Preliminary proposal to encode Sogdian in Unicode

Anshuman Pandey pandey@umich.edu

May 9, 2016

1 Introduction

This document presents a draft encoding for the Sogdian script in Unicode. The Sogdian script discussed here is the 'cursive' or 'sutra' script. Although related to 'Old Sogdian' it is proposed for encoding as a separate block on account of its structure (see L2/15-089R). The 'Sogdian Uyghur' script may be a candidate for unification with the present script, but further research is required regarding this matter.

The preliminary encoding for Sogdian is intended solely to provide a foundation for discussion. It is not a formal proposal. The character repertoire and character names are tentative. The representative glyphs are merely illustrative and are not intended to be typographically aesthetic; they are normalizations of diverse handwritten styles. Feedback is from the scholarly community is requested.

2 Script details

Sogdian is an *abjad* that is written from right to left. It is a structurally conjoining script similar to Arabic and Mongolian. As in these *abjad* systems, letters connect and change shape based upon their position within a word. Words are separated using spaces.

The script consists of 17 basic letters. Diacritic marks are used for indicating different values of base letters. A set of numbers are also attested. Signs for delimiting various sections of text are used.

The script is written both horizontally and vertically. In horizontal mode, the script is written in lines that proceed from top to bottom. In vertical mode, the writing direction is rotated 90° and lines proceed from top to bottom from the right edge of the writing surface towards the left.

2.1 Letters

Each letter is shown below in its isolated and positional forms. The alphabetic order is based upon Aramaic. The final form is tentatively selected as the isolated form because it is the most distinctive of all the forms. As is the case for structurally similar scripts such as Arabic, only the isolated form would be encoded. The the positional forms would exist only as glyphs in the font and the rendering engine would perform the necessary shaping. Although initial and medial forms may not be palaeographically distinctive, these are differentiated typographically in order to support the joining features of the script.

	Character name	Value	Final	Medial	Initial
×_	SOGDIAN LETTER ALEPH	', a, ā	×-	x	×
٦	SOGDIAN LETTER BETH	β	٦	r	3
У	SOGDIAN LETTER DALETH	δ	У	Х	2
ح	SOGDIAN LETTER HE	h	ح		
٩	SOGDIAN LETTER WAW	w, u, ū, o	٩	٩	۵
÷	SOGDIAN LETTER ZAYIN	Z, Ż	÷	÷	,
~	SOGDIAN LETTER HETH	γ, x	~	*	N
د	SOGDIAN LETTER YODH	y, i, ī, e	د	۲	د
6	SOGDIAN LETTER KAPH	k	6	ىر	ى
ゥ	SOGDIAN LETTER MEM	m	ゥ	ち	さ
4	SOGDIAN LETTER NUN	n	L	•	•
¢	SOGDIAN LETTER SAMEKH	S	ę۶	9	4
ھا	SOGDIAN LETTER PE	р	ما	ى	ى
Ŀ	SOGDIAN LETTER SADHE	c, j	لد	E	~
×	SOGDIAN LETTER RESH	r	×	×	У
×	SOGDIAN LETTER SHIN	š	×	~	8
4	SOGDIAN LETTER TAW	t	لم	b	Ь

All letters except for the following join to the left and right edges of adjacent letters:

- HE This letter is used only in final position. It does not have initial or medial forms.
- **G** KAPH This letter joins to the right in medial and final positions. The top of the curve joins to adjacent letters in initial and medial positions.
- **(** $_{\text{PE}}$ This letter joins to the right in medial and final positions. The top of the curve joins to adjacent letters in initial and medial positions.

2.2 Phonetic modifier signs

The following signs are used for indicating different pronunciations or phonetic values of letters:

	Character name
ੇ	sogdian modifier mark-1
ੁ	sogdian modifier mark-2

For example, in the Sogdian transcription of the *Nīlakantha Dhāranī*, these marks are used with base letters to represent non-Sogdian sounds:

ĥ	h	м нетн + Э
ŗ	1	🗶 resh + 🕤
ĥ	ţ	k _{TAW} + ∂

2.3 Punctuation

_

The following marks of punctuation are used in various sources for marking phrases and end of text sections (see figures 3, 4, 5).

