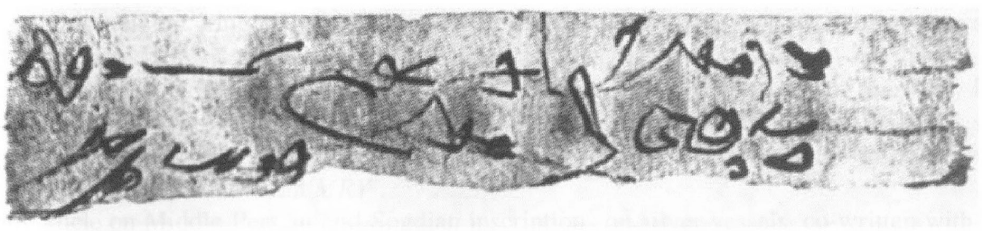
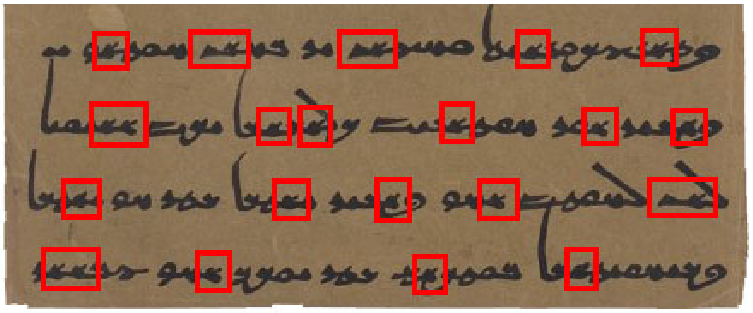
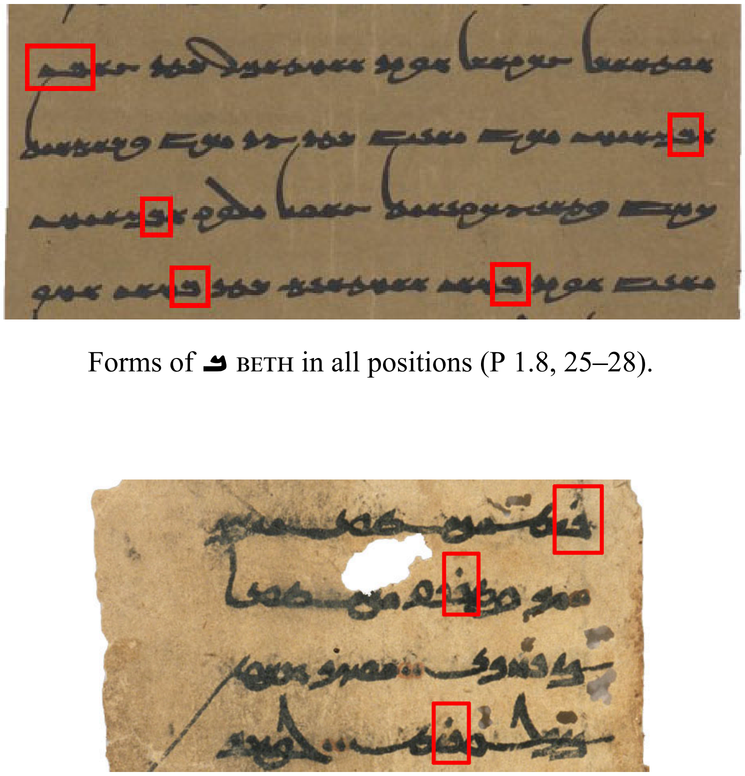


Figure 4: A manuscript fragment in the Otani collection containing the Sogdian repertoire (Livshits 2015: 231). The repertoire is an abbreviated version of the Aramaic original. The order of letters also differs and several letters are repeated: *aleph, beth, gimel, lamedh, mem, ayin, sadhe, resh, pe, lamedh, taw // nun, zayin, teth (?), lamedh, ayin / sadhe (?), taw // waw, yodh, pe, resh, shin, taw*.



Forms of 𐼰 ALEPH in all positions (P 1.3, 29–32).

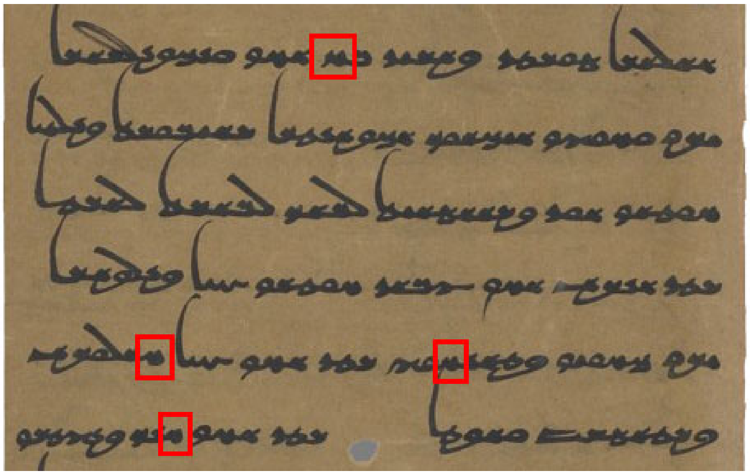
Figure 5: Specimens of *aleph*.



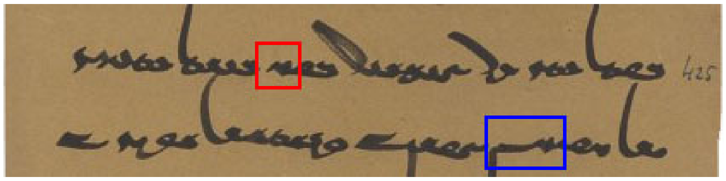
Forms of 𐼱 BETH in all positions (P 1.8, 25–28).

Forms of BETH resembling 𐼷 YODH, distinguished with the sign ◌𐽈 (So 18120 r).

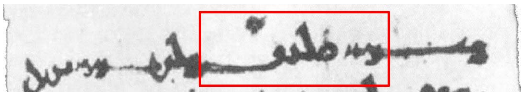
Figure 6: Specimens of *beth*.



Forms of 𐼲 GIMEL in all positions (P 1.9, 7–12).

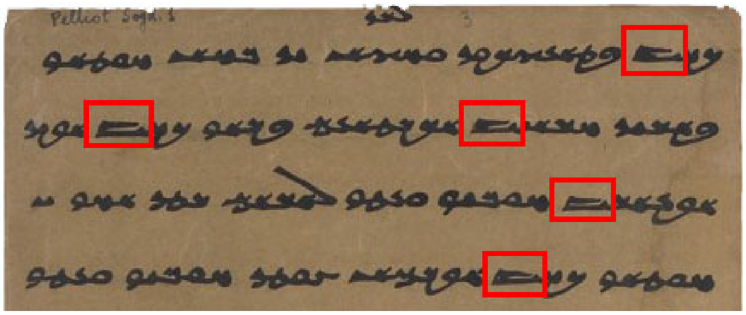


Final form of 𐼲 GIMEL (red) contrasted with variant final form of 𐼶 HETH (blue) (P 2.9v).



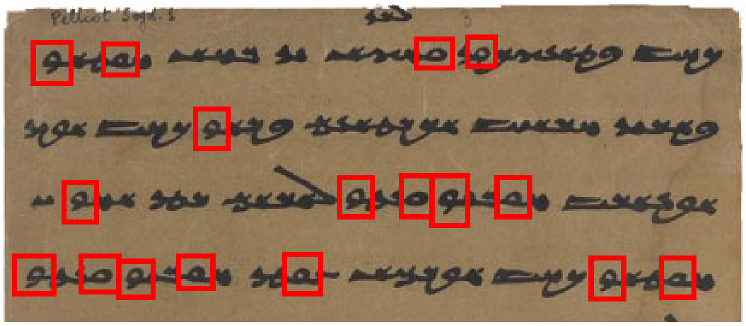
Unjoined medial 𐼲 GIMEL in the word 𐼼𐼲𐼴𐼹𐼷𐼸 *sγ-wδyk* (Mug A-16).

Figure 7: Specimens of *gimel*.



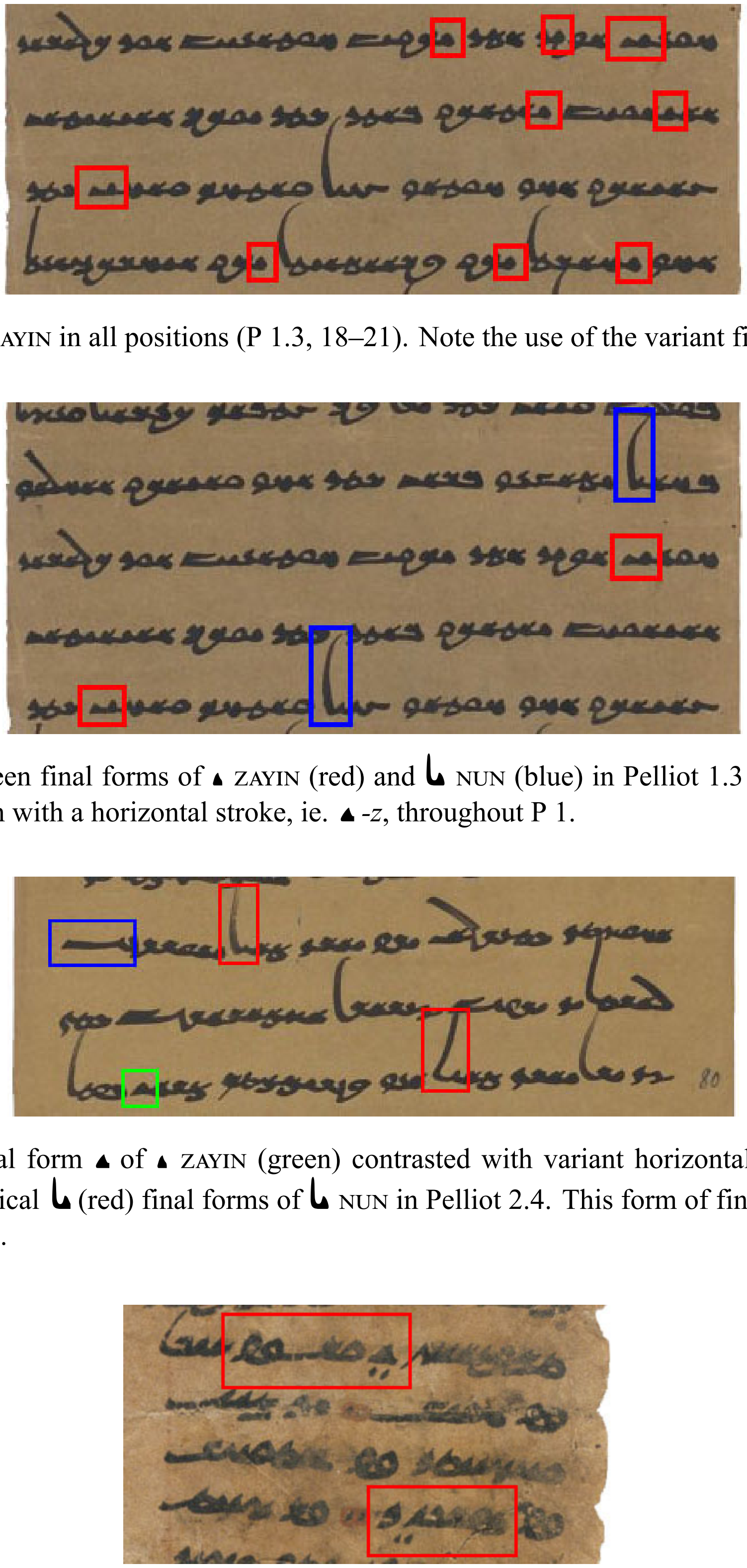
Forms of 𐼳 HE in all positions (P 1.3, 1–4).

Figure 8: Specimens of *he*.



Forms of 𐼴 waw in all positions (P 1.3, 1–4).

Figure 9: Specimens of *waw*.



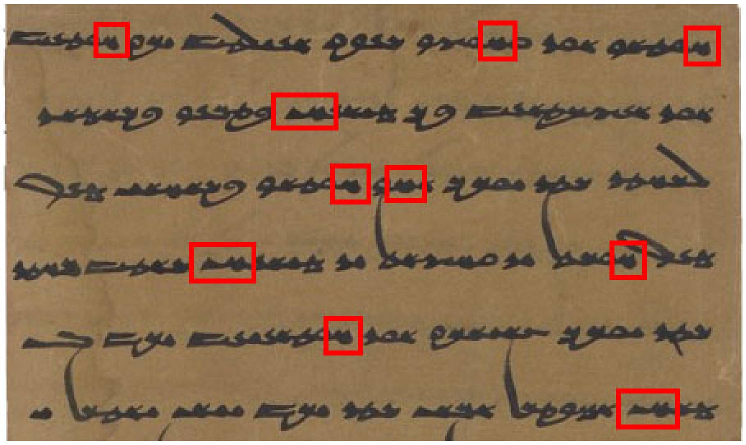
Forms of 𐼵 ZAYIN in all positions (P 1.3, 18–21). Note the use of the variant final form 𐼵.

Contrast between final forms of 𐼵 ZAYIN (red) and 𐼻 NUN (blue) in Pelliot 1.3 (17–20). Final *zayin* is written with a horizontal stroke, ie. 𐼵 -*z*, throughout P 1.

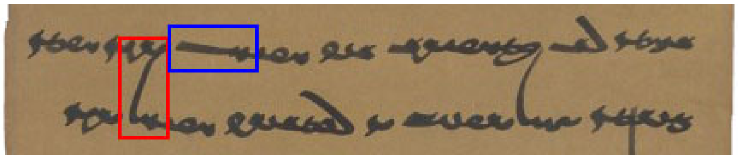
Normative final form 𐼵 of 𐼵 ZAYIN (green) contrasted with variant horizontal 𐼻 (blue) and normative vertical 𐼻 (red) final forms of 𐼻 NUN in Pelliot 2.4. This form of final *zayin* is used throughout P 2.

Unjoined 𐼵 ZAYIN with ◌𐽇 COMBINING TWO DOTS BELOW in initial and medial positions (So 18196 r). The method of representing unjoined ZAYIN is explained in section 4.2.

Figure 10: Specimens of *zayin*.

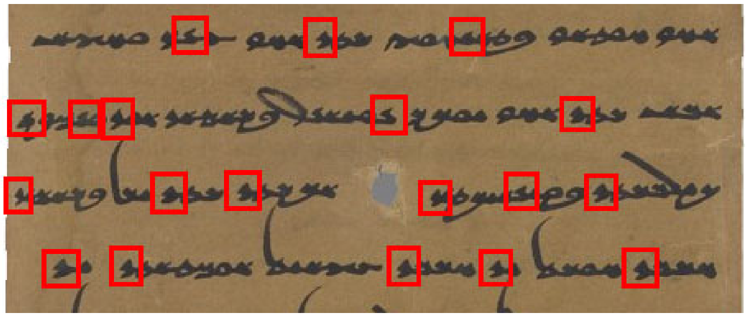


Forms of 𐼶 HETH in all positions (P 1.3v, 1–6). Note the use of the variant final form 𐼶.

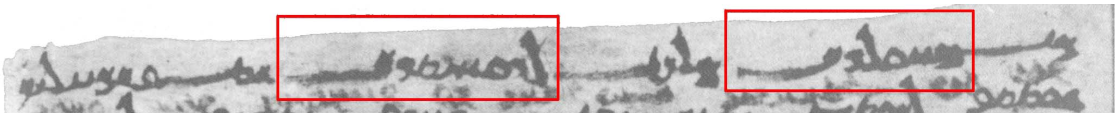


Regular 𐼶 and variant 𐼶 final forms of 𐼶 HETH (P 2.8).

Figure 11: Specimens of *heth*.

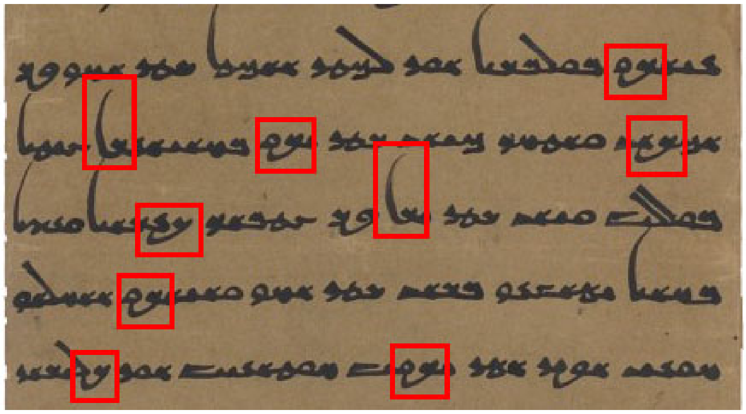


Forms of 𐼷 YODH in all positions (P 1.3v, 9–12).

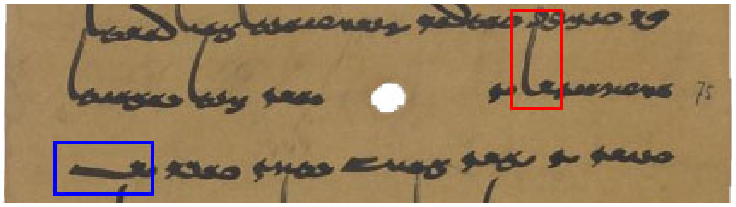


Unjoined 𐼷 YODH in the words 𐼼𐼲𐼴𐼹𐼷‌𐼸 *sγwδy-k* "Sogdian" and 𐼹𐼷‌𐼴𐼰𐽁𐽂𐼷‌𐼿 *δy-w'šty-c* "Đēwāštīc" (Mug A-18, 1). All instances of *yodh* in Mug A-18 are unjoined (see fig. 71).

Figure 12: Specimens of *yodh*.

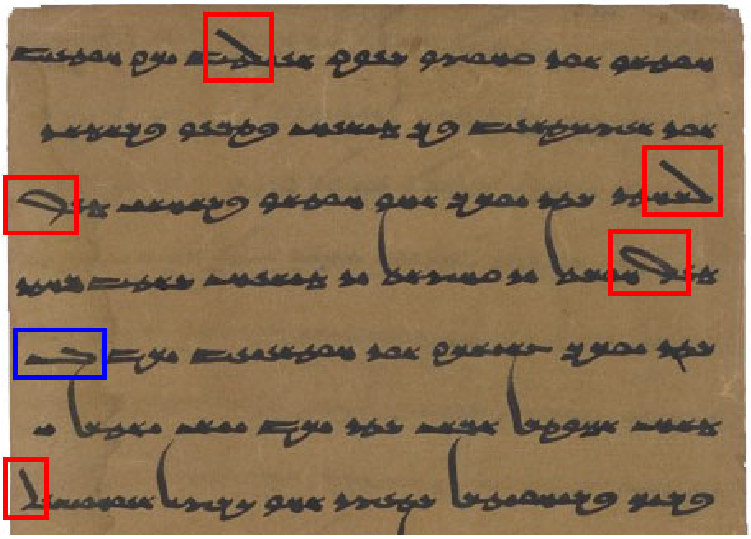


Forms of 𐼸 KAPH in all positions (P 1.3, 15–18).

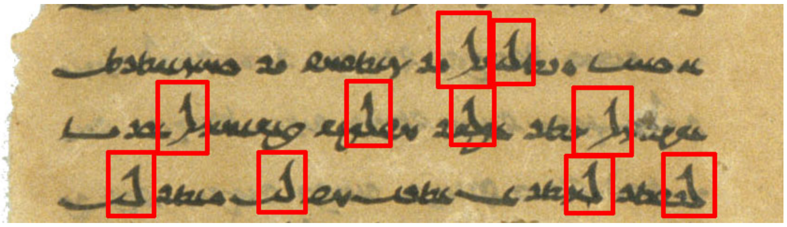


Regular 𐼸 and variant 𐼸 final forms of 𐼸 KAPH (P 2.4).

Figure 13: Specimens of *kaph*.

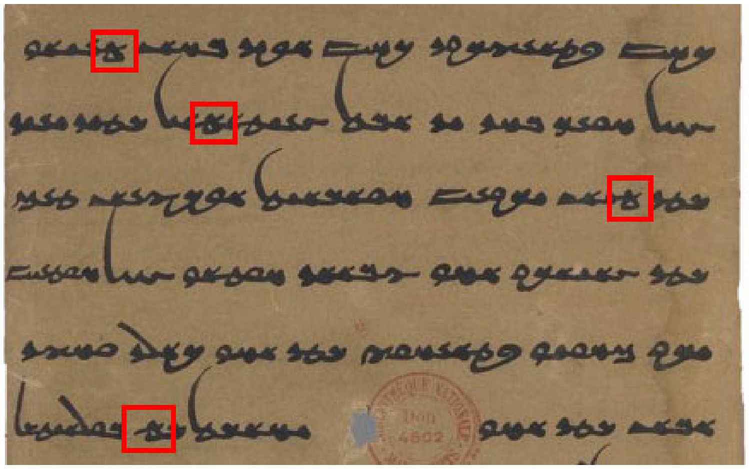


Forms of 𐼹 LAMEDH in all positions (P 1.3v, 1–7). Compare the shape of LAMEDH with 𐽒 TEN (blue). The looped form final form 𐼹 is a stylistic variant.



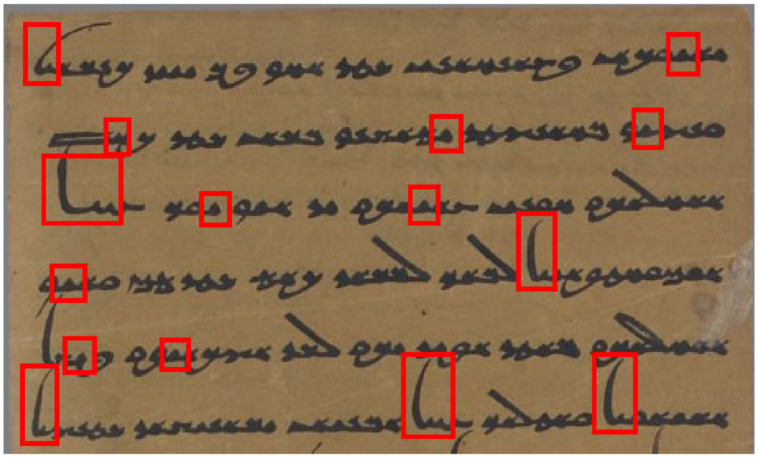
Usage of the 'hooked' form 𐼹 of 𐼹 LAMEDH in all positions (So 18248 r). This form is a glyphic variant that is used most commonly in the 'cursive' style of the script.

Figure 14: Specimens of *lamedh*.

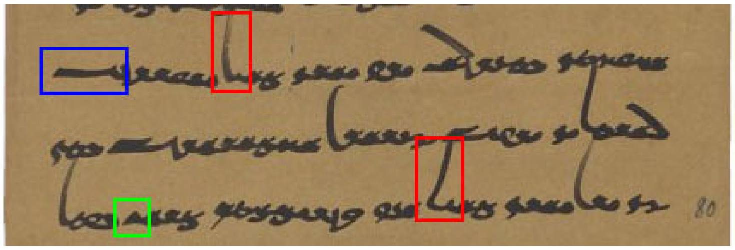


Forms of 𐼺 MEM in all positions (P 1.3, 6–11).

Figure 15: Specimens of *mem*.

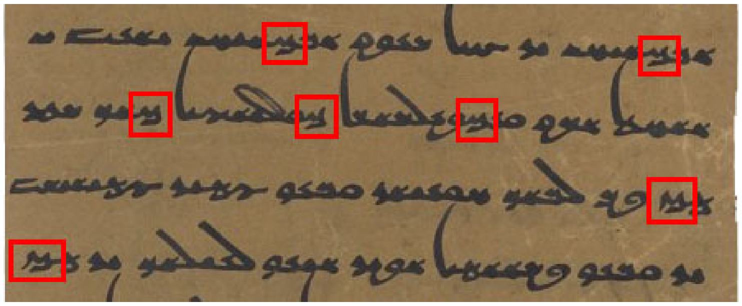


Forms of 𐼻 NUN in all positions (P 1.5v, 1–6).



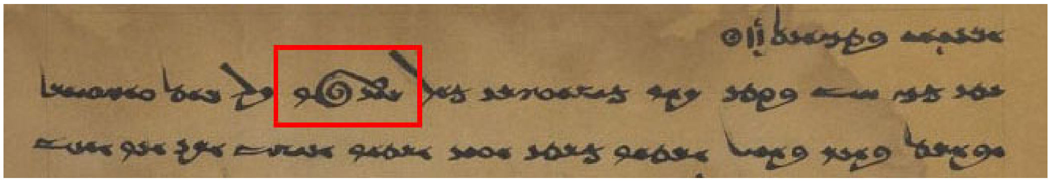
Variant final forms of 𐼻 NUN (red, blue) contrasted with final form of 𐼵 ZAYIN (green) (P 2.4).

Figure 16: Specimens of *nun*.

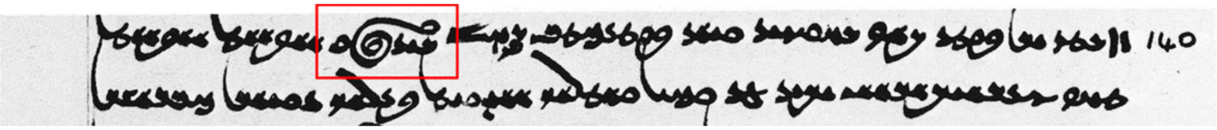


Forms of 𐼼 samekh in all positions (P 1.5v, 22–25).

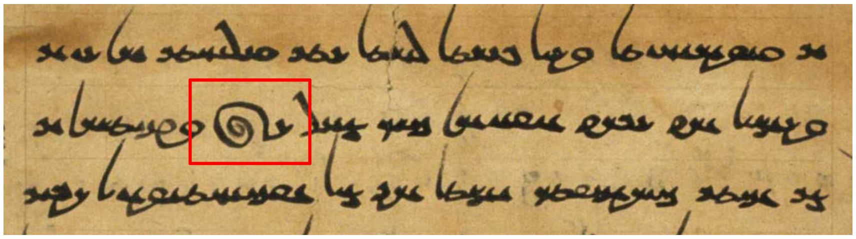
Figure 17: Specimens of *samekh*.



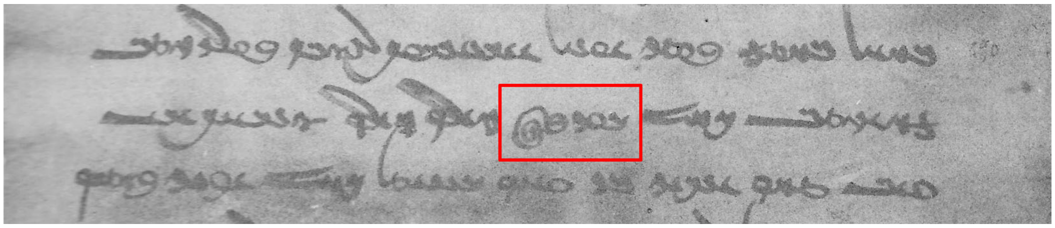
The letter *ayin* represented in the word 𐼾𐼻𐼷𐼾𐼴 *ʿNYʿW*, using 𐼾 RESH-AYIN and 𐼾 AYIN (Pelliot 6).



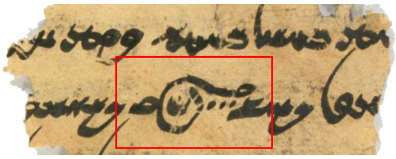
The Aramaic heterogram “said” written as 𐼾𐼻𐼷𐼾𐼴 *ʿNYʿW*, (Dhuta-sutra (Or.8212/160), line 140).



A shortened form of the Aramaic heterogram “said” represented as 𐼾𐼾 (So 20165, lines 8, 11).

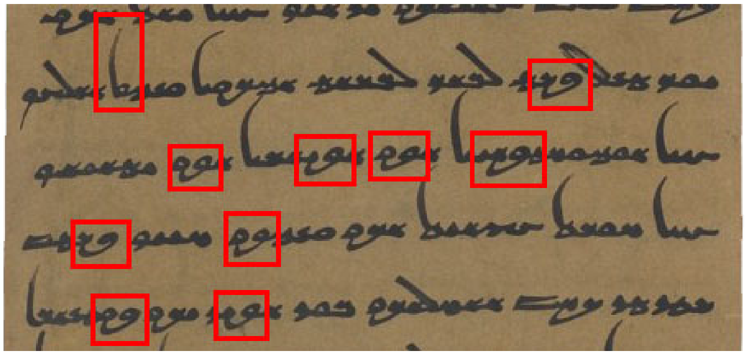


The Aramaic heterogram “said” written as 𐼾𐼻𐼷𐼾 (P7.191).



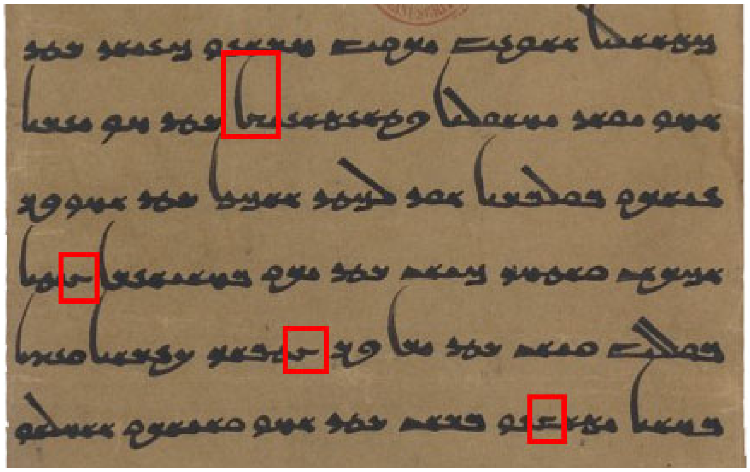
The Aramaic heterogram “said” written as ⸱⸱⸱𐼾𐼴 (So 20241a v).

Figure 18: Specimens of *ayin*. See also § 4.5 and fig. 21.



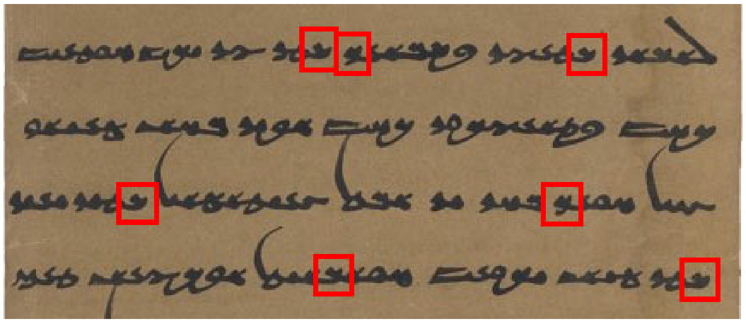
Forms of 𐼾 PE in all positions (P 1.15, 17–20).

Figure 19: Specimens of *pe*.

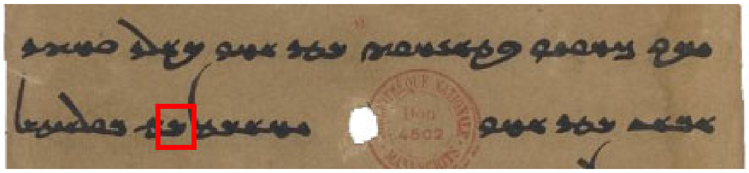


Forms of 𐼿 SADHE in all positions (P 1.3, 12–17).

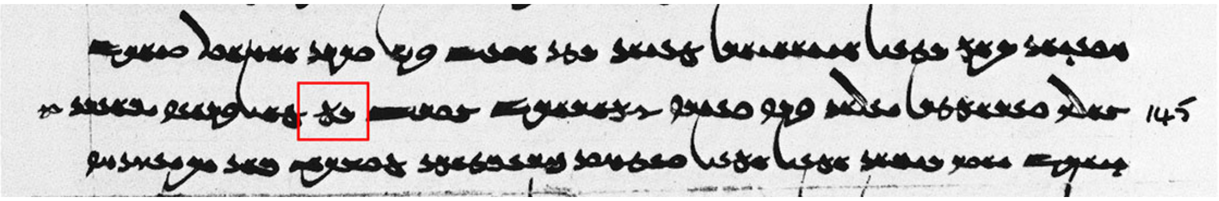
Figure 20: Specimens of *sadhe*.



Forms of 𐽀 RESH-AYIN used for representing *resh* in all positions (P 1.3, 5–8).

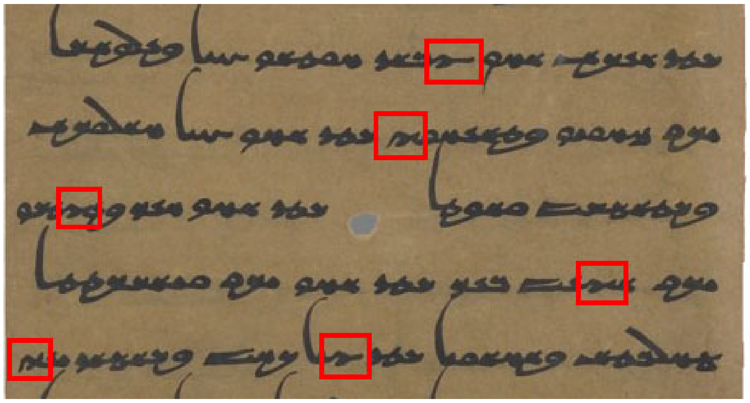


Usage of 𐽀 RESH-AYIN for representing *ayin* in the Aramaic heterogram 𐽀𐼺 *ʻM* ʻwith' (P 1.3, 11).

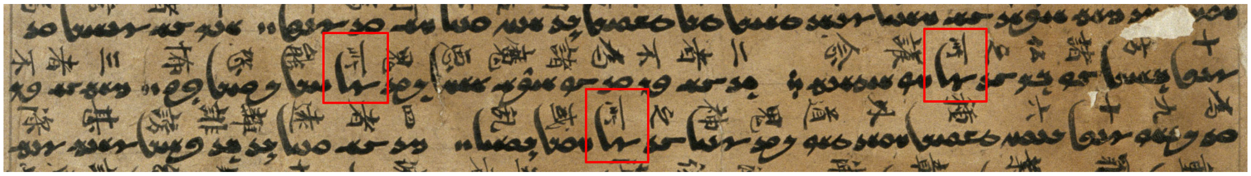


Usage of 𐽀 RESH-AYIN for representing *ayin* in the heterogram 𐽀𐼺 *ʻM* ʻwith' (Dhyana text, line 145).

Figure 21: Specimens of *resh*.

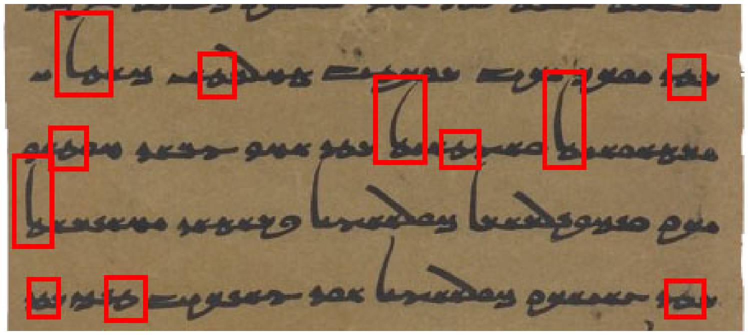


Forms of 𐽁 SHIN in all positions (P 1.9, 10–14).

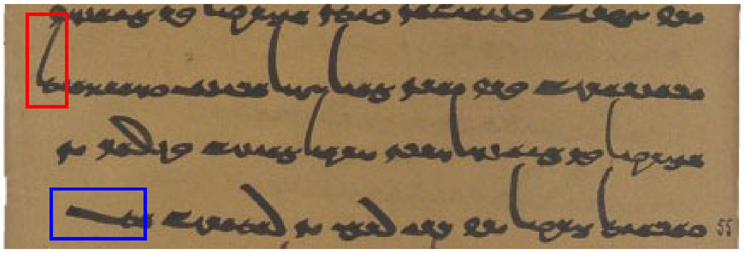


Three instances of 𐽅 PHONOGRAM SHIN used in place of the normative isolated form 𐽁 of SHIN in So 14830. The 𐽅 is used here for transcribing Chinese 所 *suǒ*, and reflects the syntactic function or the enclitic nature of 所, which is a type of relative marker (Yoshida, personal communication, 2016).

Figure 22: Specimens of *shin*.

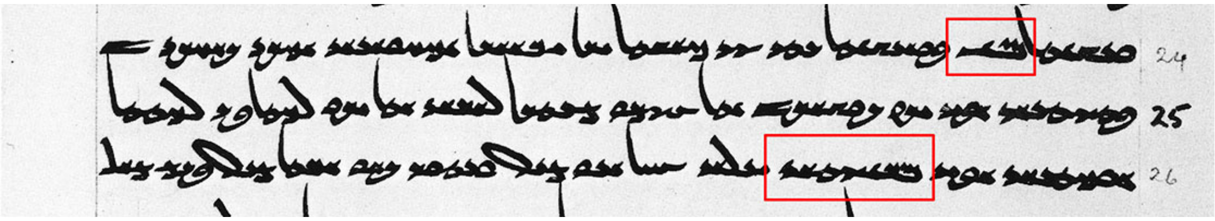


Forms of 𐼹 TAW in all positions (P 1.9, 15–18).



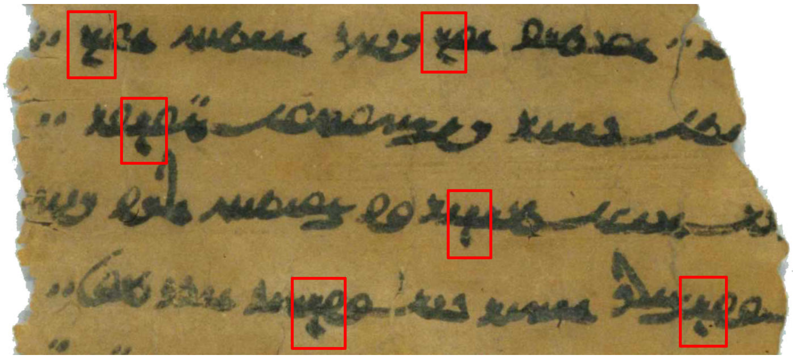
Variant final forms of 𐼹 TAW (P 2.3v).

Figure 23: Specimens of *taw*.

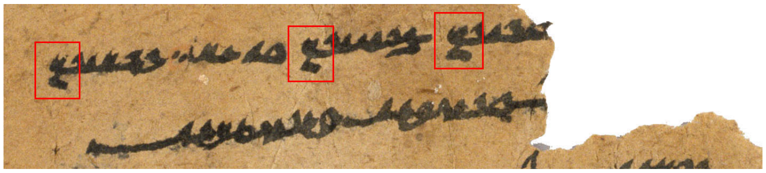


Usage of 𐽃 FETH for representing [f] (Dhyana text, lines 24, 26.).

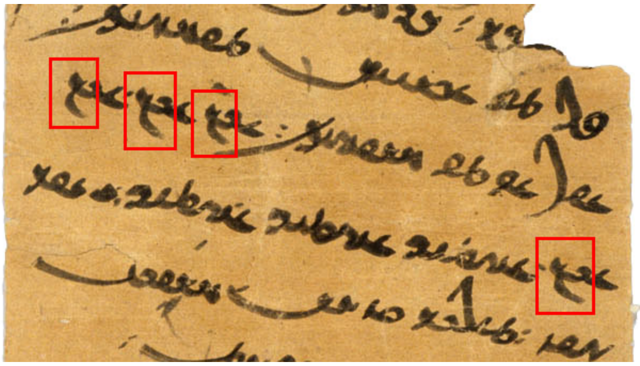
Figure 24: Specimens of *feth*.



Usage of 𐽄 LESH in medial and final positions (So 10026 v).

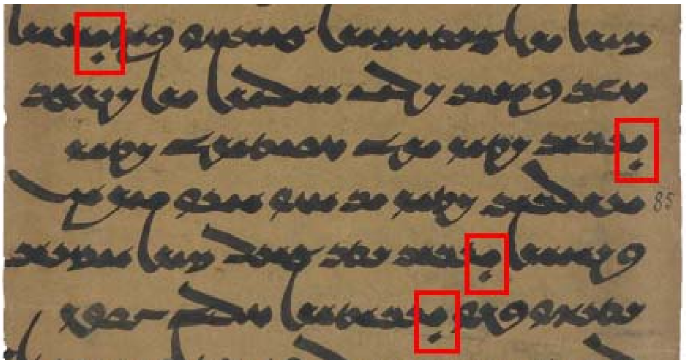


Usage of 𐽄 LESH in final position (So 10678 r).

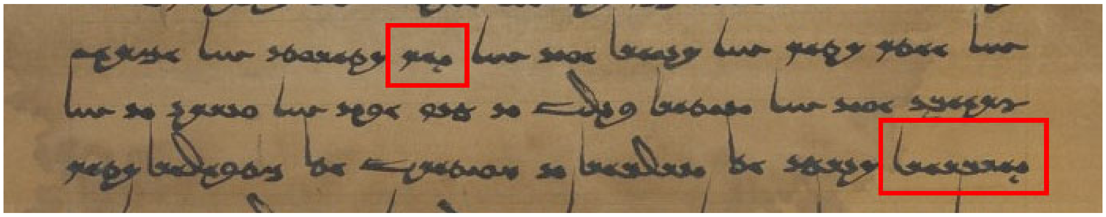


Usage of 𐽄 LESH in final position (Ch/So 20135 v).

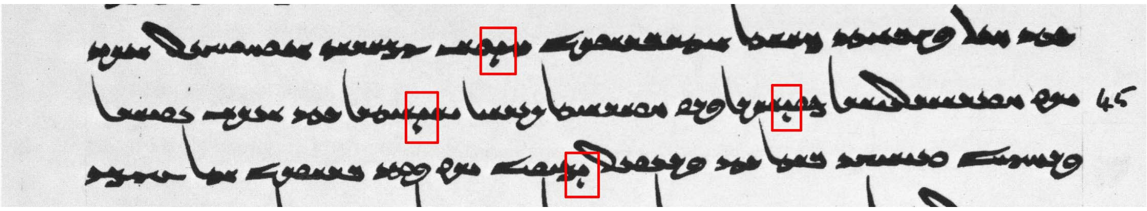
Figure 25: Specimens of *lesh*.



The sign ◌𐽆 used with 𐼵 ZAYIN for transcribing [ž] (P 3.2v).

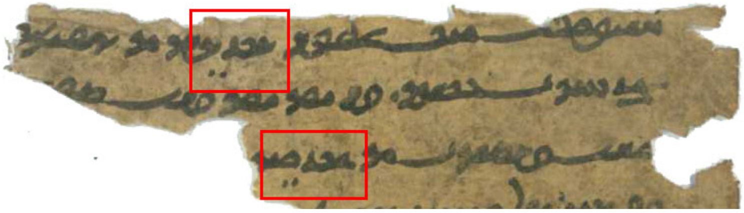


The sign ◌𐽆 may not always be perfectly round or square and may appear as a hook, as is the case here with 𐼵 ZAYIN for transcribing [ž] (P 6).

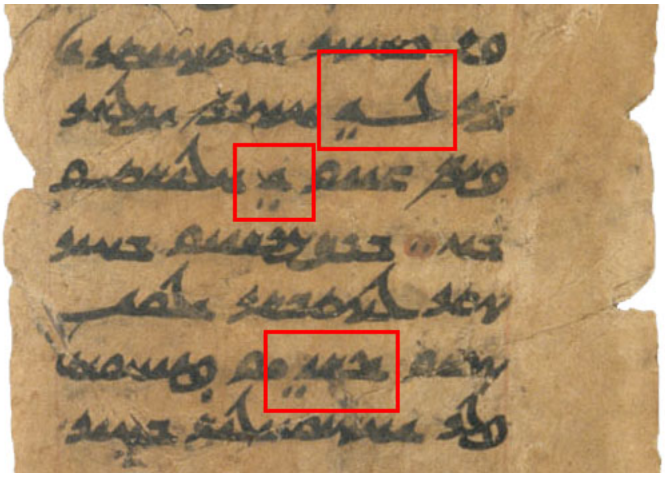


In some cases the sign ◌𐽆 may appear as a large hook or connected to the base letter with an elongated stroke, as shown here with 𐼵 ZAYIN for transcribing [ž] (Dhyana text, lines 44–46). These are to be distinguished from ◌𐽍 SOGDIAN COMBINING HOOK BELOW.

Figure 26: Usage of ◌𐽆 SOGDIAN COMBINING DOT BELOW.

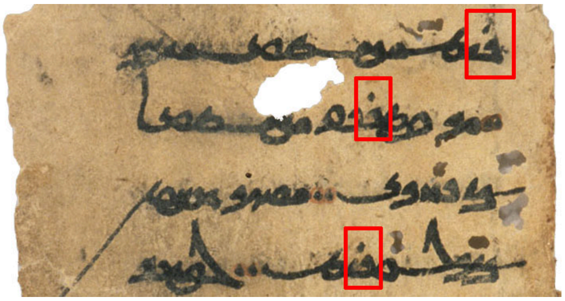


The sign ◌̤ used with 𐼵 ZAYIN for transcribing [z] (So 20226 r).

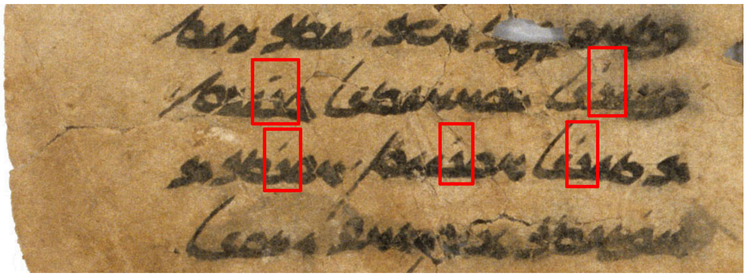


The sign ◌̤ may appear as short oblong strokes, as here with 𐼵 ZAYIN for transcribing [z] (So 18196 v).

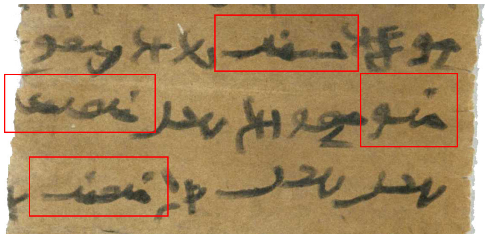
Figure 27: Usage of ◌̤ SOGDIAN COMBINING TWO DOTS BELOW.



The sign ◌𐽆 used with 𐼱 BETH in order to distinguish the letter from 𐼷 YODH (So 18120 r).

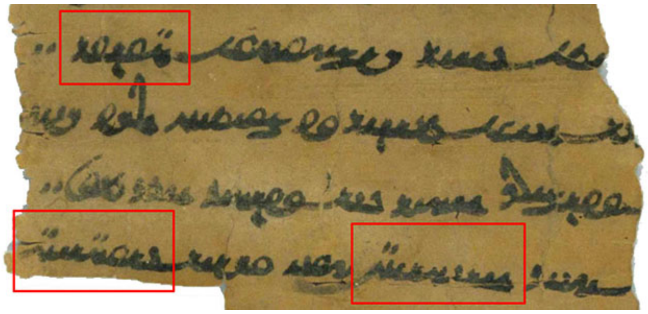


The sign ◌𐽆 used with 𐼱 BETH in order to distinguish the letter from 𐼷 YODH (So 10700b).

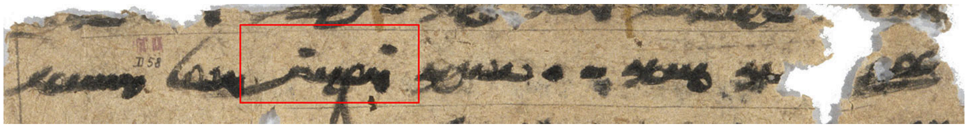


The sign ◌𐽆 used with 𐼶 HETH (So 14800 v).

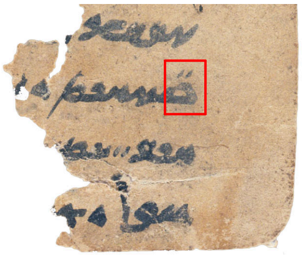
Figure 28: Usage of ◌𐽆 SOGDIAN COMBINING DOT ABOVE.



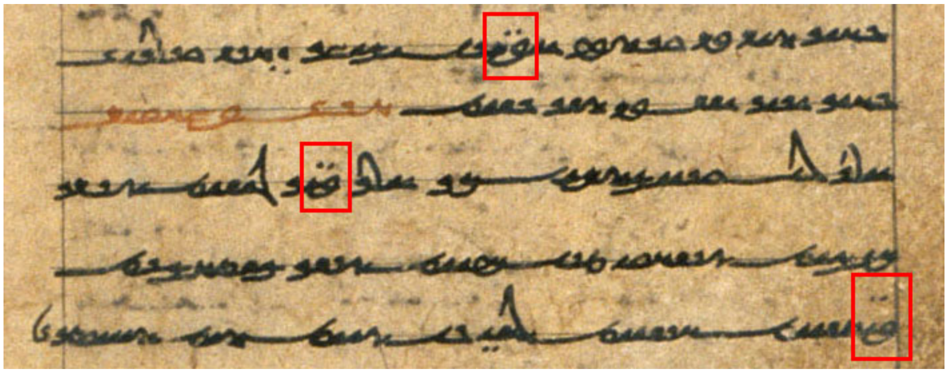
The sign ◌̈ used with 𐼶 HETH (So 10026 v).



The sign ◌̈, rendered as a dash, used with 𐼶 HETH for representing [q] (So 13881/13882 r).

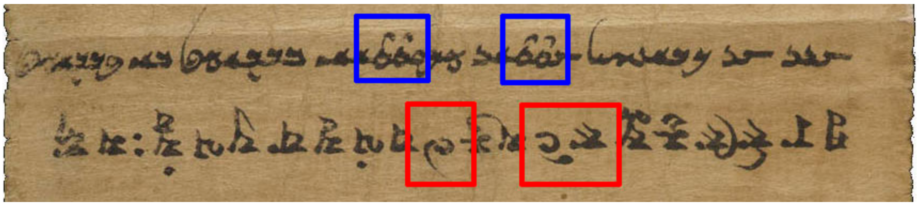


The sign ◌̈ used with 𐼱 BETH in order to distinguish the letter from 𐼷 YODH (So 20193a r).

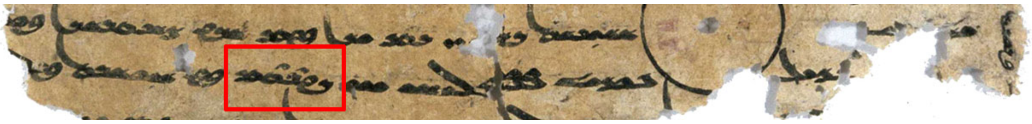


The sign ◌̈ used with 𐼾 PE for indicating [f] (So 14410 r).

Figure 29: Usage of ◌̈ SOGDIAN COMBINING TWO DOTS ABOVE.

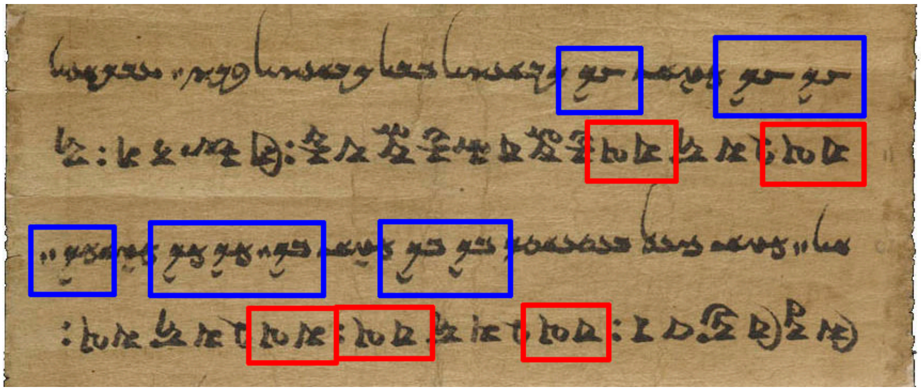


The sign ◌𐽊 used with 𐽂 TAW for transcribing the Sanskrit retroflex consonant [ṭ], eg. 𐼿𐽂𐽊𐽂𐽊𐼰𐼷 *cṭṭʾy* = °जटे *°jaṭe*, 𐼺𐼸𐼴𐽂𐽊𐽂𐽊𐼰𐼰 *mkwṭṭʾʾ* = मकुटा *makuṭā* (BL Or.8212/175). Sogdian text in blue with corresponding Siddham in red.