	Character name
	sogdian punctuation mark-1
й	sogdian punctuation mark-2
o	sogdian punctuation mark-3

Some punctuation signs occur in combination (see figure 5).

2.4 Abbreviation

The sign **(b)** occurs in a few sources as an abbreviation for an Aramaic heterogram (see figures 2 and 5). It may be an embellished form of the letter *ayin*, which is not distinctively represented in Sogdian. It is a connecting character, as shown in figure 5. It may be appropriate to encode this sign as a letter.

2.5 Numbers

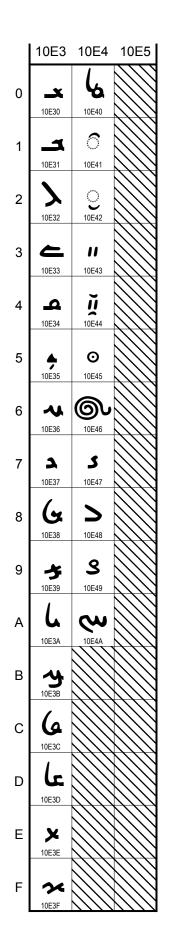
Numbers are attested in Sogdian manuscripts:

	Character name
٢	SOGDIAN NUMBER ONE
2	SOGDIAN NUMBER TEN
٤	SOGDIAN NUMBER TWENTY
~	SOGDIAN NUMBER HUNDRED

See figures 2 and 5 for examples. Numbers in Sogdian follow the same pattern of numeration as in the 'Old Sogdian' script (see L2/15-089R).

10E30

Sogdian



Representative glyphs are based upon the 'sutra' style.

Letters

10E30	×	SOGDIAN LETTER ALEPH
10E31	r	SOGDIAN LETTER BETH
10E32	Σ	SOGDIAN LETTER DALETH
10E33	ھ	SOGDIAN LETTER HE
10E34	٩	SOGDIAN LETTER WAW
10E35	4	SOGDIAN LETTER ZAYIN
10E36	Å	SOGDIAN LETTER HETH
10E37	د	SOGDIAN LETTER YODH
10E38	6	SOGDIAN LETTER KAPH
10E39	5	SOGDIAN LETTER MEM
10E3A	t	SOGDIAN LETTER NUN
10E3B	4	SOGDIAN LETTER SAMEKH
10E3C	Ģ	SOGDIAN LETTER PE
10E3D	عا	SOGDIAN LETTER SADHE
10E3E	×	SOGDIAN LETTER RESH
10E3F	ሯ	SOGDIAN LETTER SHIN
10E40	6	SOGDIAN LETTER TAW

Various signs

- 10E41 [©] SOGDIAN MODIFIER SIGN ABOVE
- 10E42 g SOGDIAN MODIFIER SIGN BELOW

Punctuation

- 10E43 " SOGDIAN PUNCTUATION MARK-1
- 10E44 ğ SOGDIAN PUNCTUATION MARK-2
- 10E45 o SOGDIAN PUNCTUATION MARK-3

Abbreviation

10E46 SOGDIAN ABBREVIATION AYIN

Numbers

- 10E47 **s** SOGDIAN NUMBER ONE
- 10E48 SOGDIAN NUMBER TEN
- 10E49 S SOGDIAN NUMBER TWENTY
- 10E4A 🛯 SOGDIAN NUMBER HUNDRED

Printed using UniBook™ (http://www.unicode.org/unibook/)

440 Pelliot Soga . S xxxx0 30 4x15460 guban. LANDER S 40 399× 1 - 0 -25 94 IN 999 - AKAR 9920 GRAD 94 9450 9130N 360- 2549951 900 240 5000 5 40 40 QKASK 1 194 E MSAN 04 Y MARCO 6.90 1240 69 4 the best are warded and 990 6 969 144 944 M QSKAM 14 305 RKA 200 XAS. 2994× 300 14 9 an 691 24 6 500 LOE WAKO QUAK 9 100 1 KAS OSENNO

Figure 1: Excerpt from the Sogdian Vessantara Jātaka (Pelliot Sogdien 1).