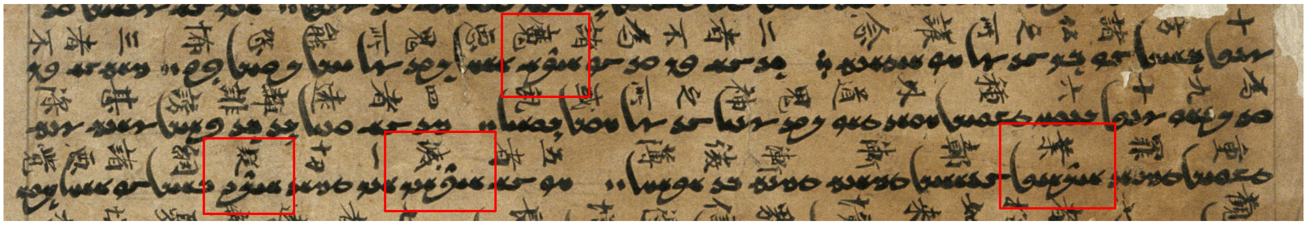
The sign ◌𐽊 used with 𐽂 TAW for transcribing the Sanskrit retroflex consonant [ṭ], eg. 𐼸𐼴𐽀𐽊𐽂𐽊𐽂𐼷 (intended 𐼸𐼴𐽀𐽂𐽊𐽂𐽊𐼷) *kwrṭty* = कोटि *koṭi* (So 14680 r).

Figure 30: Usage of ◌𐽊 SOGDIAN COMBINING CURVE ABOVE.



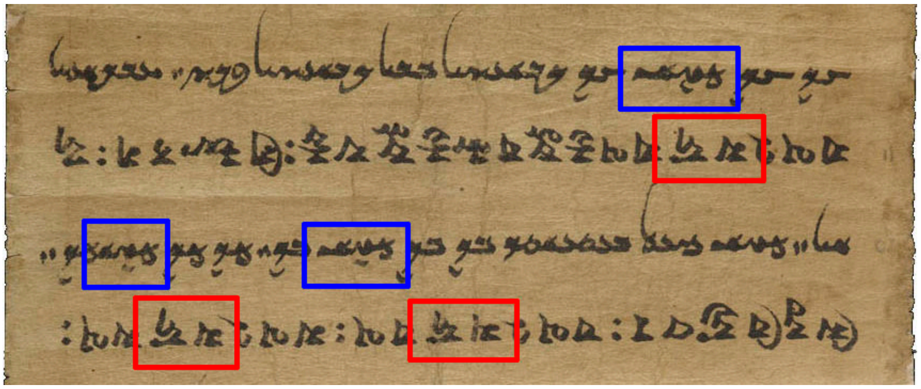
The sign ◌𐽋 used with 𐽀 RESH-AYIN for transcribing Sanskrit [l], eg. 𐼿𐽀𐽋 *cl* = चल *cala*, 𐼱𐽀𐽋 *bl* = बल *bala*, 𐼺𐽀𐽋 *ml* = मल *mala*. Sogdian text in blue with corresponding Siddham in red.

Figure 31: Usage of ◌𐽋 SOGDIAN COMBINING CURVE BELOW.

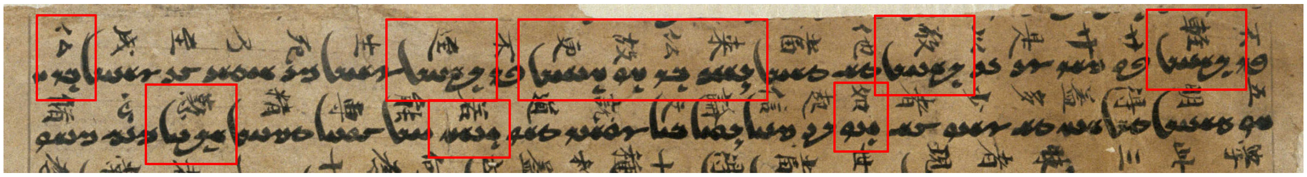


The sign ◌𐽌 used with 𐼸 KAPH and 𐼾 PE for possibly transcribing voiced Chinese consonants, ie. [g], [b].

Figure 32: Usage of ◌𐽌 SOGDIAN COMBINING HOOK ABOVE.

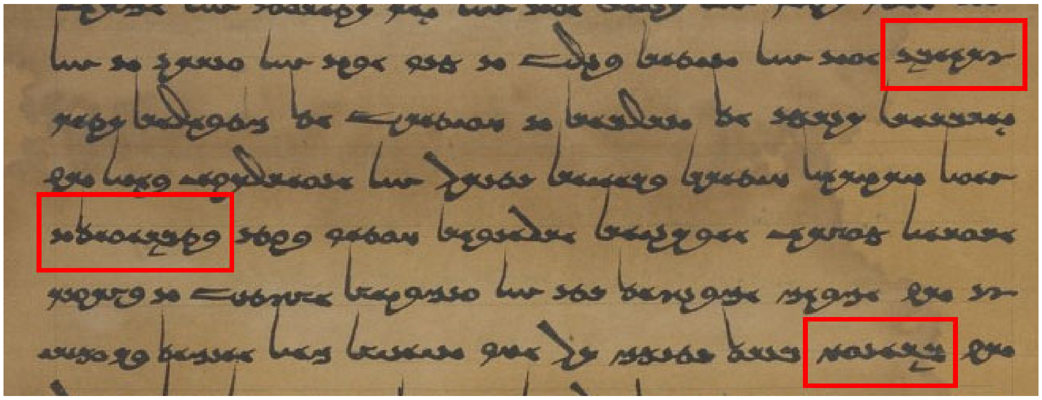


The sign ◌𐽍 used with 𐼲 GIMEL for transcribing Sanskrit ह *ha*, eg. 𐼺𐼲𐽍𐼰 = महा (highlighted red).

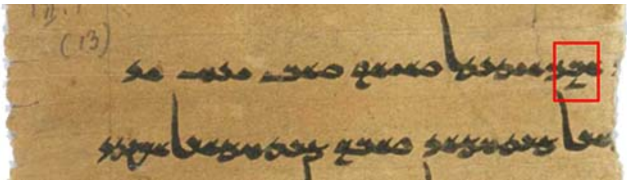


The sign ◌𐽍 used for transcribing Chinese consonants.

Figure 33: Usage of ◌𐽍 SOGDIAN COMBINING HOOK BELOW. See also fig. 34.

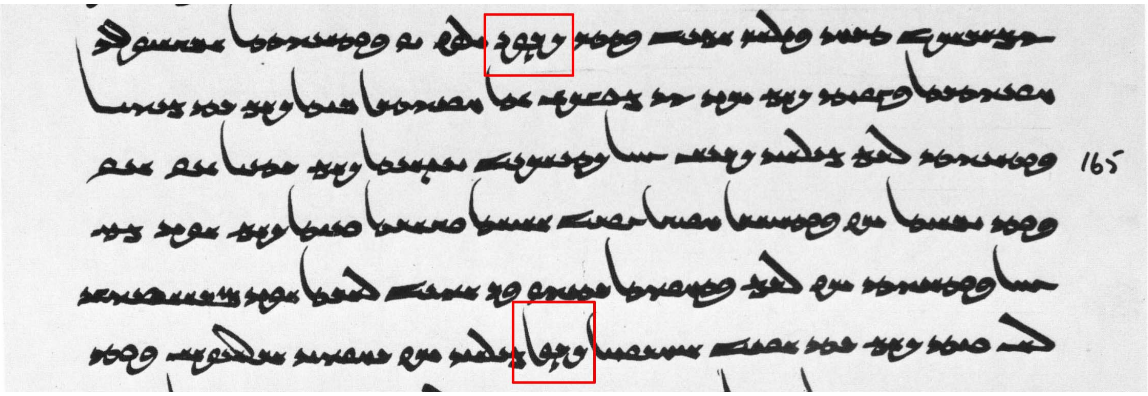


The sign ◌𐽎 used with 𐼱 beth for transcribing [f] (Pelliot 6).

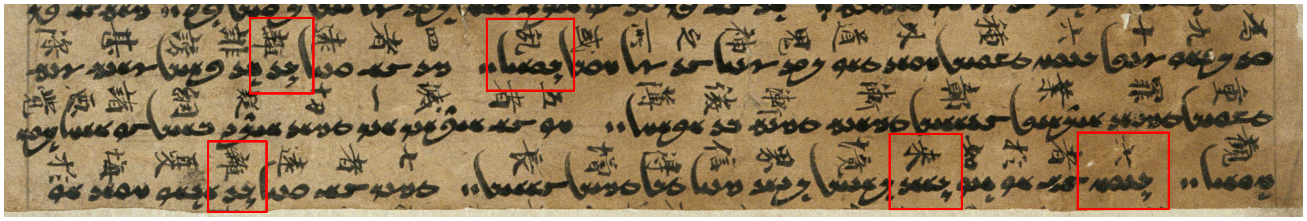


The sign ◌𐽎 used with 𐼱 beth for transcribing [f] (So 14800 v). The hooks accompanying 𐼴 waw and 𐽀 resh-ayin are ornamental strokes added to final forms, not hook diacritics.

Figure 34: Usage of ◌𐽎 sogdian combining long hook below. See also fig. 33.

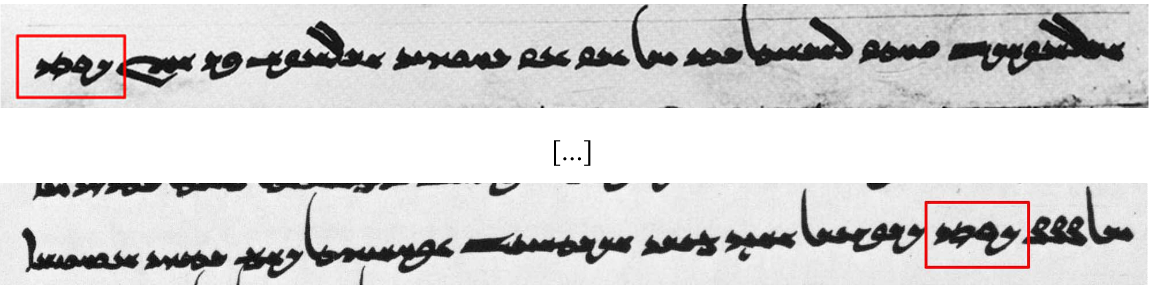


Usage of ◌𐽏 with 𐽀 RESH-AYIN for representing [l], eg. 𐼸𐽀𐽏𐼾𐼷 *klpy* and 𐼸𐽀𐽏𐼾 *klp* (Dhyana text, lines 163, 168). In these cases the sign is connected to the base letter, which is a result of the scribe not lifting the pen for writing the sign.



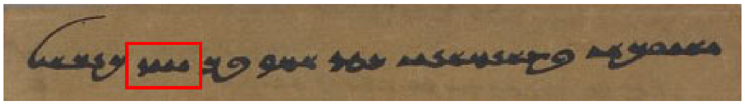
Usage of ◌𐽏 with 𐽀 RESH-AYIN for transcribing Chinese [l].

Figure 35: Usage of ◌𐽏 SOGDIAN COMBINING RESH BELOW.

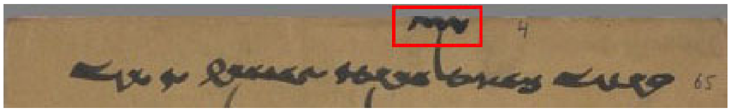


The sign ◌𐽐 used with 𐽂 TAW for representing the Sanskrit retroflex consonant [ṭ] in 𐼸𐼴𐽂𐽐𐼷 *kwṭy* = कोटि *koṭi* (Dhyana text, lines 152, 161). See fig. 30 for a different representation of *koṭi*.

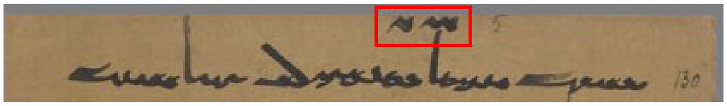
Figure 36: Usage of ◌𐽐 SOGDIAN COMBINING STROKE BELOW.



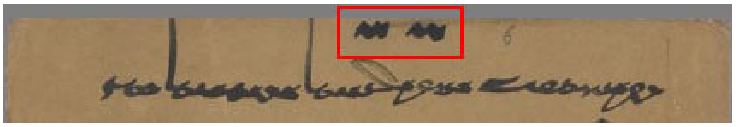
The number 4 𐼠 (Pelliot 1.5v, 1).



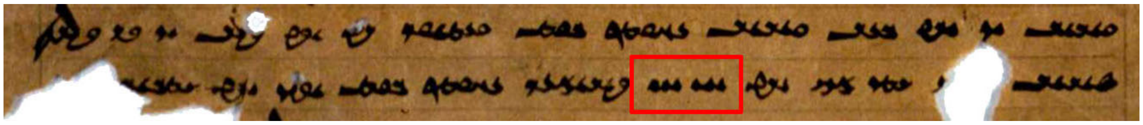
The number 4 𐼠 (Pelliot 2.4).



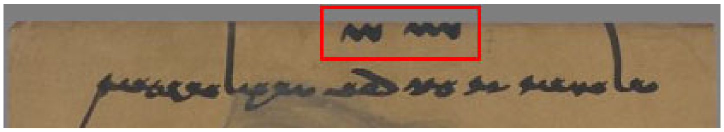
The number 5 𐼟 𐼞 (Pelliot 2.5).



The number 6 𐼟 𐼟 (Pelliot 2.6).

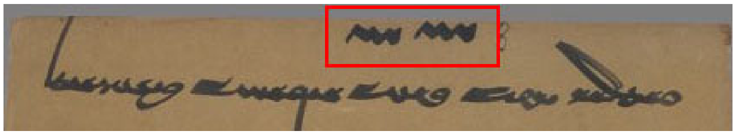


The number 6 𐼟 𐼟 (So 20164 v).

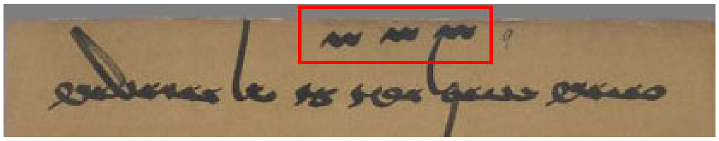


The number 7 𐼠 𐼟 (Pelliot 2.7).

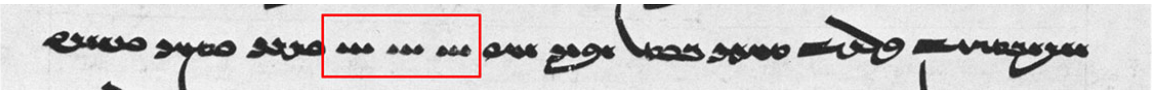
Figure 37: Specimens of numbers (1/6).



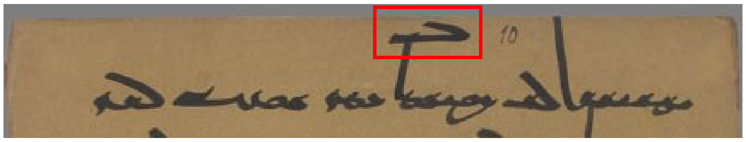
The number 8 𐽑𐽑𐽑𐽑 𐽑𐽑𐽑𐽑 (Pelliot 2.8).



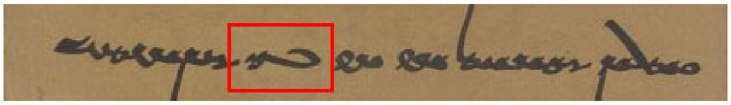
The number 9 𐽑𐽑𐽑 𐽑𐽑𐽑 𐽑𐽑𐽑 (Pelliot 2.9).



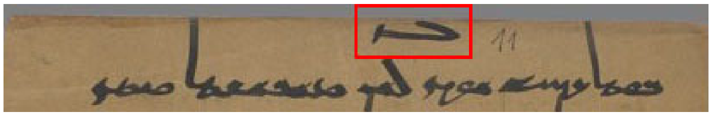
The number 9 𐽑𐽑𐽑 𐽑𐽑𐽑 𐽑𐽑𐽑 (Vimalakīrtinirdeśasūtra, line 17).



The number 10 𐽒 (Pelliot 2.10).

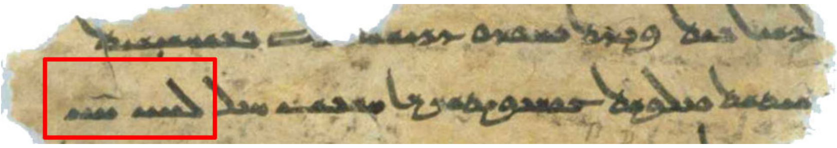


The number 10s 𐽓 (Pelliot 2.3v, 24).

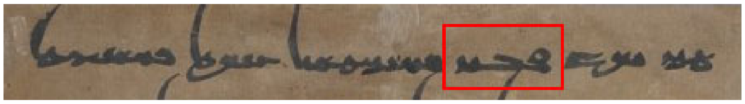


The number 11 𐽒𐽑 (Pelliot 2.11).

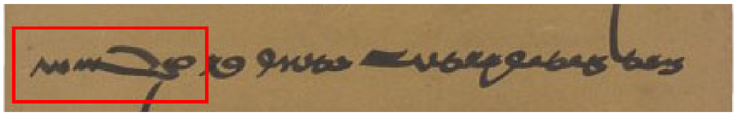
Figure 38: Specimens of numbers (2/6).



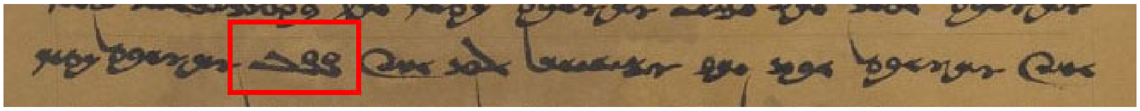
The number 18 𐽑𐽑𐽑 𐽒𐽑𐽑𐽑𐽑𐽑 (So 10921 v).



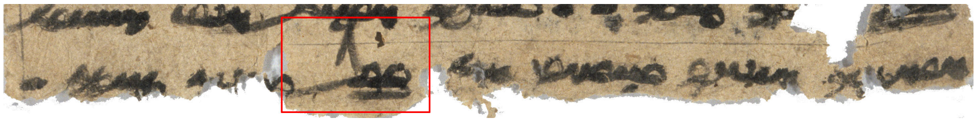
The number 32 𐽓𐽒𐽑𐽑 (Pelliot 1.30v, 1).



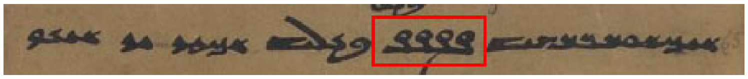
The number 36 𐽑𐽑𐽑 𐽓𐽒𐽑𐽑𐽑 (Pelliot 2.8v).



The number 50 𐽓𐽓𐽒 (P 6).

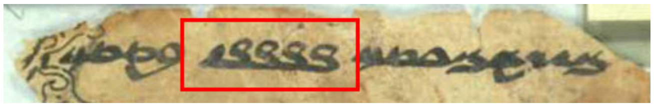


The number 50 𐽓𐽓𐽒 written with the alternate form 𐽓 of 𐽓 (So 13881/13882 r).

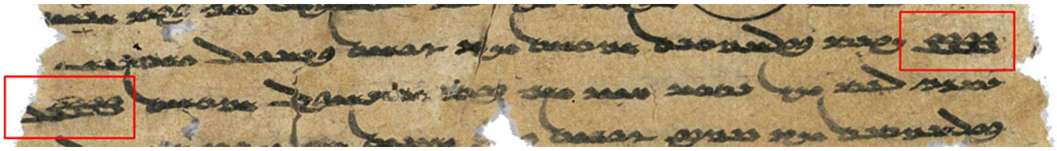


The number 80 𐽓𐽓𐽓𐽓 (Pelliot 1.5, 1).

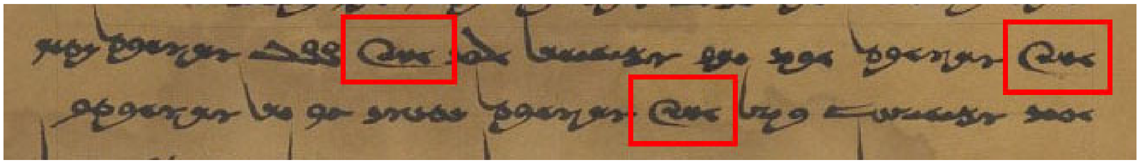
Figure 39: Specimens of numbers (3/6).



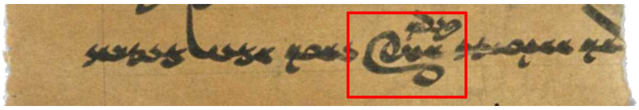
The number 81 𐽓𐽓𐽓𐽓𐽑 (So 18160 verso).



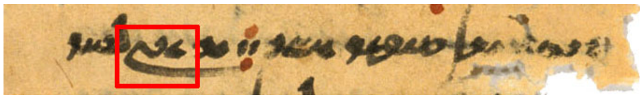
The numbers 80 𐽓𐽓𐽓𐽓 and 90 𐽓𐽓𐽓𐽓𐽒 (?) expressed using the alternate form 𐽓 of 𐽓 (So 14680 v).



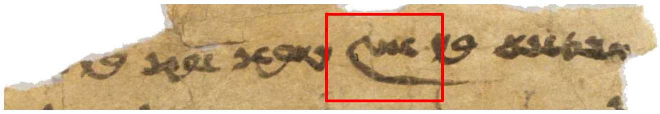
The number 100 𐽔 (P 6).



The number 100 𐽔 written using a stylized form (So 14800 r).

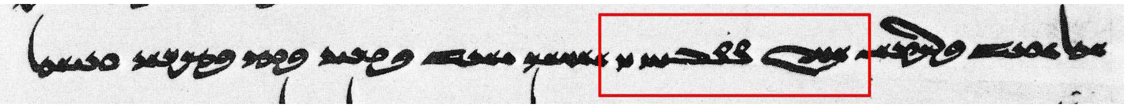


The number 100 written using a variant form 𐽔 of 𐽔 (So 18300 v).

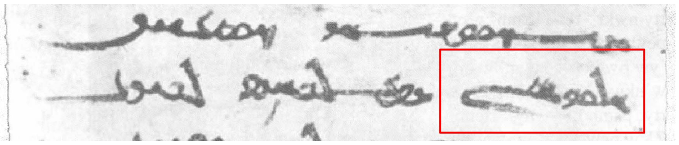


The number 100 𐽔 (So 16122).

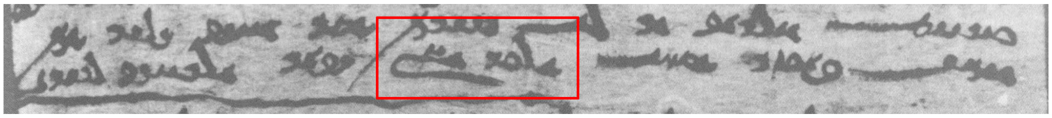
Figure 40: Specimens of numbers (4/6).



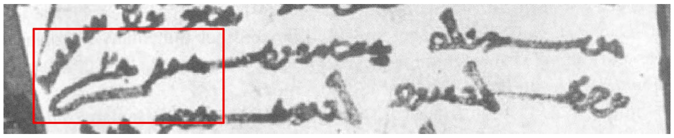
The number 155 𐽔 𐽓𐽓𐽒𐽑𐽑𐽑 𐽑𐽑 (from Dhyana text, line 93).



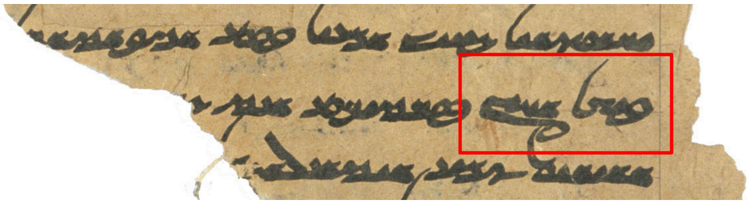
The number 200 𐼰𐼹𐼴 𐽔 written using a variant of 𐽔 (Mug A-3).



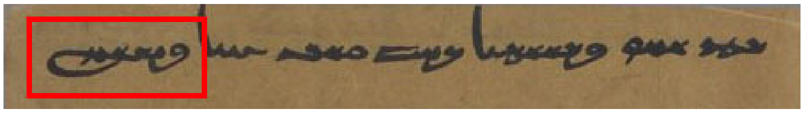
The number 200 𐼰𐼹𐼴 𐽔 written using a variant of 𐽔 (Mug A-18).



The number 300 𐽑𐽑𐽑𐽔 written using a variant of 𐽔 (Mug A-2).

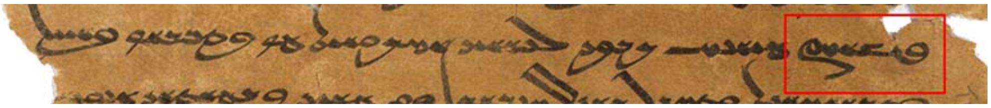


The number 500 𐼾𐼻𐼿 𐽔 (So 14485).

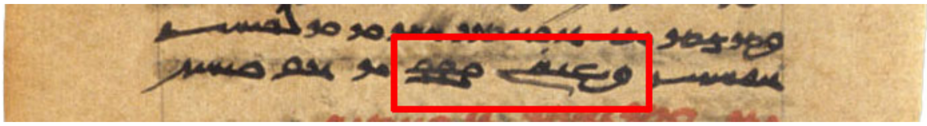


The number 500 𐼾𐼻𐼿𐽔 (P 1.15).

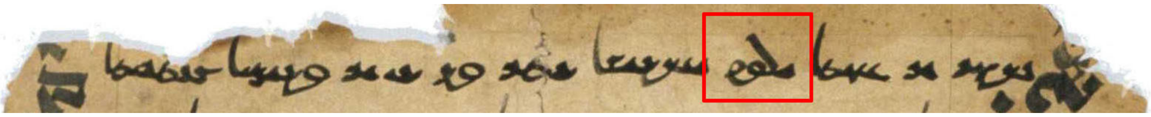
Figure 41: Specimens of numbers (5/6).



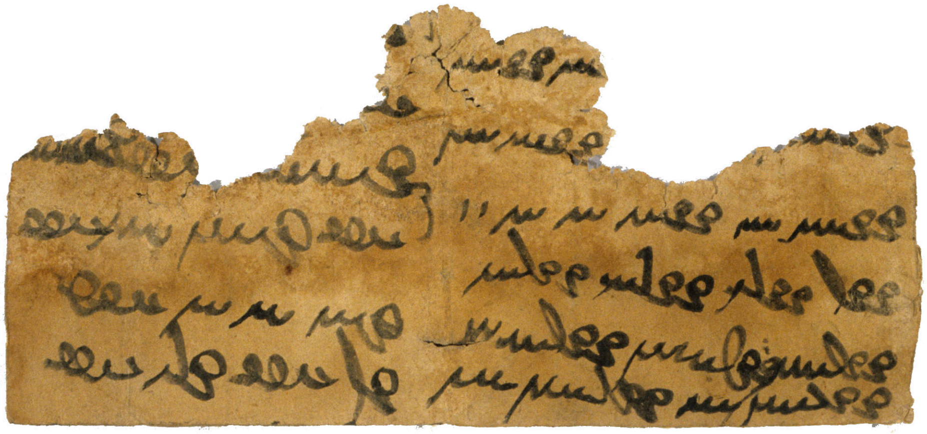
The number 500 𐽑𐽑𐽑𐽑𐽑𐽔 (So 18311 v).



The number 560 𐽑𐽑𐽑𐽑𐽑𐽔 𐽓𐽓𐽓 (So 14570 r).

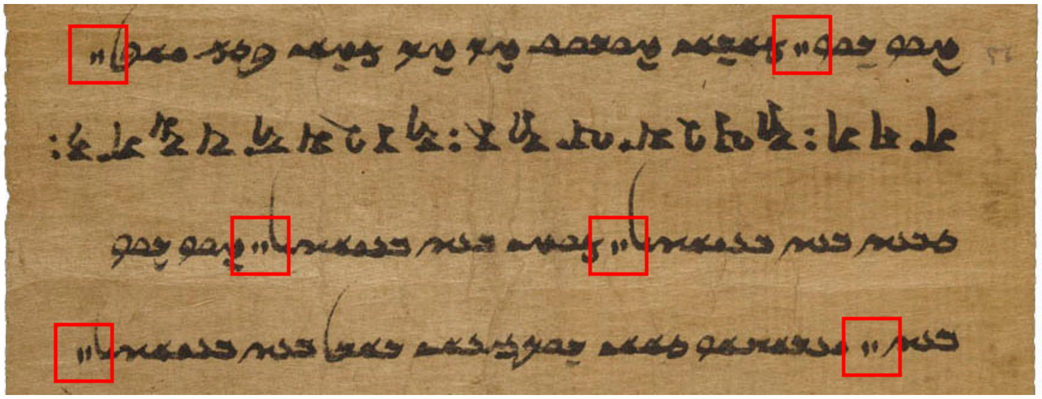


The number 1000 𐽑𐼹𐼾𐼴 *1LPw* in the expression of 8000 𐼰𐽁𐽂 𐽑𐼹𐼾𐼴 (So 20165).

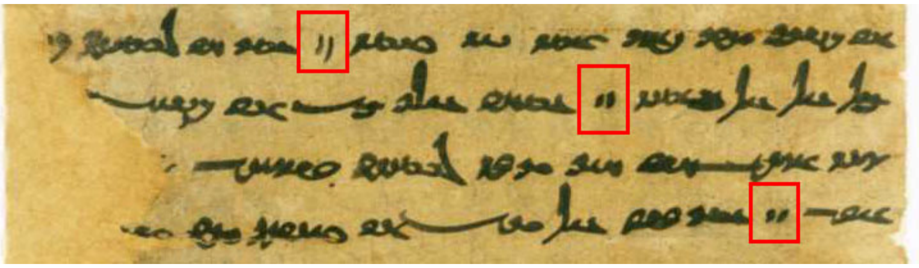


Fragment showing various numbers (Ch/So 20513 v).

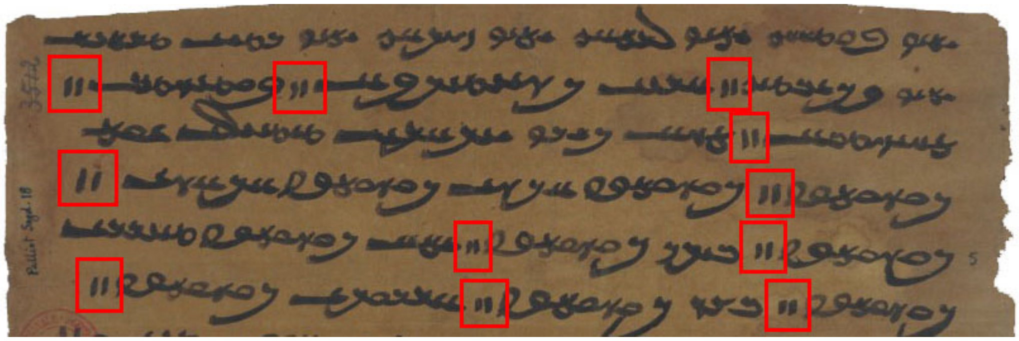
Figure 42: Specimens of numbers (6/6).



𐽕 PUNCTUATION TWO VERTICAL BARS (BL Or. 8212/174)

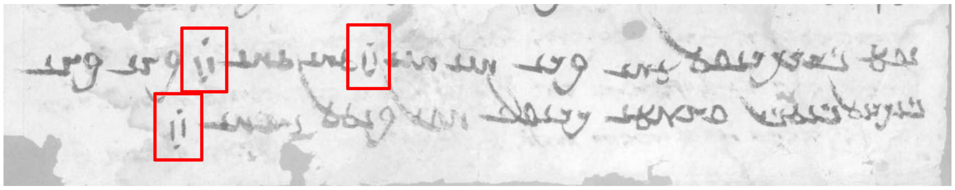


𐽕 PUNCTUATION TWO VERTICAL BARS (Ch/So 20182 v).

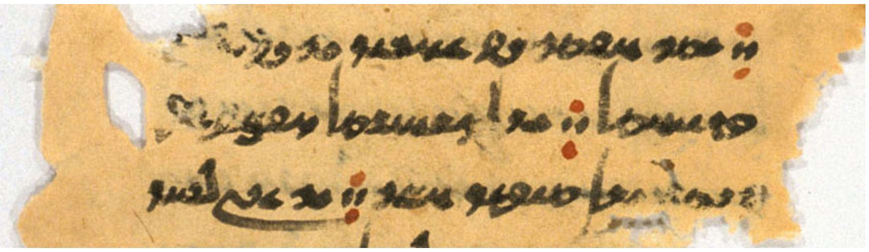


𐽕 PUNCTUATION TWO VERTICAL BARS (P 18).

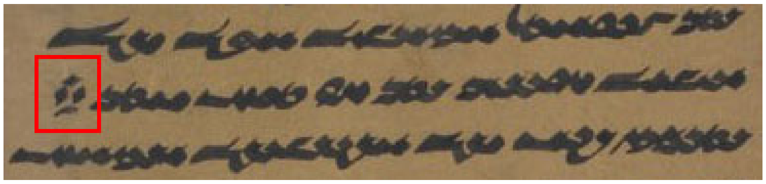
Figure 43: Specimens of punctuation signs proposed for encoding.



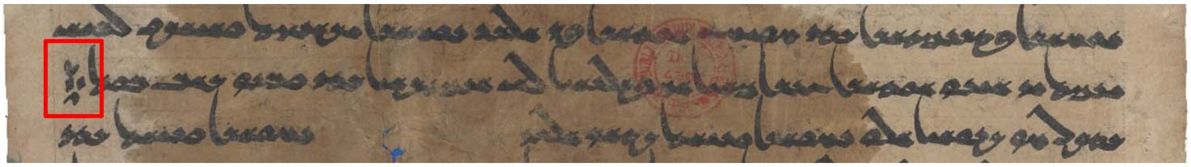
𐽖 punctuation two vertical bars with dots (So 10006 v)



𐽖 punctuation two vertical bars with dots with dots in red ink (So 18300 v).

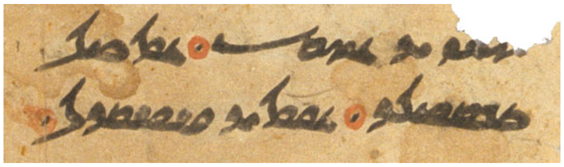


Variant form 𐽖 of 𐽖 punctuation two vertical bars with dots (P 3).

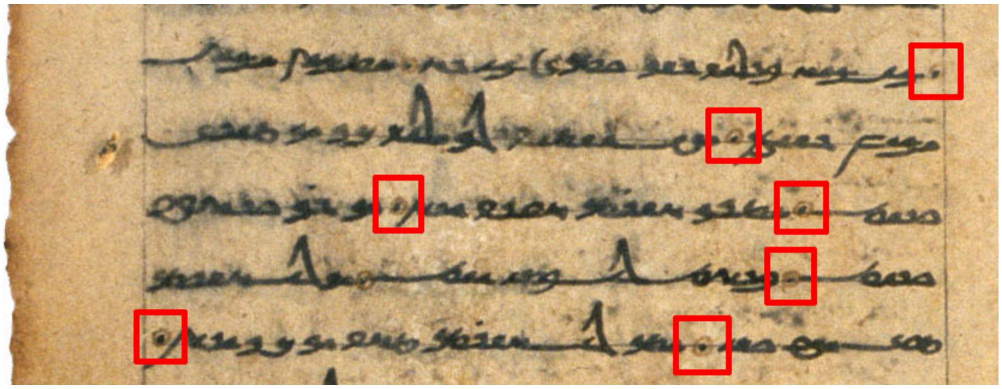


Variant form 𐽖 of 𐽖 punctuation two vertical bars with dots (P 22).

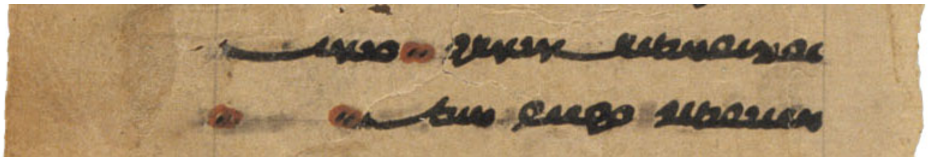
Figure 44: Specimens of punctuation signs proposed for encoding.



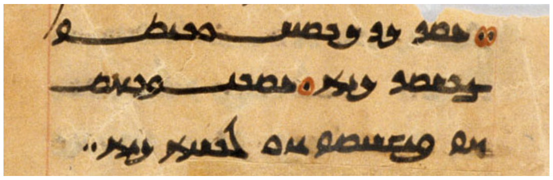
𐽗 PUNCTUATION CIRCLE WITH DOT with circle in red ink (Ch/So 20208).



𐽗 PUNCTUATION CIRCLE WITH DOT with circle in red ink (So 14410).

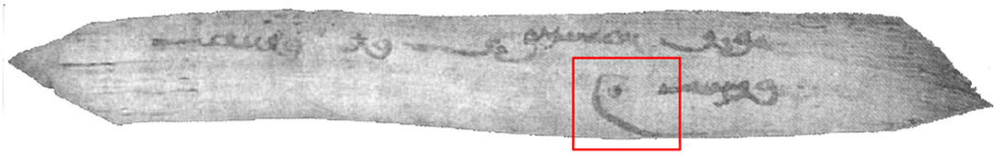


𐽘 PUNCTUATION TWO CIRCLES WITH DOTS with circles in red ink (So 14615 r).

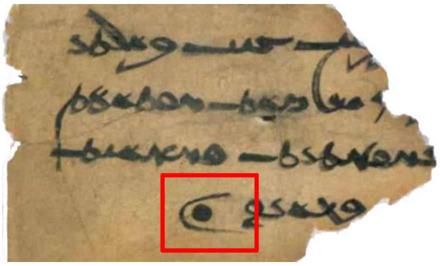


𐽗 PUNCTUATION CIRCLE WITH DOT and 𐽘 PUNCTUATION TWO CIRCLES WITH DOTS with circles in red ink, as well as 𐽕 PUNCTUATION TWO VERTICAL BARS (So 10100(e) r).

Figure 45: Specimens of punctuation signs proposed for encoding.

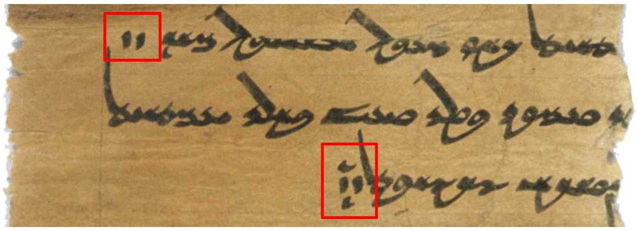


𐽙 PUNCTUATION HALF CIRCLE WITH DOT (Sogdian document no. 1 from Chilkhujra).

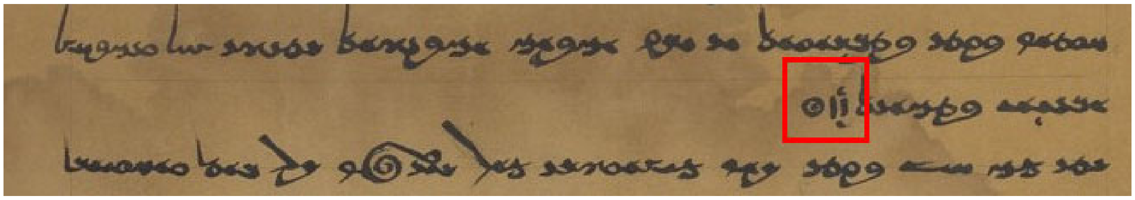


𐽙 PUNCTUATION HALF CIRCLE WITH DOT (So 14700 (16a) r).

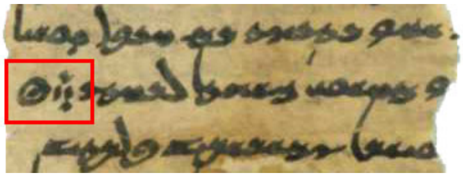
Figure 46: Specimens of punctuation marks proposed for encoding.



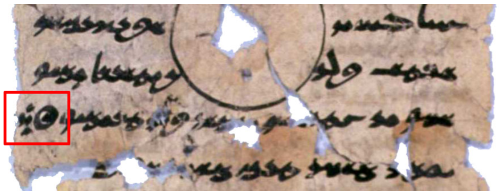
𐽕 PUNCTUATION TWO VERTICAL BARS and 𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS (So 14800 r).



𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS and 𐽗 PUNCTUATION CIRCLE WITH DOT (P 6, line 17).

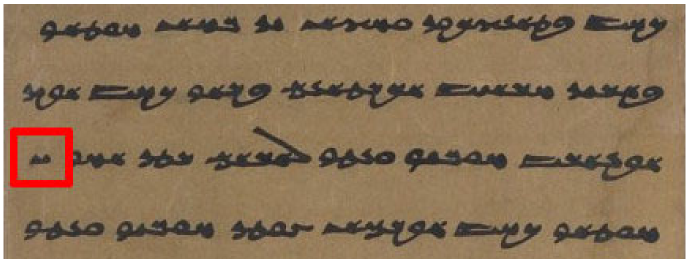


𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS and 𐽗 PUNCTUATION CIRCLE WITH DOT (So 10100 (i) v).

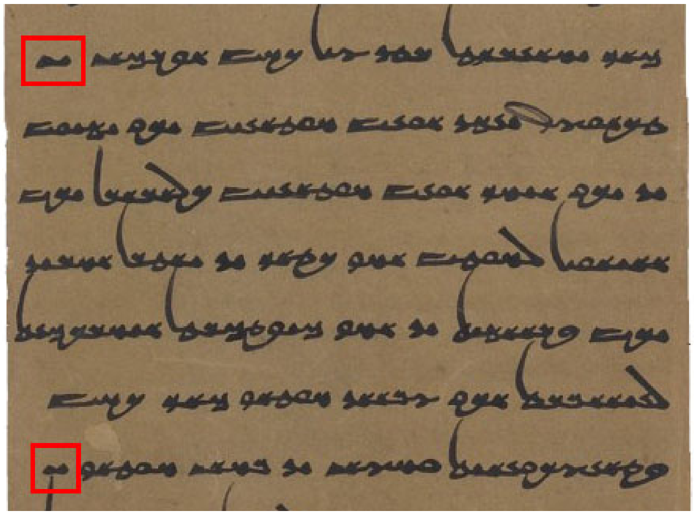


𐽗 PUNCTUATION CIRCLE WITH DOT and 𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS (So 18400 v).

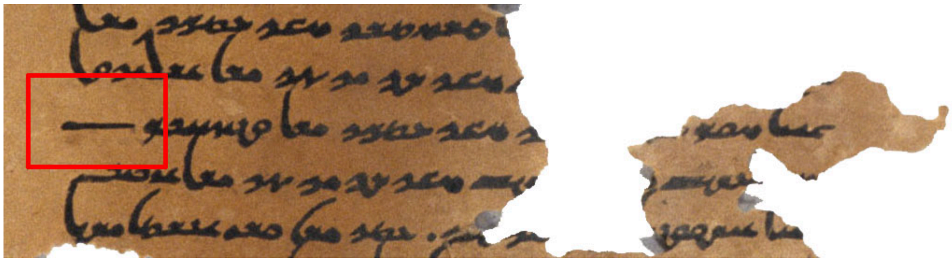
Figure 47: Specimens of punctuation marks proposed for encoding.



variant of LINE FILLER (P 1.3, line 3).

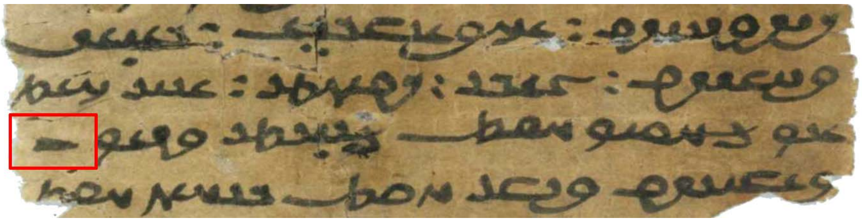


LINE FILLER (P 1.3, lines 23, 29).

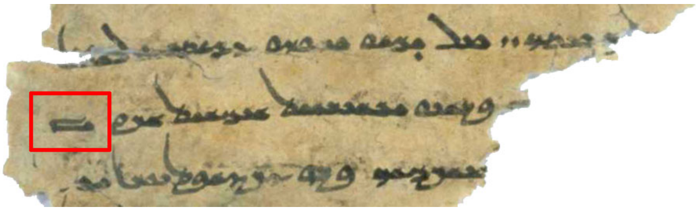


variant of LINE FILLER (So 18220 r).

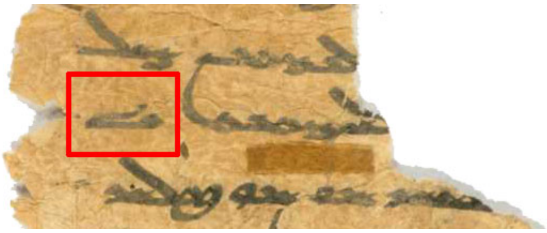
Figure 48: Additional punctuation, not proposed for encoding at present.



variant of LINE FILLER (So 14842a-f v).

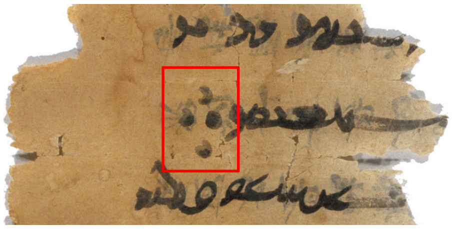


he-like variant of LINE FILLER (So 10921 v).

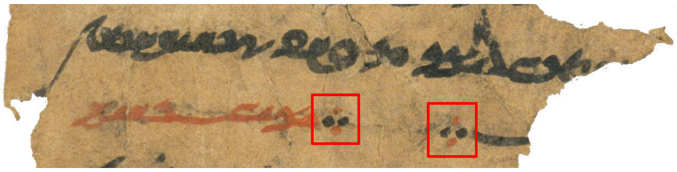


he-like variant of LINE FILLER (So 20152 r).

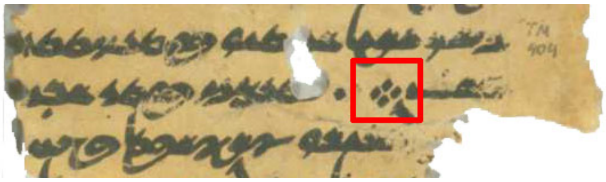
Figure 49: Additional punctuation, not proposed for encoding at present.



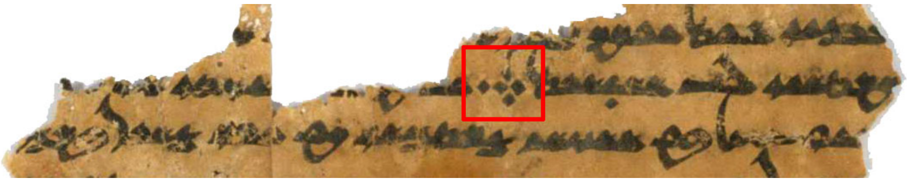
❖ ‘four dots’ (So 20150 v).



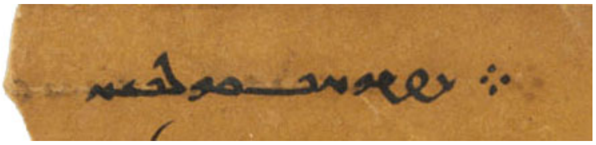
❖ ‘four dots’ with the above and below dots colored in red ink (So 10650(15) v).



❖ ‘four dots’ (So 18273 v). A ⬩ ‘dot’ is also present in the excerpt.

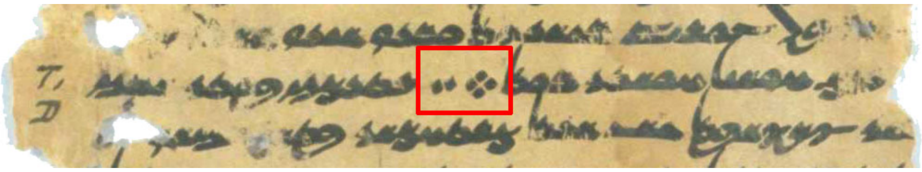


❖ ‘four dots’ followed by a ⬩ ‘dot’ (So 10028 v).

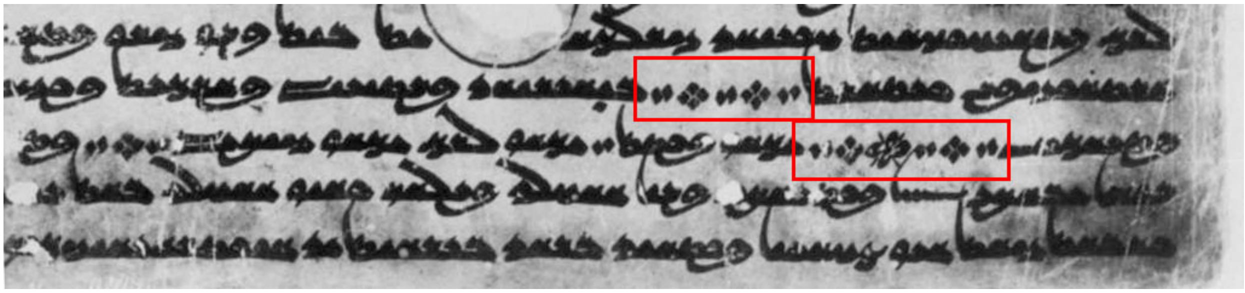


❖ ‘four dots’ at the beginning of a title (So 18224).

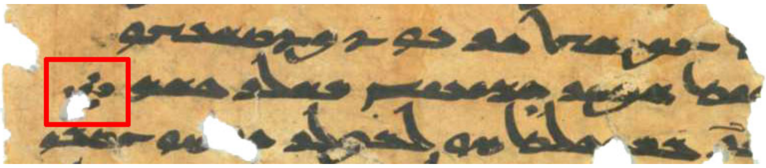
Figure 50: Additional punctuation, not proposed for encoding at present.



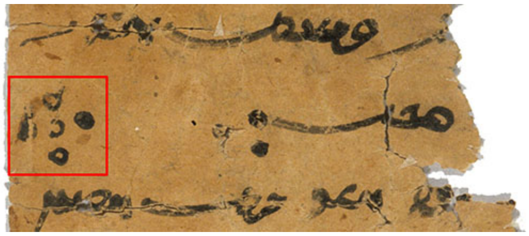
❖ ‘four dots’ and ǁ PUNCTUATION TWO VERTICAL BARS (So 10660 v).



Colophon separated from main text using combinations of ǁ PUNCTUATION TWO VERTICAL BARS, ❖ ‘four dots’, and ❁ ‘four petal fleuron with rays’ (So 18242).