Anshuman Pandey

and are to age present age for a star and this arrive par ses promoved & 60 126 6 my lander page pre ----624060 lang Jug se 0 051 34 254 es lans UN QUO SA 36. + Just week (armore exe ligger 0 510 HORMANN AGUE 34 26 4 JAG LOKES 613450 -24 24 0 440 60 ley 19 31 is buch year sos and the seal o made man gues was she been any bruns 404 shing shere burges HORMA MAM 36 446 surged (while see you buyged show

Figure 2: Fragment of the Sogdian *Saṃghaṭa Sūtra* (So 20165 r). The abbreviation (occurs in lines 6, 11. The number علوه '1000' occurs in the word علوه *ILPw* in line 2.

GAN NECESCALII AC NEW ANE ANE TELESCHOOKNEN! Annow Chesing Swith - sig " Same A sucres " Loto YOUR TEAN ENDER ENDER विद्य रुति र किम वरं : का म द मं: लि का द मंथि। AT MAL GAS PASS AND AND STAT FORMER (भद्रसम्बदान्द्र अवियंत्रिमक सिर्धेने : वर्ग בריין אילאאין ברבן היהן גיעליייבא "שכט בכט שיט ביט वर्रेषु मरः रावविमवा मध्य अद्र : सुर य व कि म הראוקר גראים הראוקר קא קא קצאר איקהארעיבלו גל ECENER SA GOORDES WEEKERSALES मा रिम्म हरेश्व के यम यम यम यूम यूम (भ द्व द्व दा वा र र य भ र य थ र य भ אין איצעיים אל אל אל אלים יידים אבי אל אל אלי אריביים בריי אל אל אלי איבי אריי איבי אל אל אלי אליי איבי איבי א オペシャダイマをみずまないまでしまいろうこう - it - it same - it berny way berny bis " anthrong

Figure 3: Excerpt (3/9) of the $N\bar{l}akantha Dh\bar{a}ran\bar{i}$ (BL Or.8212/175). The Sogdian text is to be read by rotating the 90° clockwise. The consonant modifier signs $\hat{\circ}$ occurs in line 13 and $\hat{\circ}$ occurs in lines 1, 10, 11, 13, 15, 17.

Anshuman Pandey

क के में याग म में भी उवा गा यग म ANTOCH Cardens of the बार आग देश ठ में जुर दे मुरि म रवि 506999 त्रमुरः दु मुजाँ भाकृ Greenelicer anos waters brok erest when Podi 3 स म ग

Figure 4: Excerpt (8/9) of the $N\bar{\imath}lakantha Dh\bar{a}ran\bar{\imath}$ (BL Or.8212/175). The Sogdian text is to be read by rotating the 90° clockwise. The punctuation sign μ occurs in lines 2, 4, 8; the sign $\tilde{\mu}$ occurs in lines 7, 8, 10.

many chees amored ent ades 195005 1910345 462000 5)91 2000 06309 PELLIOT SOGDIEN 149 95 1595 200 95960 90 1000 956au 1409 6 singo grader so -----NHAN 9 135 51 3 930 957 9 Las = 3 965 1 - stands was h ALKASO and and and and 30 39.000 and a ward and a property of the property 4× 410 400 ×9 ~ +94 no represed room 6000 6000 and syges syntage heras seeper see you'd the wind accepted sound a seeper ero so and there are a source of and annous the 13430 90 as an edge so and any second as soon who have and so where and a star your present south a prese Kang give skage for so beause of shop gabou Olibrary of anosta GR is sanceres any 3009 -10 15 3 5 3 63 695.90 Long 4569 and real adde of and mer convert an same a series Day soon bargeres 25000 800 4000 69 anger and fre serve mon and see and les an ere soon been 100 go envos parger Que ty a muric becan sear because good source but seconer 9 and bere brease sore constant a bog ong george suger and for and social man has

Figure 5: Excerpt from Pelliot Sogdien 6. The abbreviation (occurs in line 18. A sequence of the punctuation signs $\breve{\mu}$ and \odot occurs in line 17. The number \checkmark '100' occurs in lines 23 and 24; the number \checkmark '50' occurs in line 23.