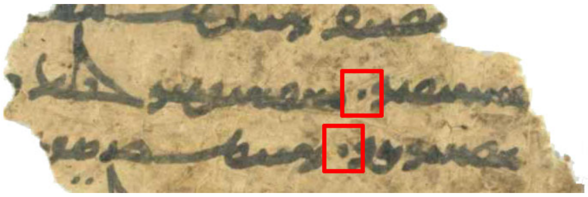


Presumably the sign ⁙ ‘five dots’ (So 18290 v).

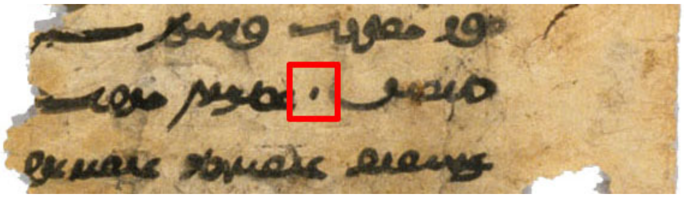


Sign similar to ⁙ ‘five dots’ (Ch/So 20512 v).

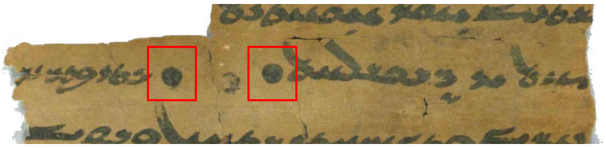
Figure 51: Additional punctuation, not proposed for encoding at present.



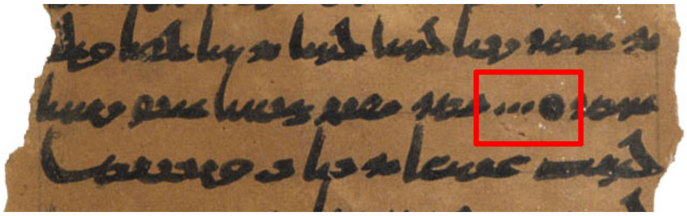
⬩ dot (So 11500 r).



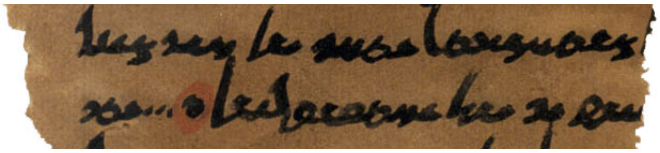
⬩ dot (So 13399(a) v).



● large dot (So 10026 r).

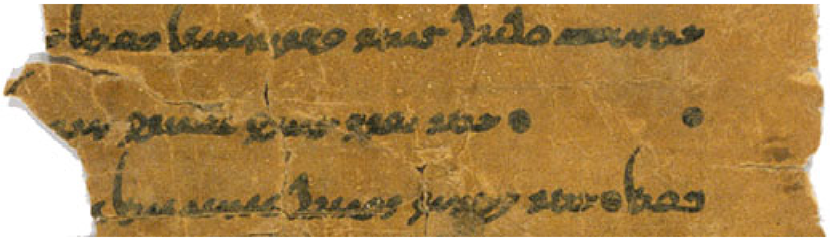


● large dot and three ⬩⬩⬩ small dots (So 20229 v).

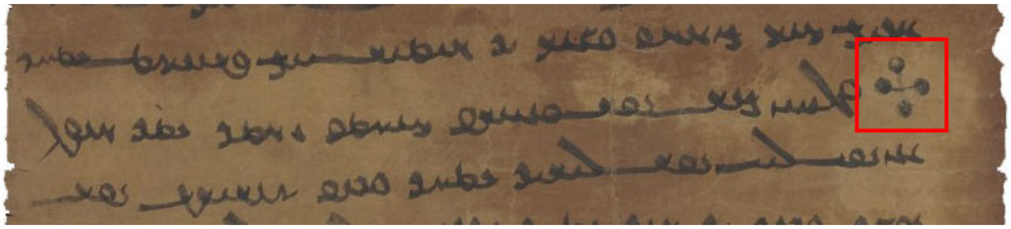


● large dot circled in red followed by three ⬩⬩⬩ small dots (So 20229 r).

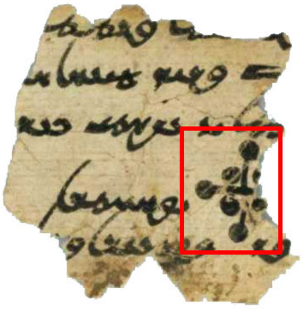
Figure 52: Additional punctuation, not proposed for encoding at present.



● large dots used for sectioning (So 11400 r).

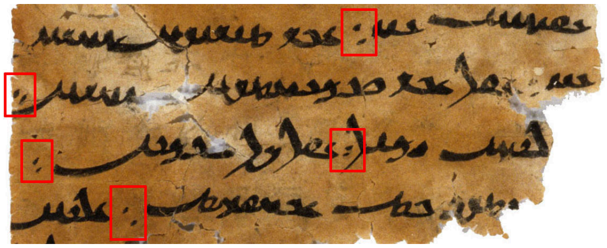


⁜ ‘cross with four dots’ (P 12).

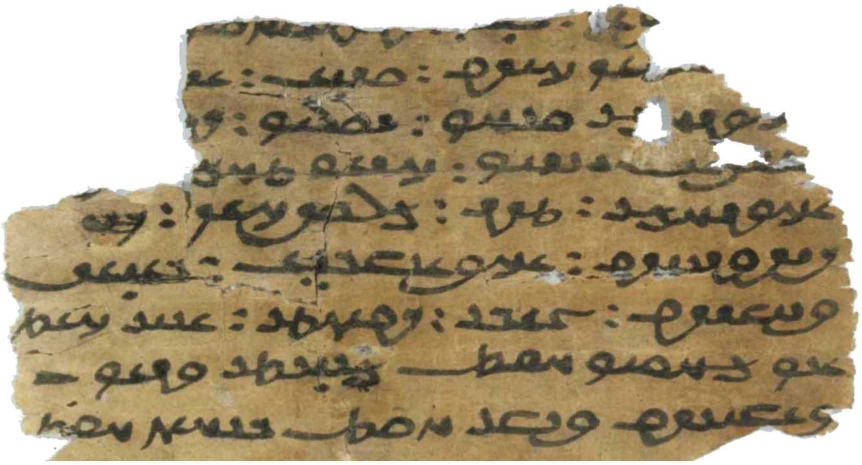


Presumably the sign ⁜ ‘cross with eight dots’ (So 10000(4) r).

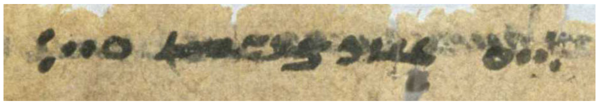
Figure 53: Additional punctuation, not proposed for encoding at present.



Usage of ∶ colon-like punctuation (Ch/So 20501).

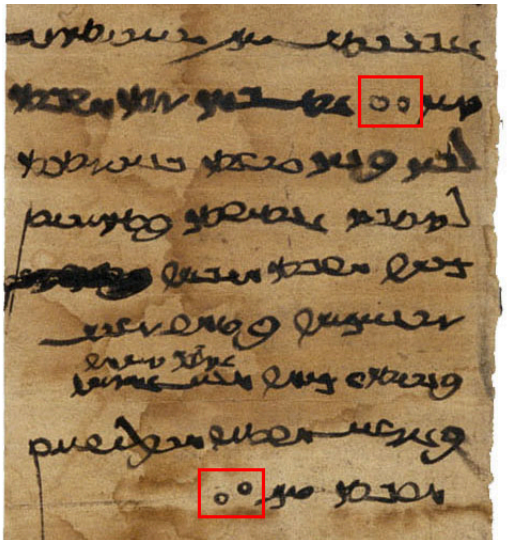


Usage of ∶ colons (Ch/So 14852a-f v).

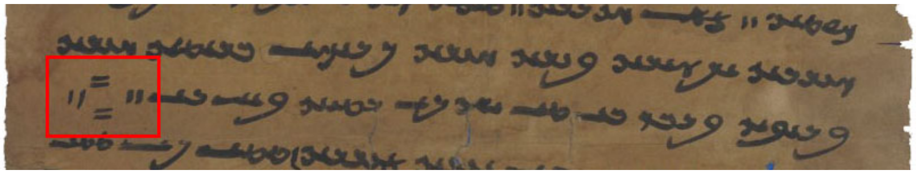


Usage of ∶ colons, ⬩ dots, ○ circles: ∶⬩⬩○ and ○⬩⬩∶ (So 14638 v).

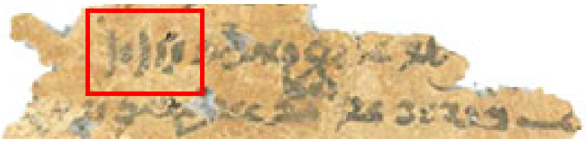
Figure 54: Additional punctuation, not proposed for encoding at present.



Usage of ○○ circles as section marks (S0 14730 v).

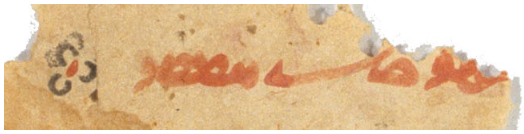


‖=‖ four bars, or instances of 𐽕 PUNCTUATION TWO VERTICAL BARS around two horizontally oriented bars (P 18).

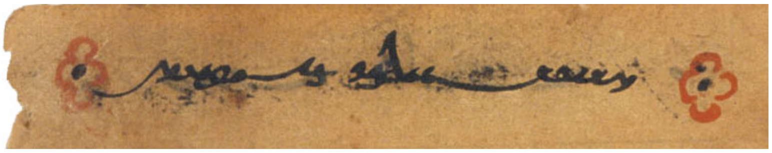


𐽖 PUNCTUATION TWO VERTICAL BARS WITH DOTS followed by |‖ ‘three bars’ (So 20195).

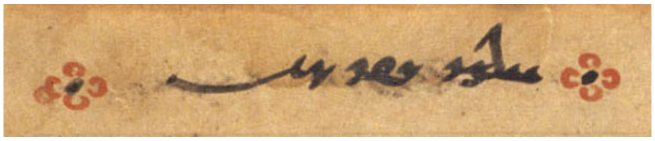
Figure 55: Additional punctuation, not proposed for encoding at present.



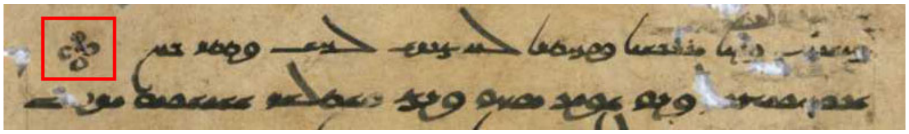
Title with ☘ ‘three petal fleuron’ (So 20208ab r).



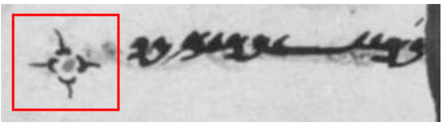
Title with ✥ ‘four petal fleuron’ (So 14441 r).



Title with ✥ ‘four petal fleuron’ (So 14570 r).

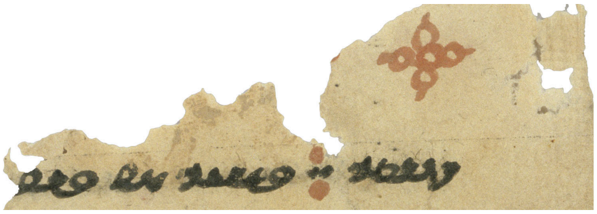


✥ ‘four petal fleuron’ (So 15201 v).

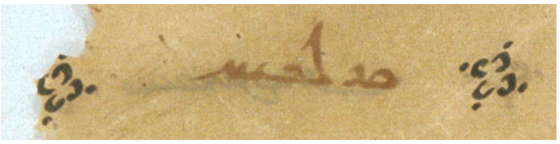


Fragment of title with ✣ ‘four petal fleuron with outer rays’ (So 14445 v).

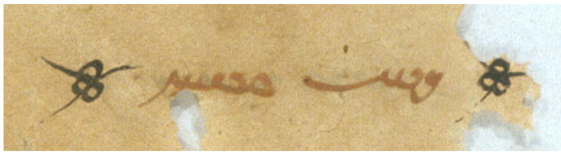
Figure 56: Specimens of text ornaments (1/2). These are not proposed for encoding at present.



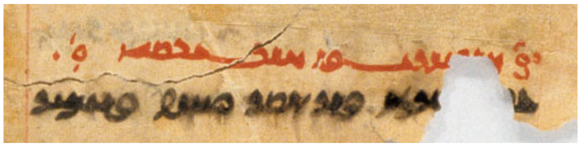
✥ ‘four petal fleuron with outer rays’ and ⁞ punctuation two vertical bars with dots with dots in red (So 18055 v).



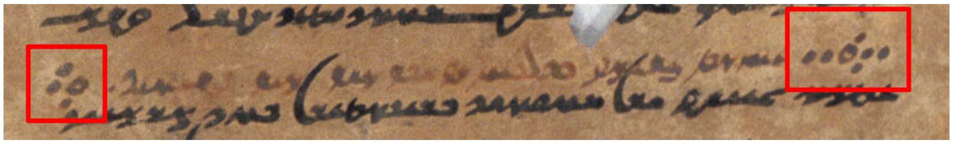
Title with ⁘ ‘four petal fleuron with dots’ (So 18248 r).



Title with ✤ ‘three petal fleuron with rays’ (So 18248 r).

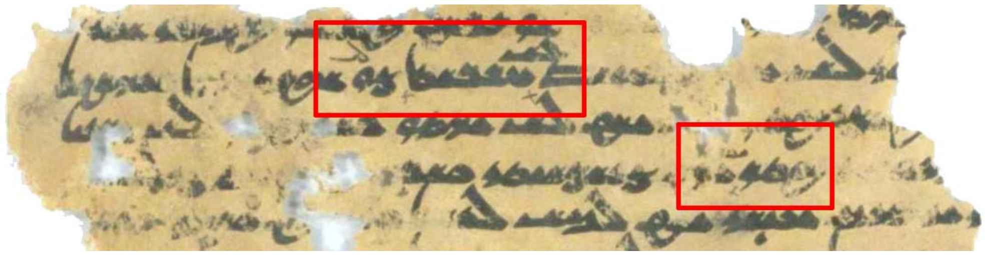


Title with ⁖ ‘right-facing three petal fleuron’ and ∴ ‘left-facing three petal fleuron’ (So 10100(e) r).

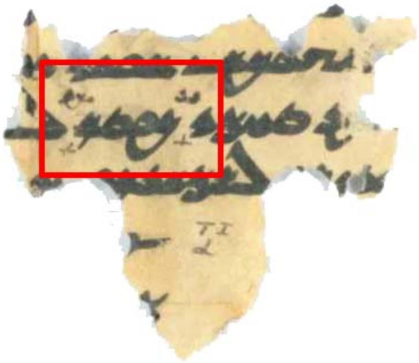


Variations of ⁖ ‘right-facing three petal fleuron’ and ∴ ‘left-facing three petal fleuron’ with additional • dots (So 14638 v).

Figure 57: Specimens of text ornaments (2/2). These are not proposed for encoding at present.



Usage of + for insertion of a word (So 10600 v).



Usage of + for insertion of a word (So 10395 v).

Figure 58: Specimens of editorial signs. These are not proposed for encoding at present.

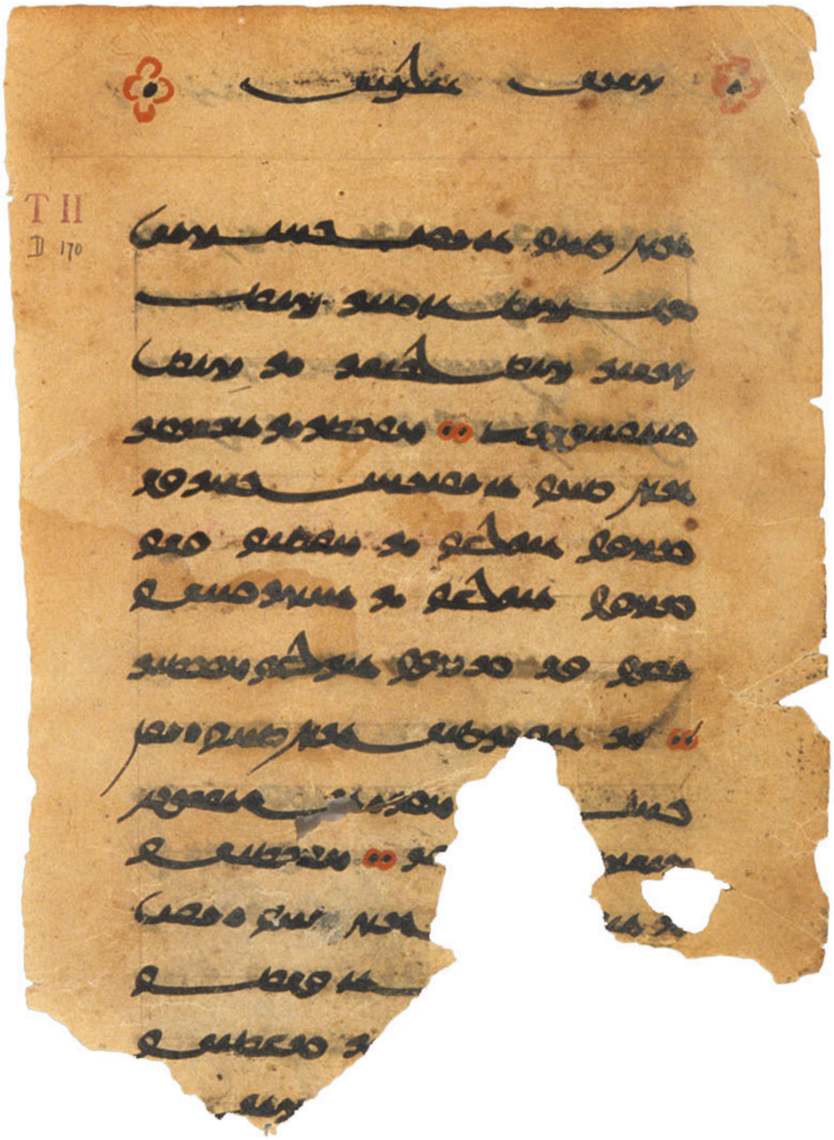


Figure 59: Elongation used for justification in a cursive Sogdian manuscript (So 14441 v).

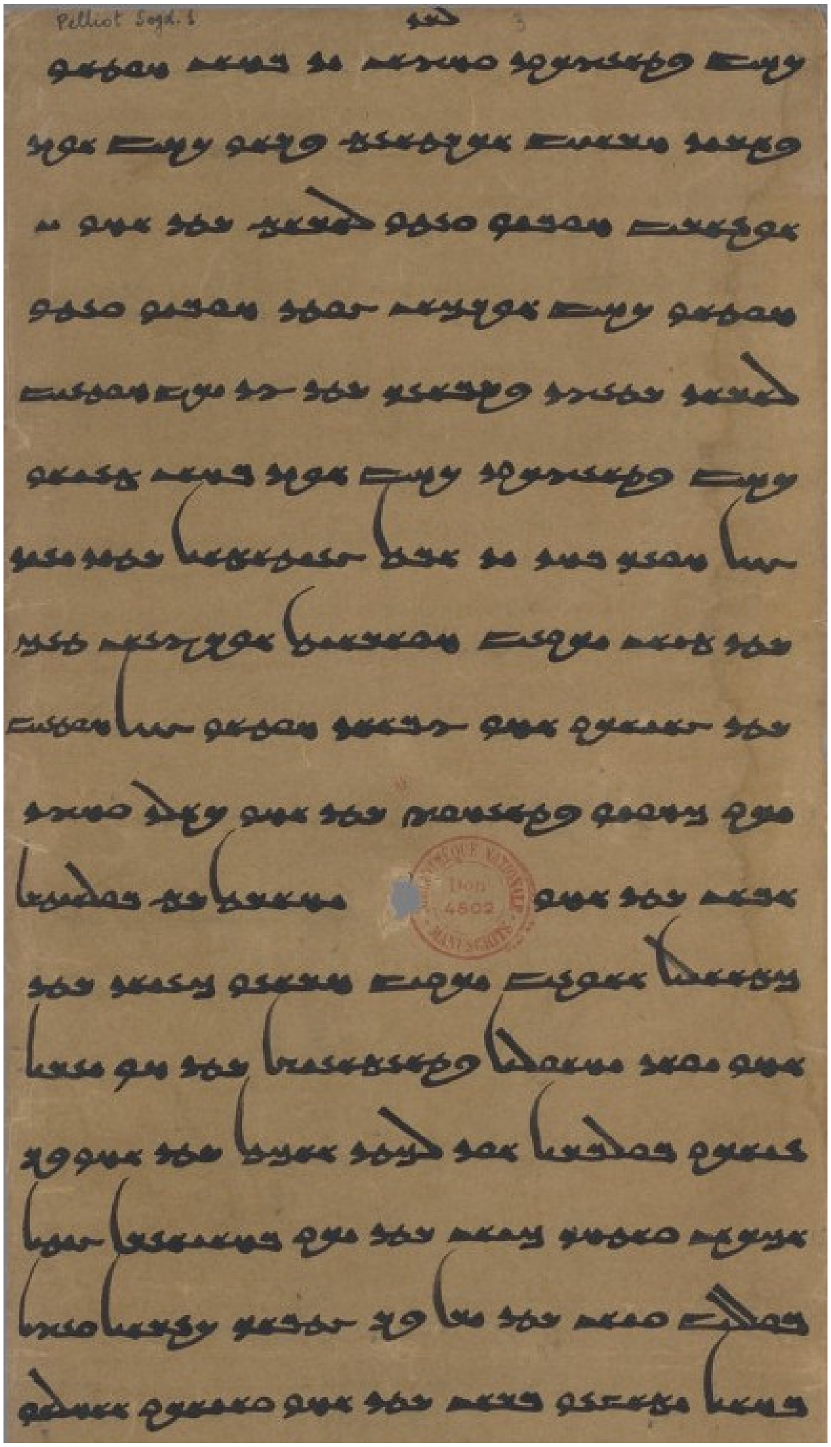


Figure 60: Excerpt from the *Vessantara Jātaka* (Pelliot Sogdien 1). Formal script.

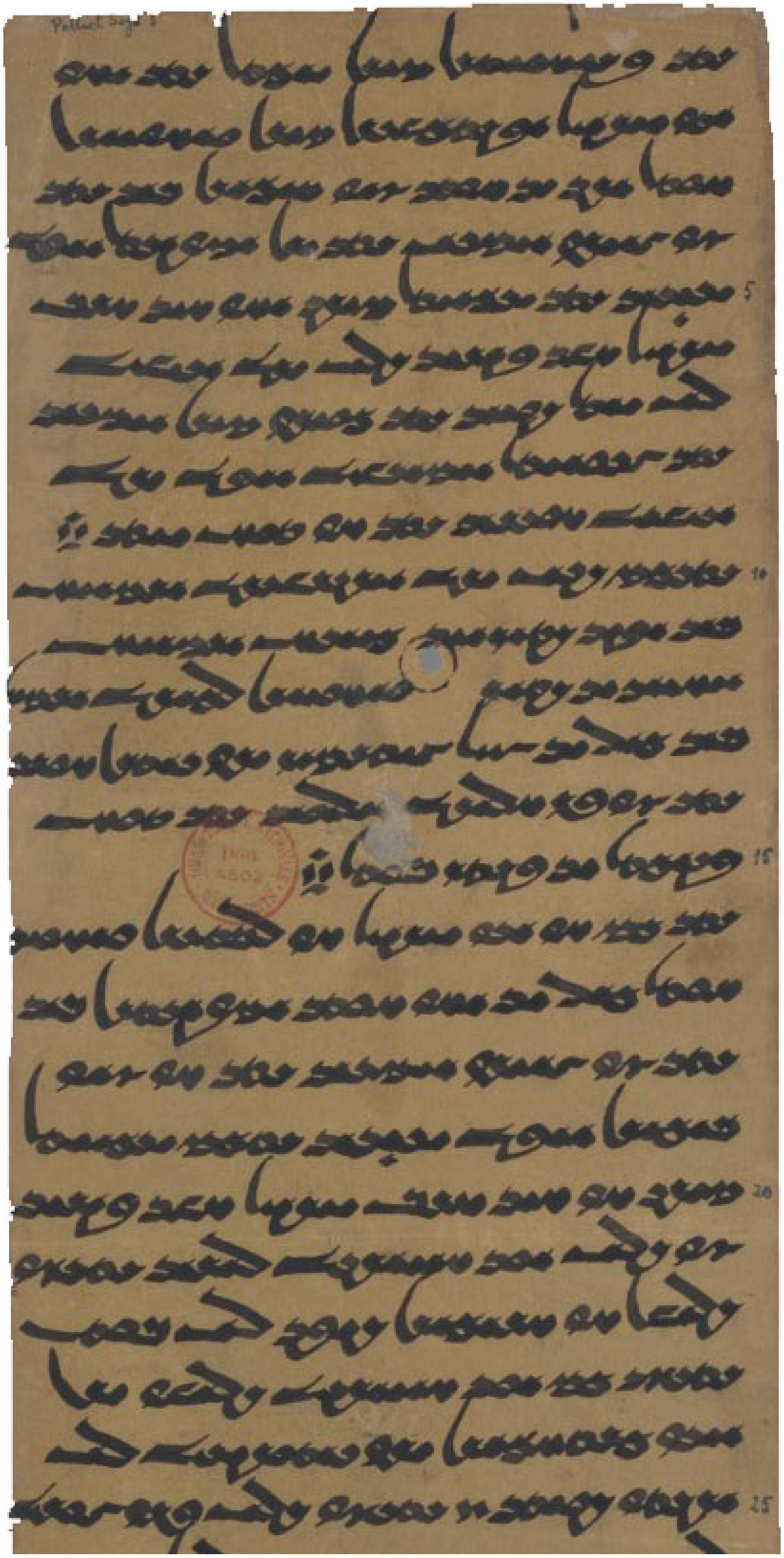


Figure 61: Excerpt from Pelliot Sogdien 3. Formal script.

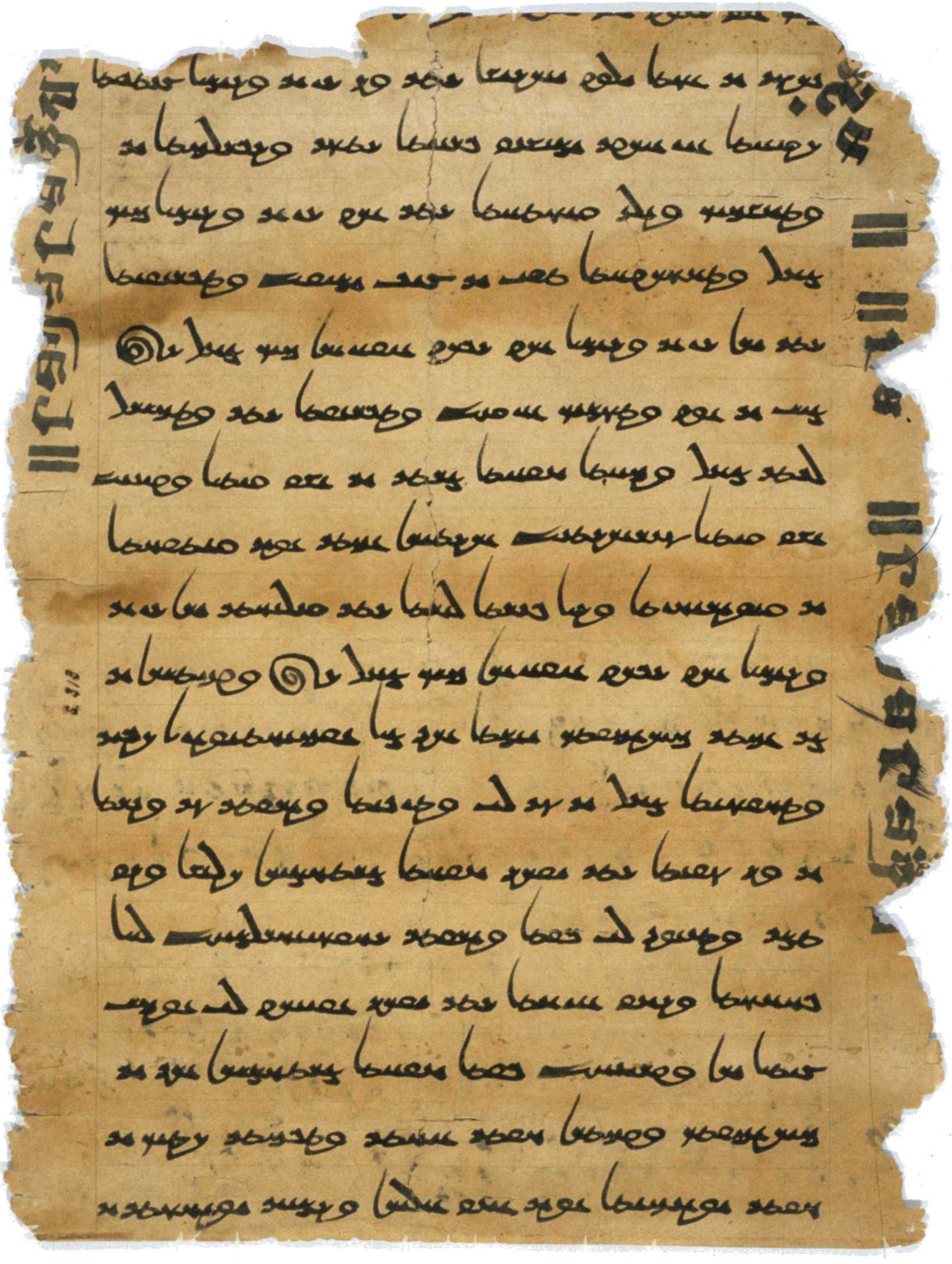


Figure 62: Fragment of the *Saṃghāṭa Sūtra* (So 20165 r). Formal script; Turkestani Brahmi in the margins.

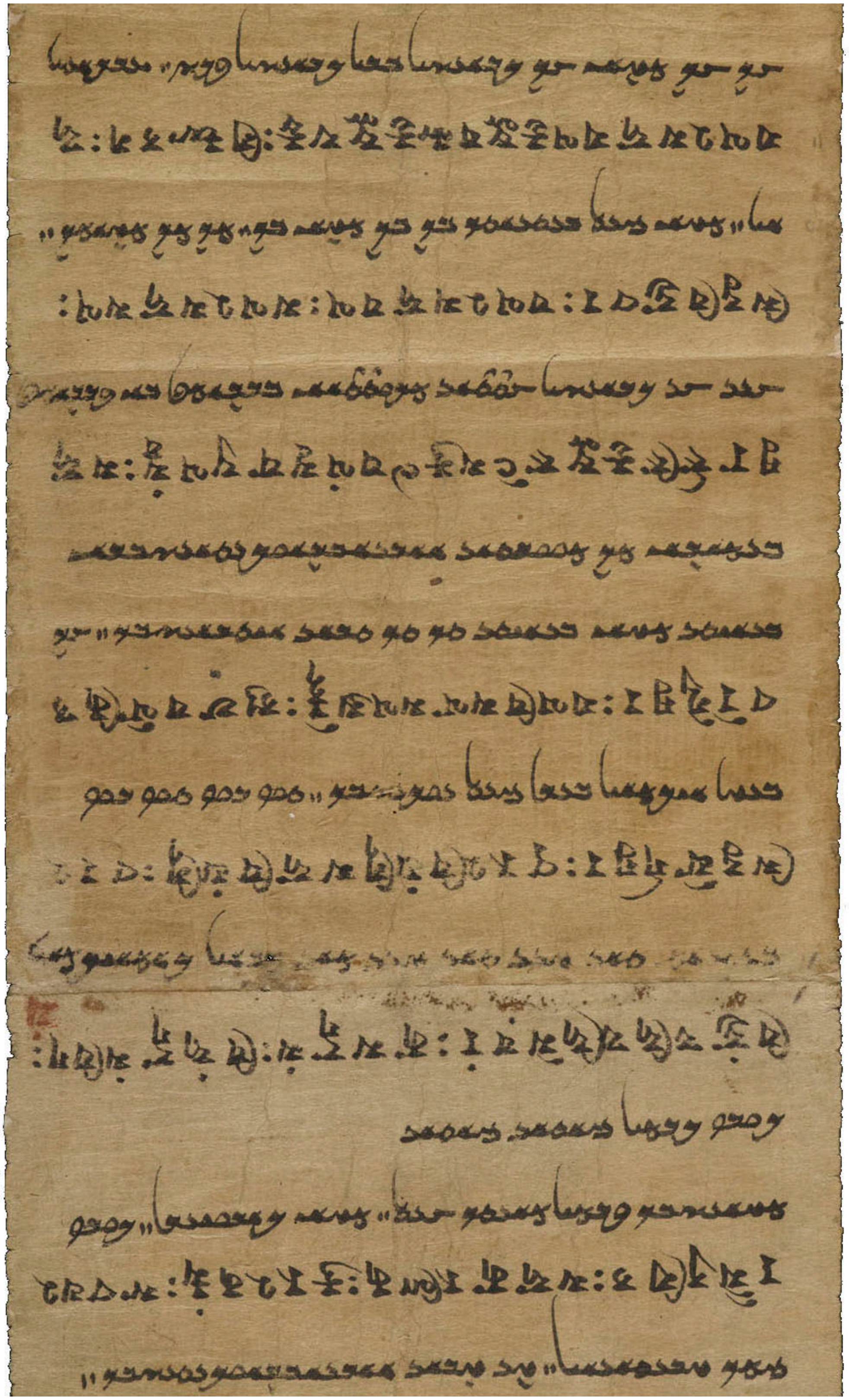


Figure 63: Folio of *Nīlakaṇṭha-dhāraṇī* in formal Sogdian and Siddham scripts (BL Or.8212/175).

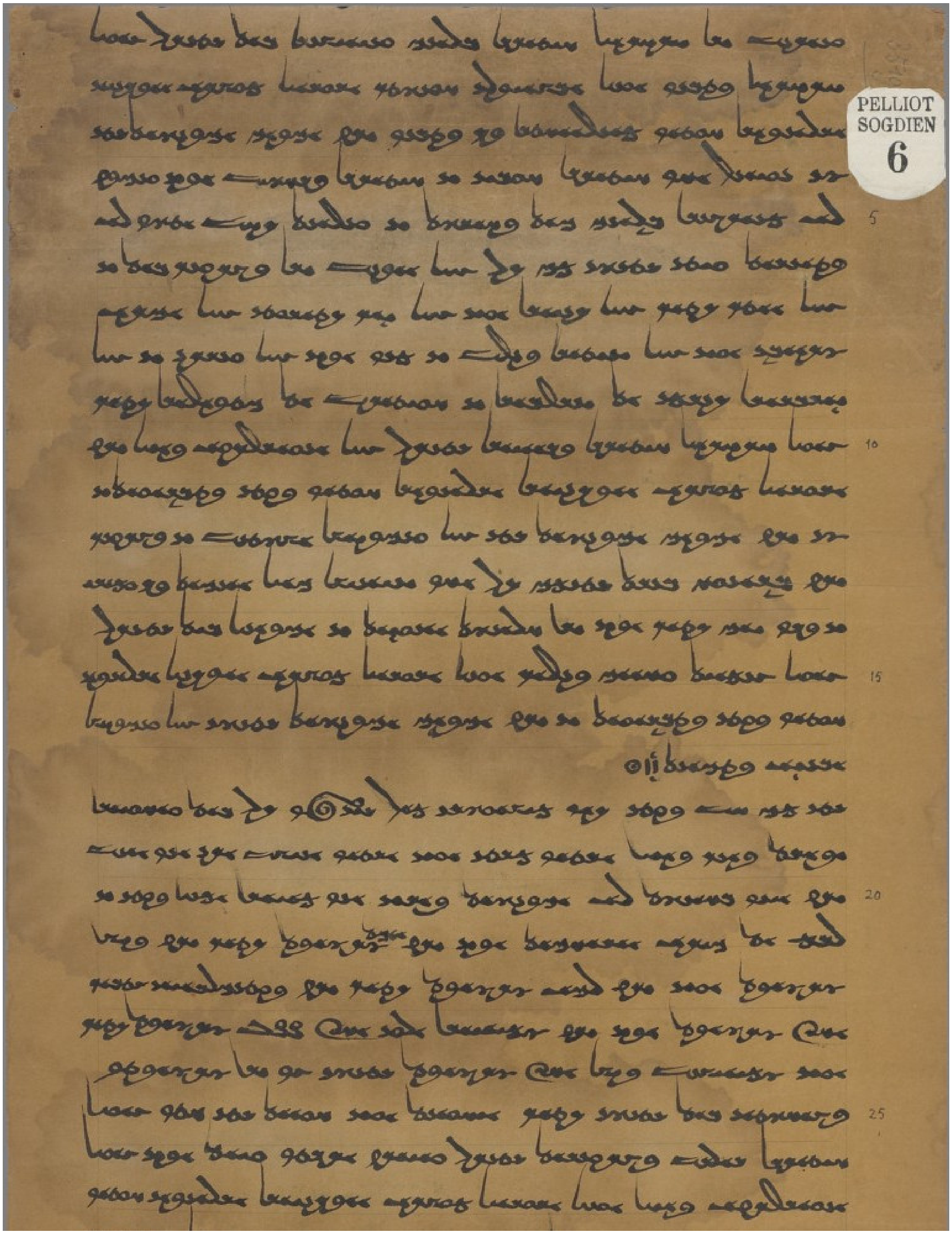
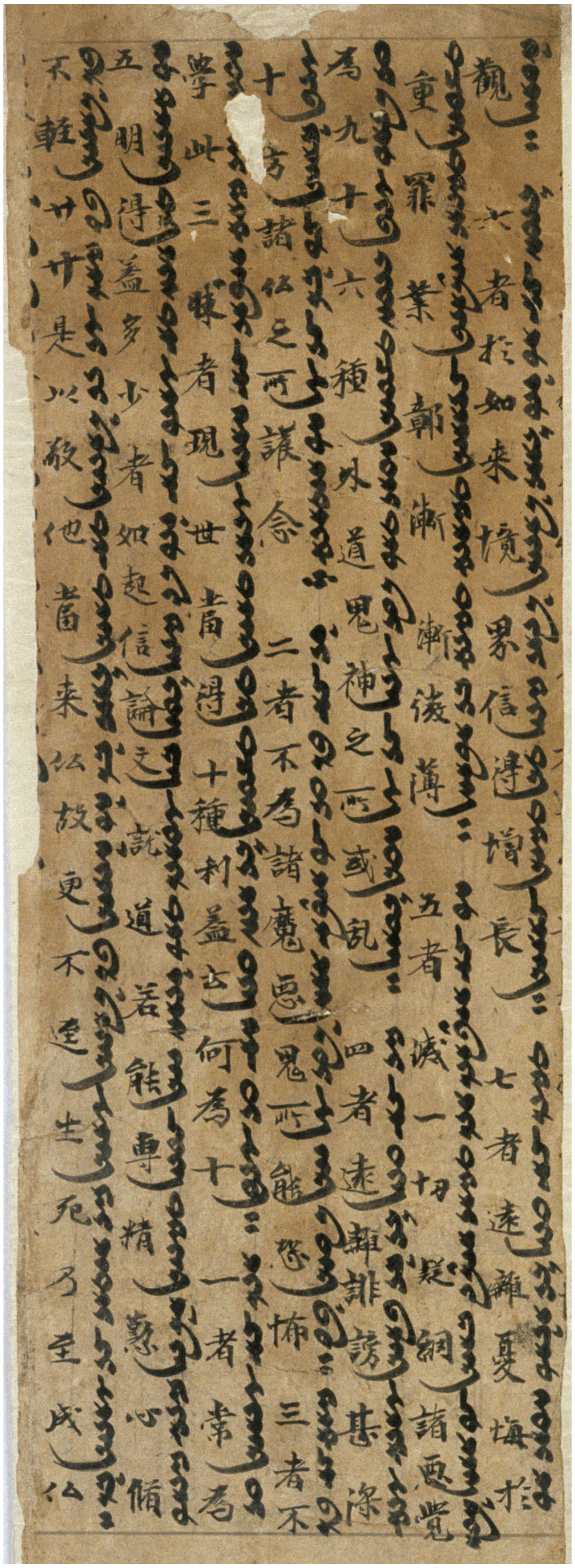

PELLIOT
SOGDIEN
6

Figure 64: Excerpt of the *Bhaiṣajya-guru-vaiḍūrya-prabhāta-tathāgata-sūtra* (Pelliot Sogdien 6). Formal script.



不輕卄卄是以敬他當来仏故更不違生死乃至成仏

五明得益多少者如起信論之說道若能專精懃心脩

學此三昧者現世當得十種利益云何為十一者常為

十方諸仏之所護念二者不為諸魔惡鬼所能恐怖三者不

為九十六種外道鬼神之所惑亂四者遠離誹謗甚深

重罪業鄣漸漸微薄五者滅一切疑網諸惡覺

觀六者於如来境界信得增長七者遠離憂悔於

Figure 65: Transcription of a Chinese text in Sogdian (So 14830). Formal script.

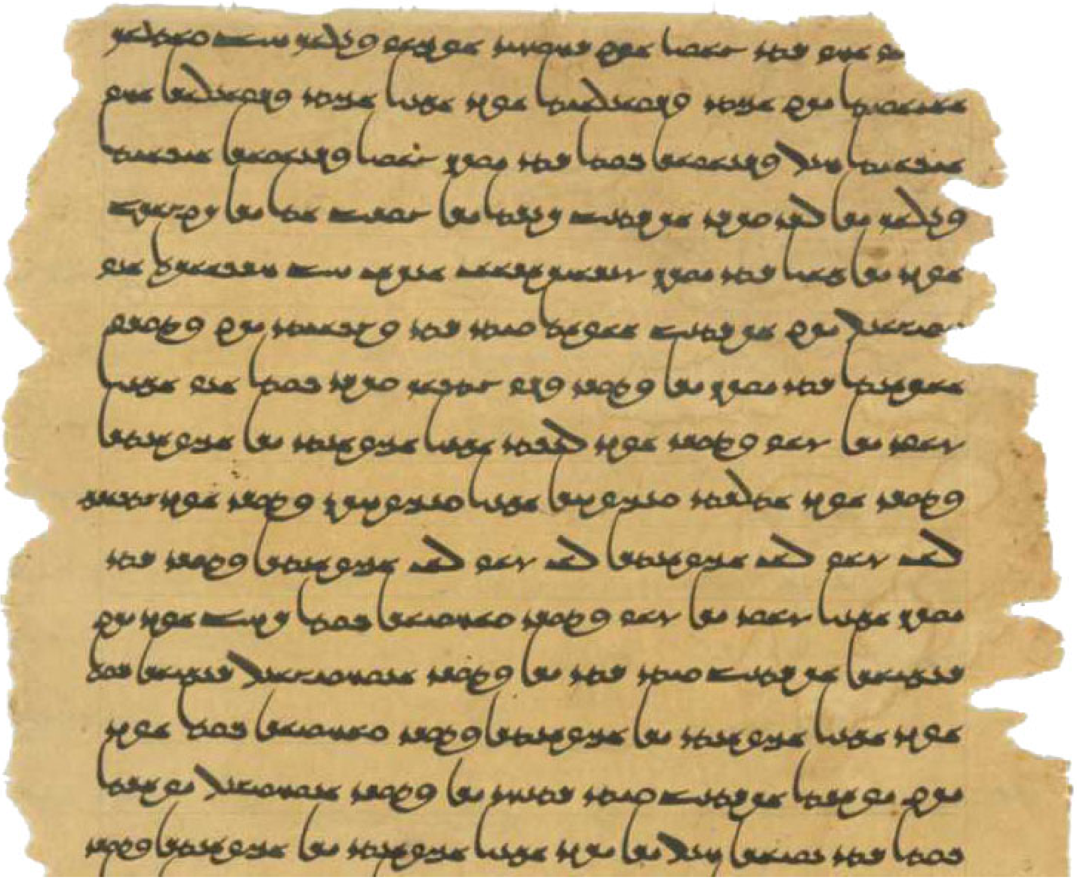


Figure 66: Excerpt of folio in formal Sogdian script (So 14851).

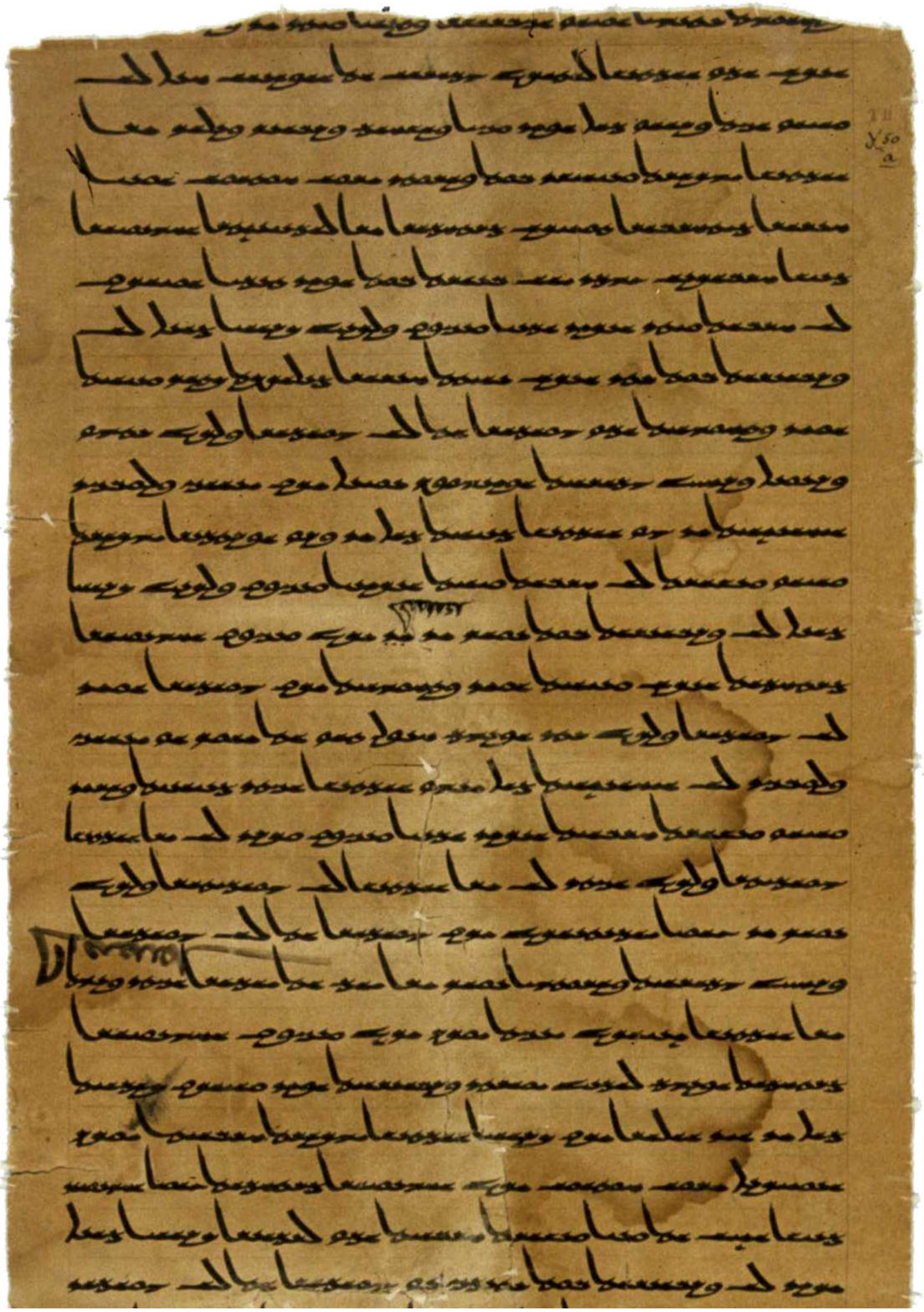


Figure 67: Excerpt of folio in formal Sogdian script (So 14850).

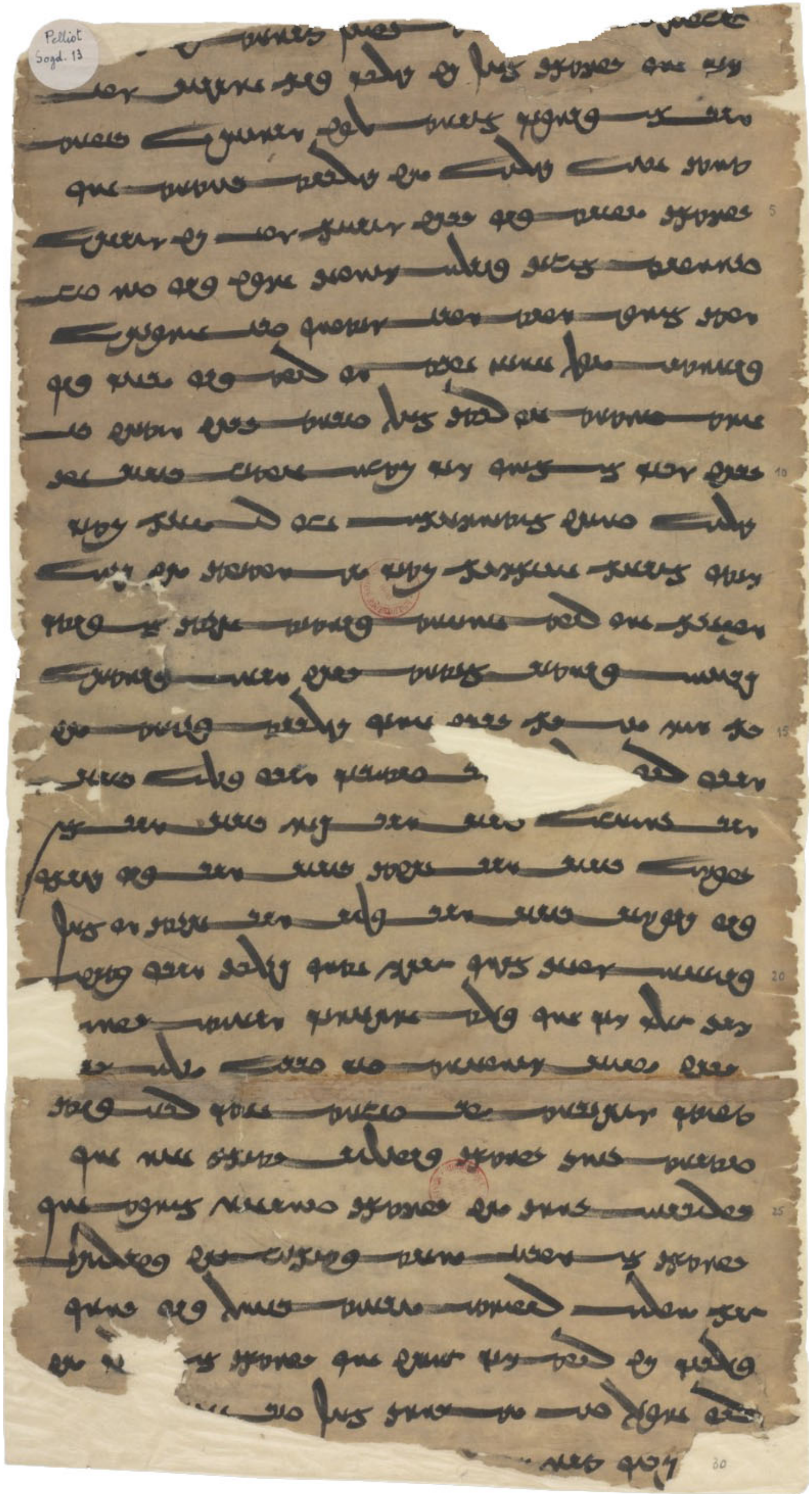
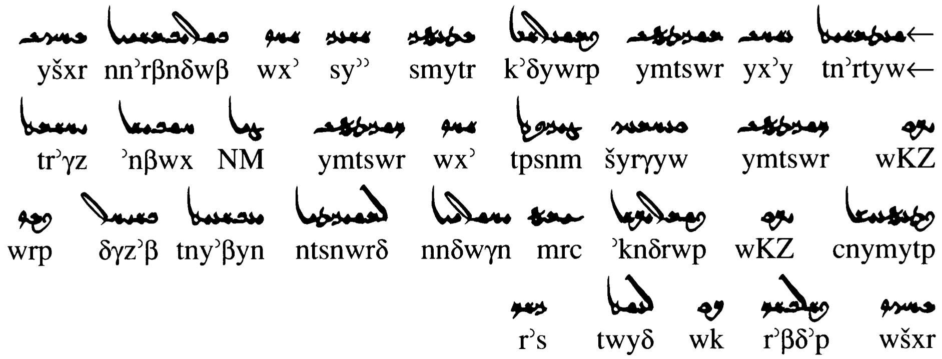

Pelliot
Sogd. 13
5
10
15
20
25
30

Figure 68: Fragment of the 'Story of Rustam' in cursive Sogdian script (Pelliot 13). Line 24 through the first three words of line 28 are given as a printed specimen in fig. 69.

Story of Rustam


yšxr nn’rβnδwβ wx’ sy’’ smytr k’δywrp ymtswr yx’y tn’rtyw←
tr’γz ’nβwx NM ymtswr wx’ tpsnm šyrγyw ymtswr wKZ
wrp δγz’β tny’βyn ntsnwrδ nnδwγn mrc ’knδrwp wKZ cnymytp
r’s twyδ wk r’βδ’p wšxr

1. Transliteration:	wytr’nt	y’xy	rwstmy	prwyδ’k	rtyms
2. Normalization:	wītarand	yaxī	Rustami	parwēδē	rti-mas
3. Gloss:	IMPF.they.departed	brave	Rustam.GEN	to.seek	and-then

1.	’’ys	’xw	βwδnβr’nn	rxšy	ZKw	rwstmy
2.	āyas	axu	βōδan-βarān	Raxši	awu	Rustami
3.	came	the.NOM	perception-bearing	Raxš.NOM	the.ACC	Rustam.ACC

1.	wyγryš	mnspt	’xw	rwstmy	MN	xwβn’	zγ’rt
2.	wīγrēš	manspat	axu	Rustami	čon	xuβna	žγart
3.	IMPF.he.woke	IMPF.arose	the.NOM	Rustam.NOM	from	sleep.ABL	quickly

1.	ptymync	ZKw	pwrδnk’ crm	nγwδnn	δrwnstn	nyβ’ynt
2.	ptīmēnč	awu	pu^{r}δang-čarm	nγōδan	δrūn-stan	nīβēnd
3.	IMPF.he.donned	the.ACC	leopard-skin	garment	bow-container	IMPF.he.tied

1.	β’zγδ	prw	rxšw	p’δβ’r	kw	δywt	s’r
2.	βāžγaδ	par-ō	Raxšu	pāθfār	kū	δēwt	sār
3.	IMPF.mounted	on-the.ACC	Raxš.ACC	IMPF.hurried	to-	demon.PL	-ward

‘They (the demons) departed in search of the brave Rustam. Then came the perceptive(?) Rakhsh (his horse) and woke Rustam. Rustam arose out of his sleep, quickly donned (his) leopard-skin garment, tied on his bow-case, mounted Rakhsh, and hurried toward the demons.’

—From a (Manichean?) version of the story of Rustam (Benveniste 1940A, pls. 193–94, 1940B: 135; Sims-Williams 1976: 54–57).

Figure 69: Example of printed Sogdian (from Skjærvø 1996: 530). A manuscript containing the text specimen is shown in fig. 68.

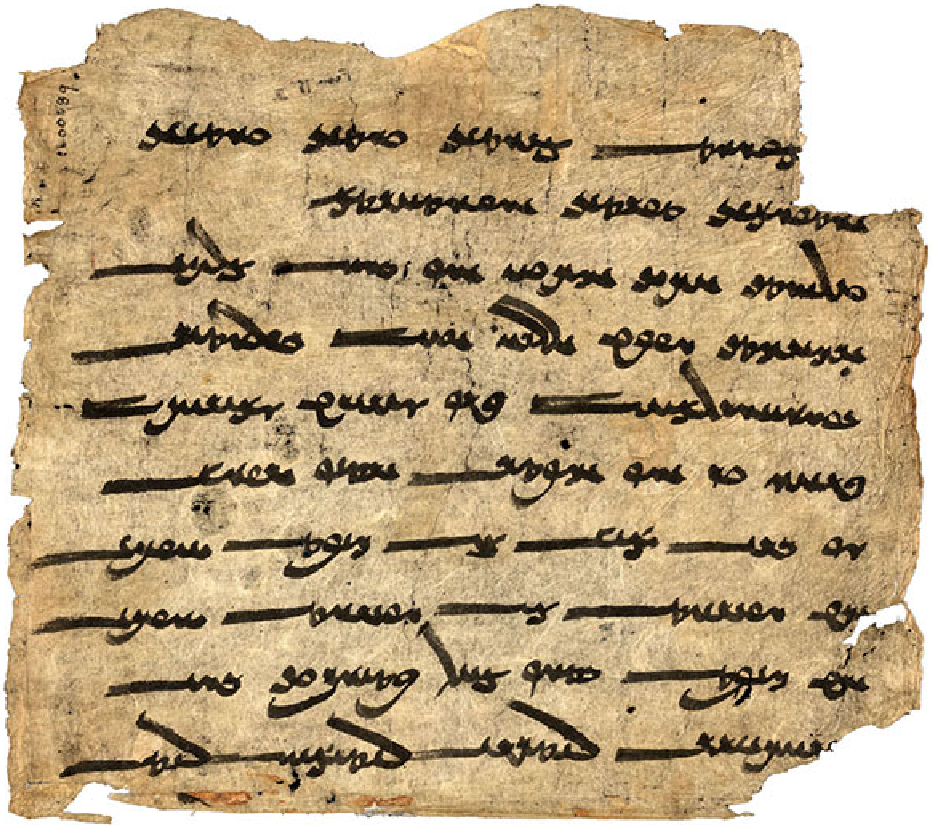


Figure 70: The Zoroastrian prayer, *ashem vohu* (*aṣ̌əm vohū*), in cursive Sogdian script, 10–11th c (Or.8212/84 recto).

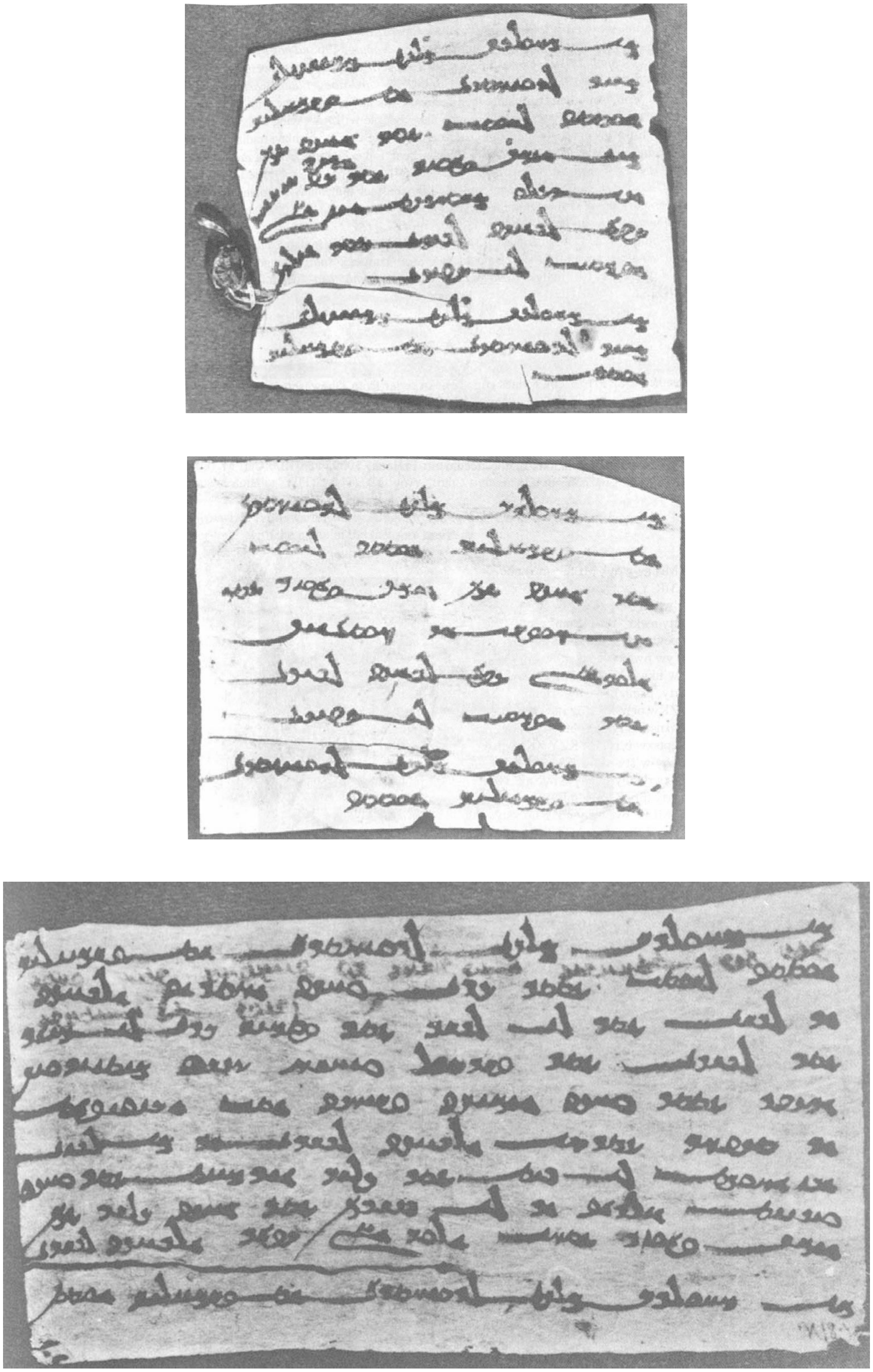


Figure 71: Letters from Đēwāštīc, ruler of Panjakant, to the *pr›m›nδ›r* (= *framāndār*) Awat (Mugh A-2, A-3, A-18 (recto); reproduced in Livshits 2015: 113, 115, 109). Cursive script.

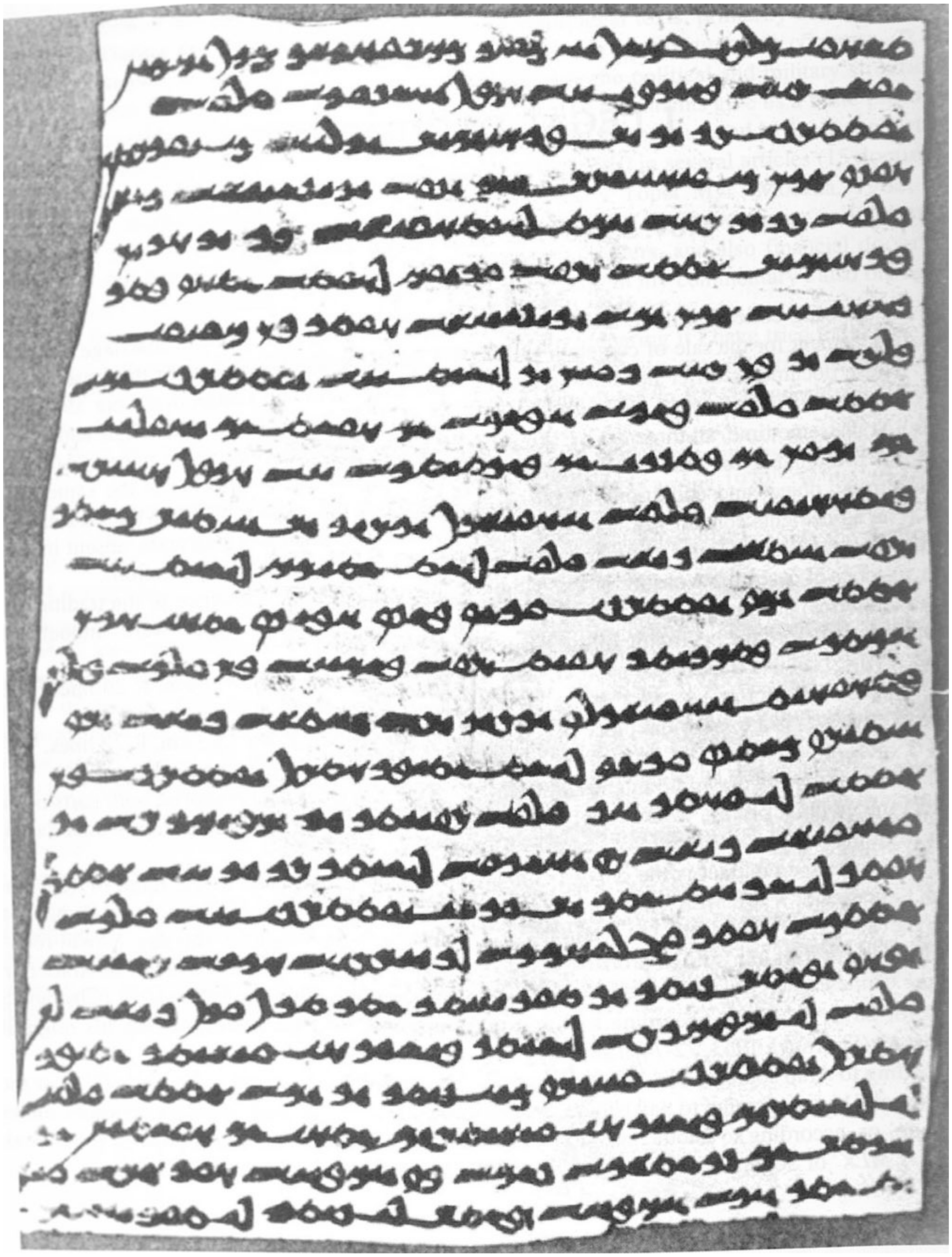


Figure 72: Marriage contract (Mugh Nov. 3 recto; reproduced in Livshits 2015: 18). Cursive script.

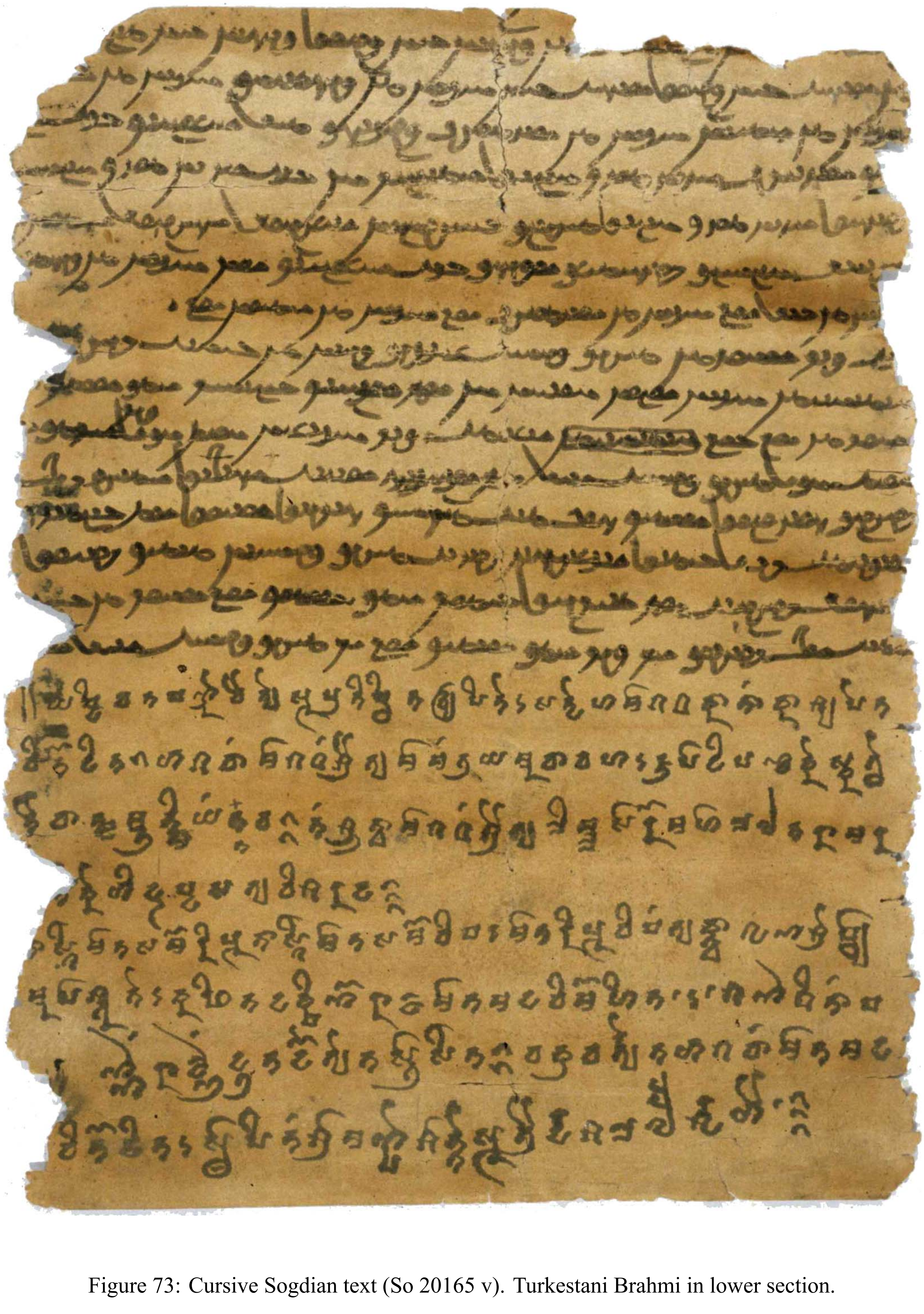


Figure 73: Cursive Sogdian text (So 20165 v). Turkestani Brahmi in lower section.

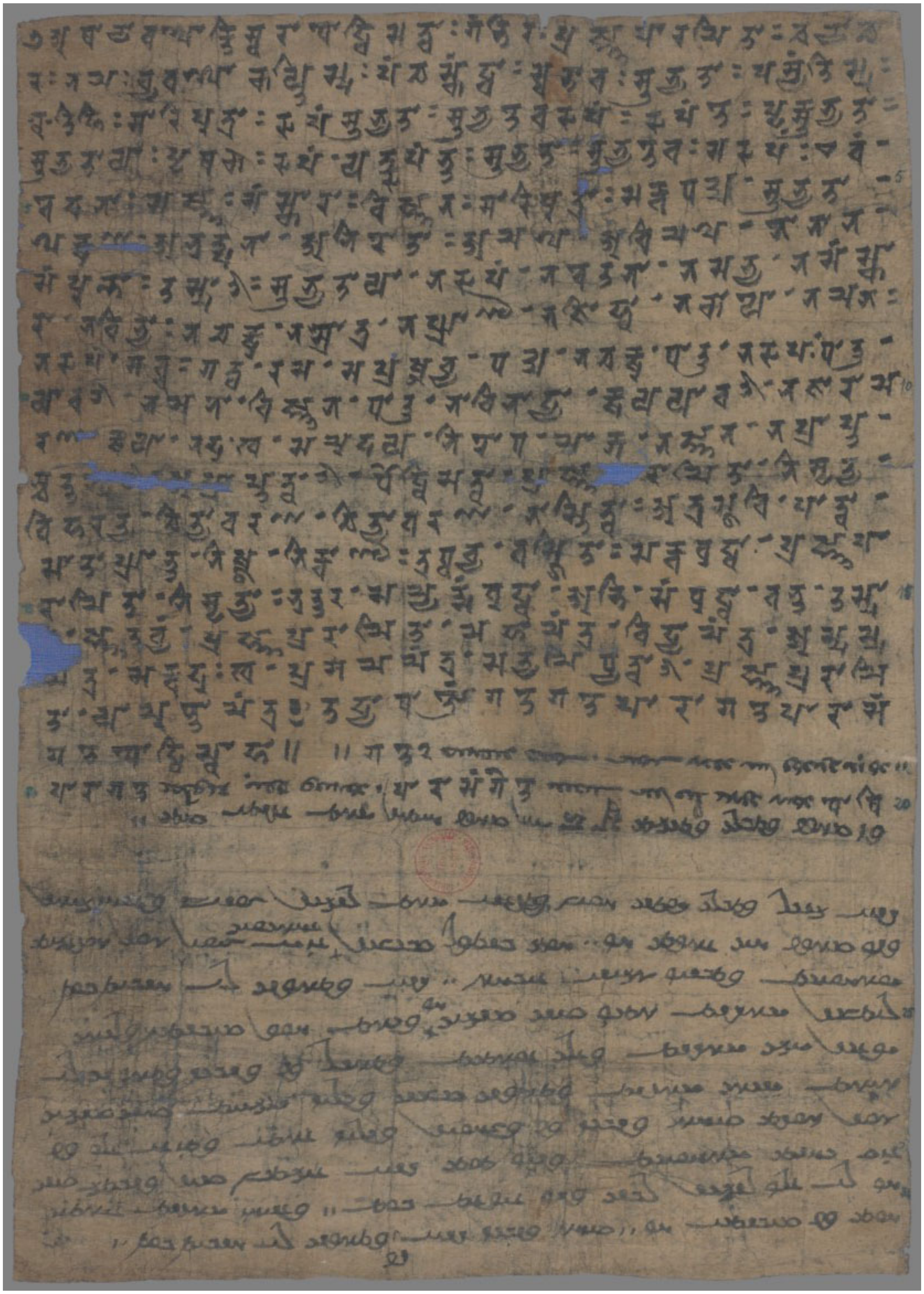


Figure 74: The *Prajñā-pāramitā-hṛdaya-sūtra* and *Pañca-vimśatikā-prajñā-pāramitā-nāma-dhāraṇī* in Siddham or a variety of Central Asian Brahmi and cursive Sogdian (Pelliot 16).

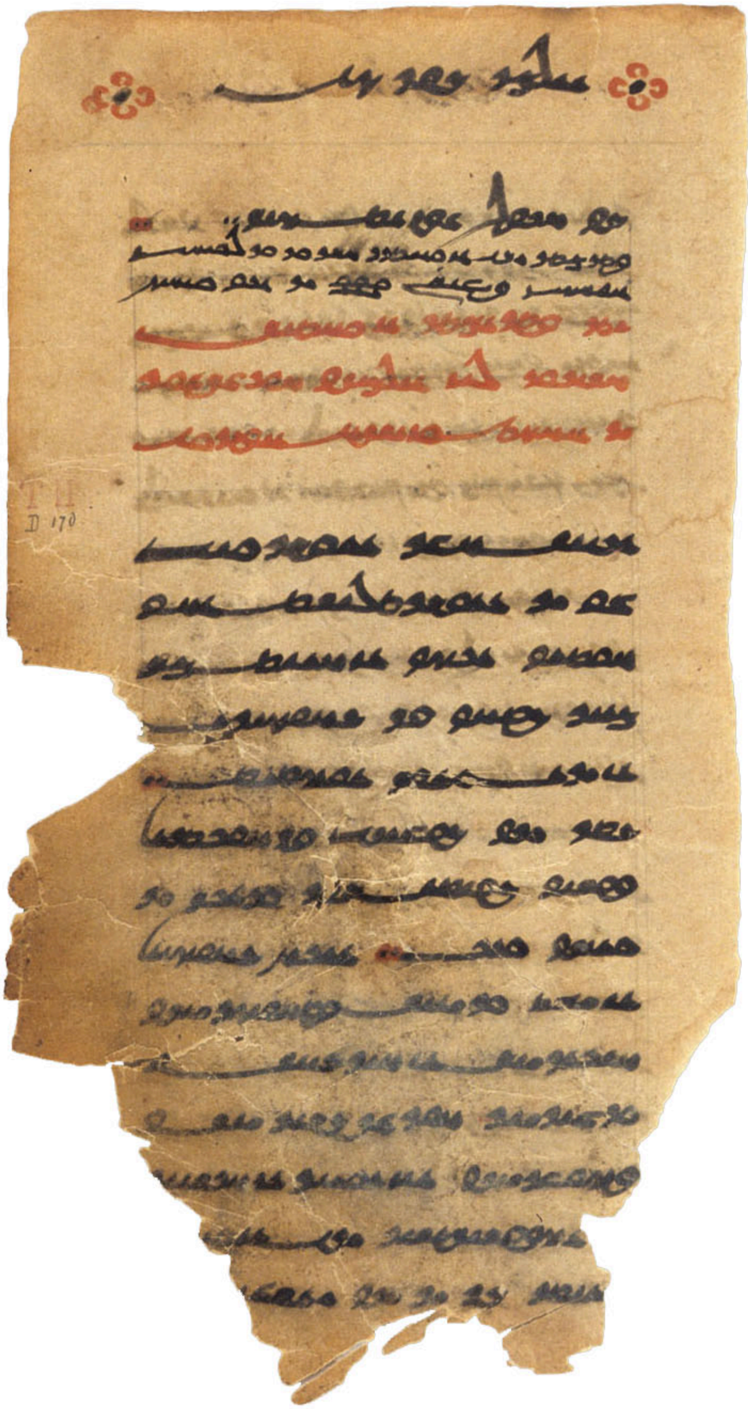


Figure 75: Folio from a cursive Sogdian manuscript (So 14570 recto).

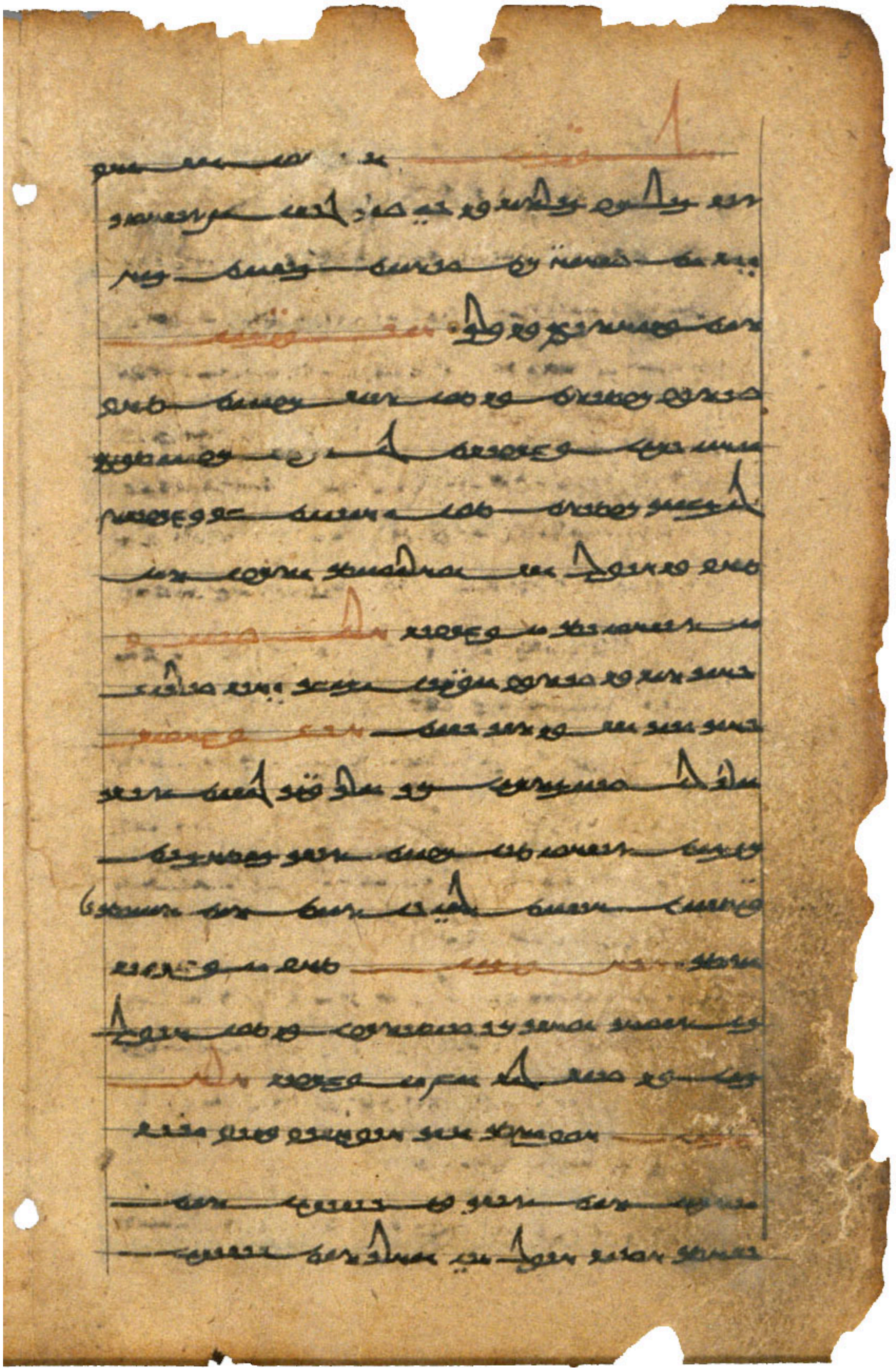


Figure 76: Excerpt from a cursive Sogdian manuscript (So 14410 verso).

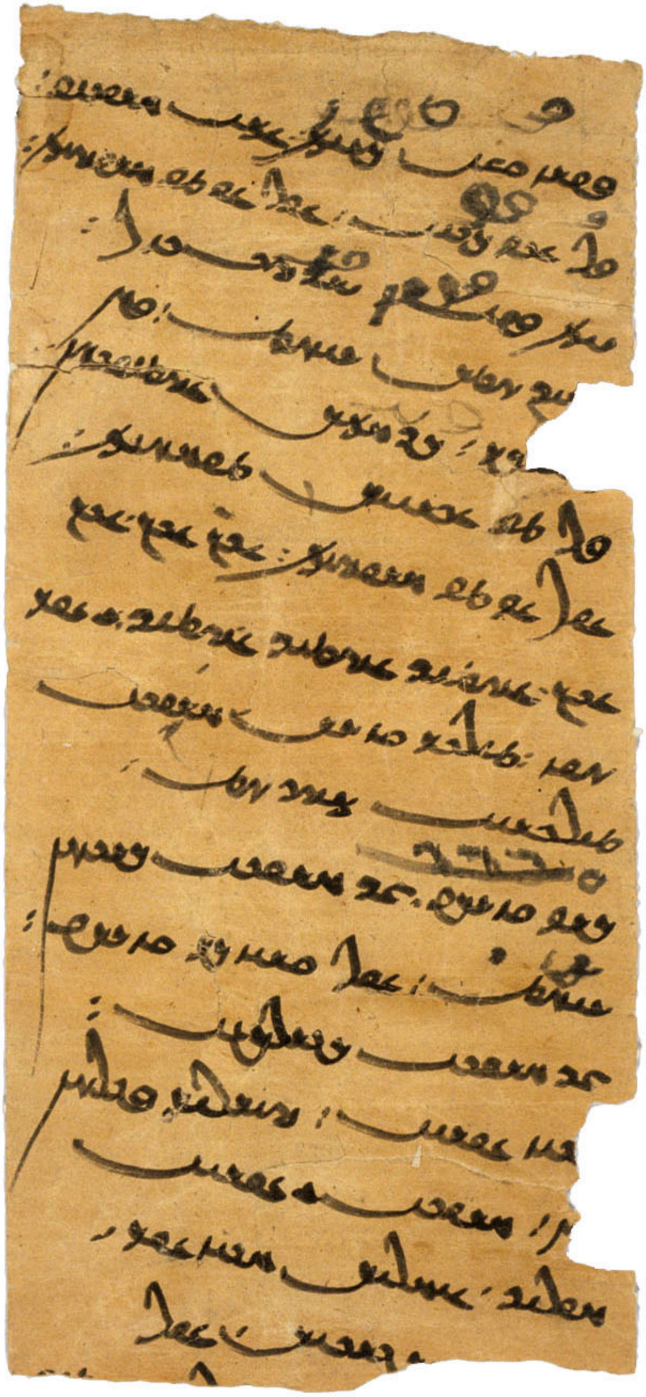


Figure 77: Cursive Sogdian manuscript (Ch/So 20135 verso).

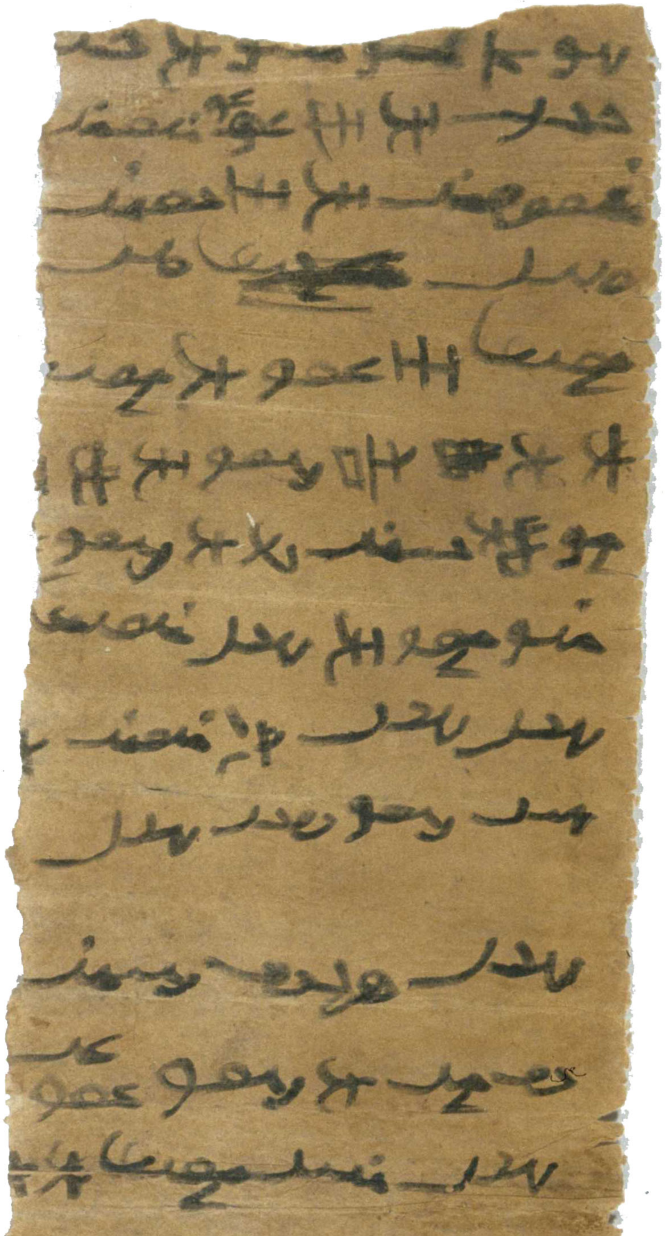


Figure 78: Cursive Sogdian with Chinese interspersed (excerpt of Ch/So 14800 verso). The actual orientation of this folio is likely to be vertical.

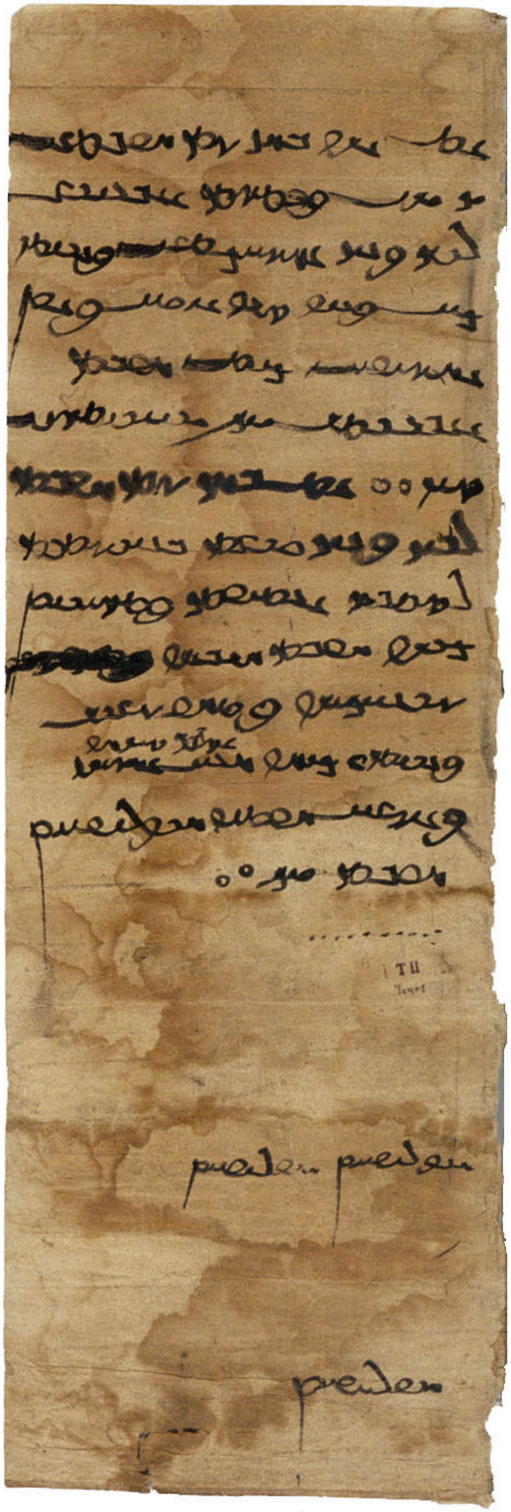
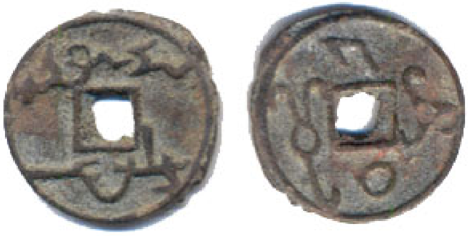
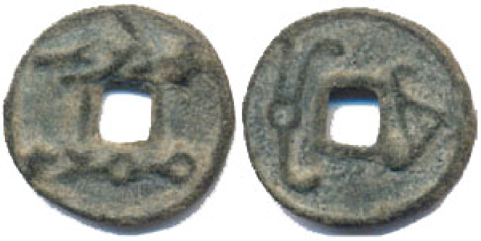


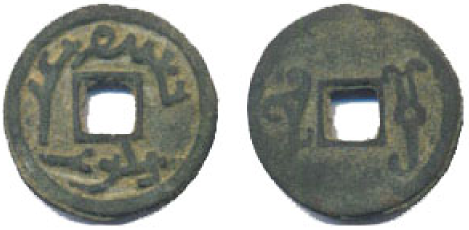
Figure 79: Excerpt from a cursive Sogdian manuscript (Ch/So 14730 verso).



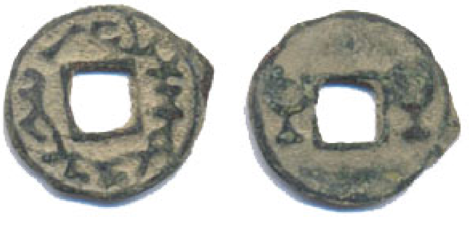
Coin from c. 642/655 CE. Reverse: 𐽁𐼷𐽁𐼾𐼷𐽀 𐼺𐼹𐼸𐼰 *šyšpyr MLKʾ*. Obverse: Four *tamgha* around the center punch.



Coin from 7th century CE. Reverse: 𐼺𐼹𐼸𐼰 𐼴𐼵𐼴𐽀𐼲 *MLKʾ wzwrγ*. Obverse: Two *tamgha* on the sides of the center punch.



Coin from c. 650/655, no later than 696 CE. Reverse: 𐼱𐽀𐼶𐼴𐼺𐼰𐼻 𐼺𐼹𐼸𐼰 *βrxwmʾn MLKʾ*. Obverse: Two *tamgha* on the sides of the center punch.



Coin from 8th century CE. Reverse: 𐼻𐼻𐼷𐼰𐼱𐼷𐼰𐽂 𐼼𐼺𐼷𐼹𐼻𐼿 *nnyʾβyʾt smyδnc*. Obverse: Two *tamgha* on the sides of the center punch.

Figure 80: Sogdian inscriptions on coins (from “Coins of Central Asia”).

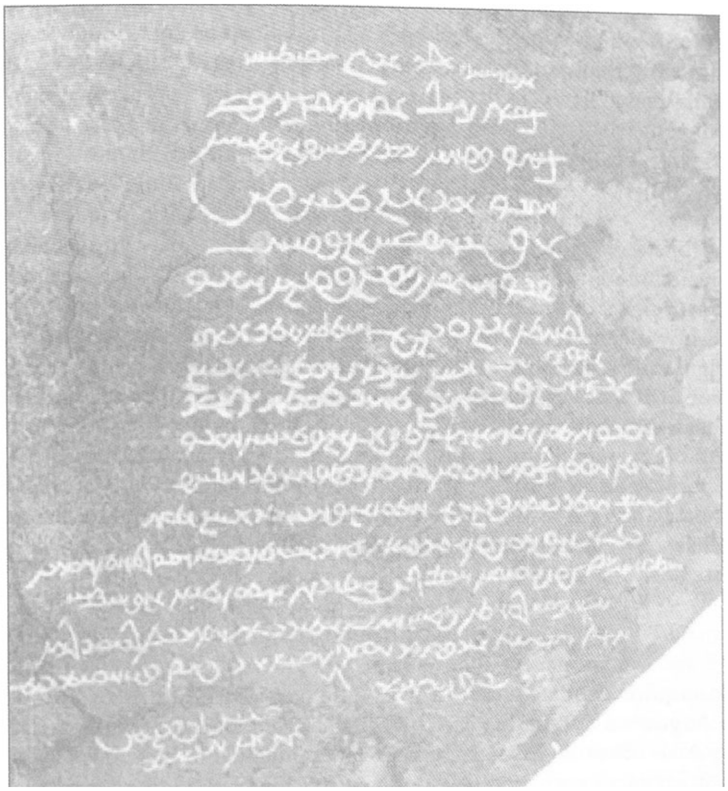


Figure 81: Sogdian inscription from Kulan-sai (I-6) (reproduced in Livshits 2015: 289).

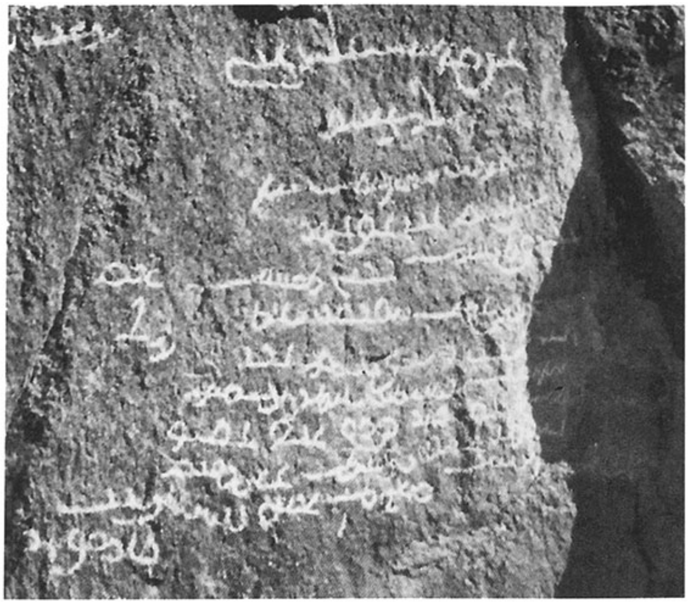
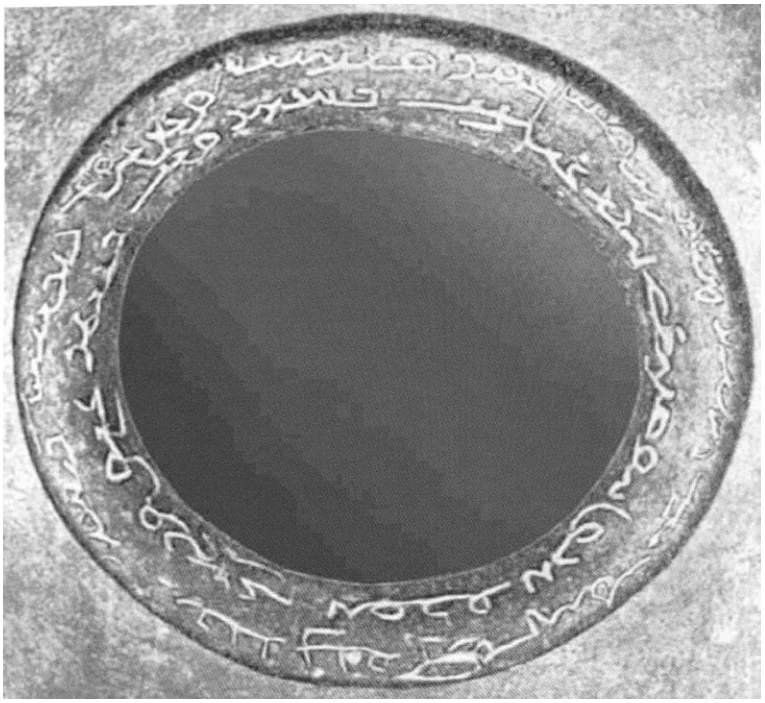


Figure 82: Sogdian inscription from Terek-sai (II-A) (reproduced in Livshits 2015: 296).



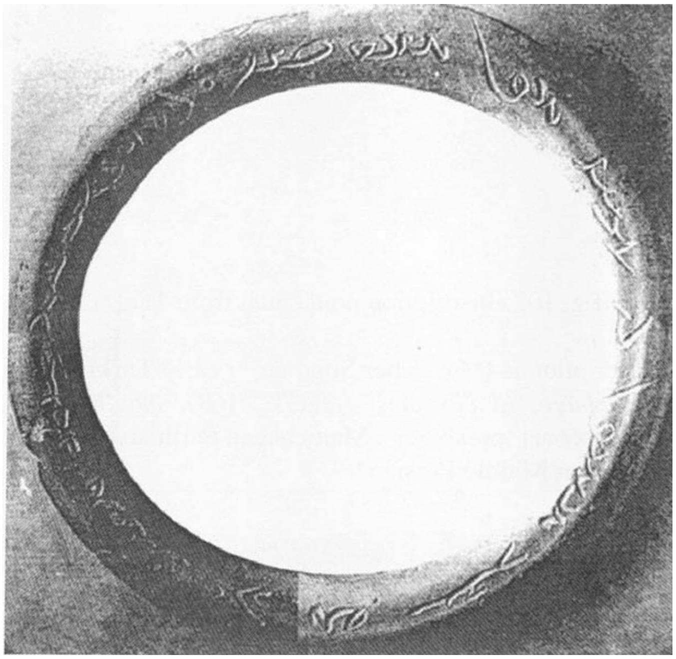


Figure 83: Sogdian inscriptions on pots (*khum*) found at Novopokrova (above) and Krasnaya Rechka (below) (reproduced in Livshits 2015: 272, 274).

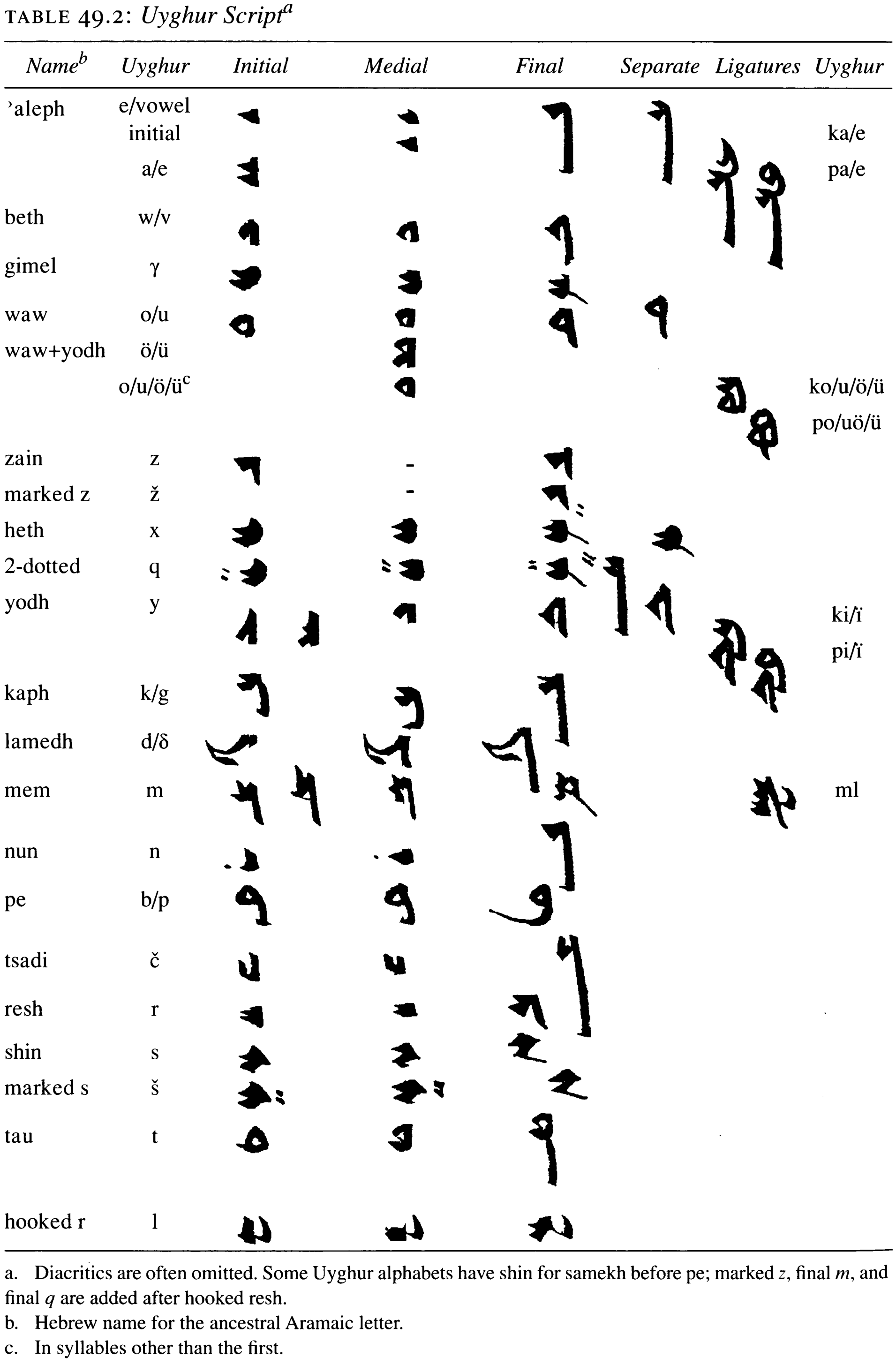


TABLE 49.2: *Uyghur Script*[a]

Name[b]	*Uyghur*	*Initial*	*Medial*	*Final*	*Separate*	*Ligatures*	*Uyghur*
'aleph	e/vowel initial						ka/e
	a/e						pa/e
beth	w/v						
gimel	γ						
waw	o/u						
waw+yodh	ö/ü						
	o/u/ö/ü[c]						ko/u/ö/ü
							po/uö/ü
zain	z		-				
marked z	ž		-				
heth	x						
2-dotted	q						
yodh	y						ki/ï
							pi/ï
kaph	k/g						
lamedh	d/δ						
mem	m						ml
nun	n						
pe	b/p						
tsadi	č						
resh	r						
shin	s						
marked s	š						
tau	t						
hooked r	l						

[a] Diacritics are often omitted. Some Uyghur alphabets have shin for samekh before pe; marked *z*, final *m*, and final *q* are added after hooked resh.

[b] Hebrew name for the ancestral Aramaic letter.

[c] In syllables other than the first.

Figure 84: Table showing letters of the Uyghur script (from Kara 1996: 540).

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 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: *2017-01-25*
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: *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: *42*
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? *Yes*

If YES, is a rationale for such use provided? *Yes*

If YES, reference: *Combining characters for diacritics*

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: