Developing a Success Proposal for Encoding a Script in Unicode

149 35 X

Internationalization and Unicode Conference #42 11 September 2018

04.8 374

Anshuman Pandey Script Encoding Initiative, UC-Berkeley

About me

- Product Strategist, Avant, Chicago
 - Fin-tech analytics
- Researcher, Script Encoding Initiative, UC-Berkeley
 - Develop Unicode standards for writing systems of the world
- Education
 - Post-Doctoral Researcher (Linguistics), UC-Berkeley, USA
 - PhD (History), University of Michigan, Ann Arbor, USA

About me: Unicode

- Developed encodings for
 - 24 scripts
 - 4 number systems
 - various individual characters
- Work in progress
 - 25 script proposals
 - research on 100+ scripts

About me: Unicode

- Developed encodings for
 - 24 scripts
 - 4 number systems
 - various individual characters
 - 2 emoji
- Work in progress
 - 25 script proposals
 - research on 100+ scripts



diya lamp



auto rickshaw

(tuk tuk)

Why?: Enable Digital Humanities Foundations

- Representation of 'visible language', eg. script / writing system
 - Often ignored in traditional and applied linguistics and language studies
 - Key to access and presentation of textual sources in the original
- Display and input of language (keyboard, stylus, etc.)
 - Enable users to input, search, and view text uniformly on devices
- Natural language processing (NLP) and machine learning (ML)
 - Information extraction and analysis using programmatic methods
- Goal: all languages as digitally native as English

'Full Stack' Digital Support for Languages

- Character-encoding standard
- Locale data
- Font / typeface
- Keyboard / input methods
- Text layout and formatting
- Database support
- System and browser support
- NLP tools and algorithms



'Full Stack' Digital Support for Languages

Character-encoding standard

- Locale data
- Font / typeface
- Keyboard / input methods
- Text layout and formatting
- Database support
- System and browser support
- NLP tools and algorithms



Script Encoding in the DH Stack



1210 E gus gus 232 052 002 541 533 Casse 36.72 4025 7230 43 6550 your person and march areas

24 67

13 36'

70 =

Script Encoding Process

1800

1 705--519

ZPERK

23 730

inn.

3220 7

allow a

Script Encoding Process: Overview

1. Identification: Native users, scholars, others identify a script not yet encoded in Unicode



- 2. Development: Research script and develop script proposal
- **3. Review:** Unicode Technical Committee reviews proposals; may request changes; vote to approve or disapprove
- 4. Publication: Publication of script in Unicode standard
- 5. Implementation: Create fonts, keyboards, update software (Noto project; individuals)
- 6. Iterative development: Repeat for new character additions

Script Encoding Process: Focus

- **1.** Identification: Identify a script not yet encoded in Unicode
- 2. Development: Techniques for researching scripts and models

the set of the state of the set of the set

Script Encoding Process

120

22220

AUN

Identification

2.20

730

Script Encoding Process: Identification

- Existing character-encoding standards
- Request from a native user community
- Proposal from scholarly user community
- Submission from writing-systems researchers and enthusiasts
- Fieldwork and archival research

Identifying a script

حساب جنسی سیاقی(ترتیب دهدهی)

Planned: Persian Siyaq

``		1+++_9+++) * * - 9 * *	۱+- ٩+	۹–۱	1+-9+	۹-۱	
خروار	خروار	خروار	خروار	خروار	خروار	من	من	ش
، ۱ <i>و</i>	•••	الن	66	a		4	• 6	١
•• /	•• <u>المسار</u>	ا عمين م	•.,	عــــا	3) •	d Y	< /۱	٢
٠. ٢	••	••,	ن ا ہے۔۔۔۔ (\square	ىر) • •	م	دى•	٣
•• ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	•••	لىر يى	•• ~~~~ 81	J	••6	0	•81	۴
· · · · · · · · · · · · · · · · · · ·	•••~~	مريد	6.,		••60	مـــه	60	۵
٠.,	•••	•., ²	6. <u> </u>	<u> </u>	•• (۹	•4	۶
<u>د.</u> الا	•••••••	•.,	••,8	$\bigcup_{i=1}^{n}$	• • 8	d_1	•8	v
•	•	•	•• ، ۔۔۔۔ ن	$(_)$	•••	4	٩U	٨
٠٠/ ٢٠	•,	•••	··ک سر	د م	د.(ک	مح	•8	٩

Identification

ب جنسی سیاقی(تر تیب دهدهی)

				Pa	10 60 ED	- sounde	19 60.00	Altread in	phase anotes to fice
۹۰۰,۰۰۰ – ۱۰۰,۰۰۰ خروار	۹۰,۰۰۰ – ۱۰,۰۰۰ خروار	۹۰۰۰۹–۱۰۰۰ خروار	۹۰۰ – ۱۰۰ خروار	- Lu	se sul	الله حصم	0 400	- 24	and a services and and and
<i>د</i> . ا	••ام	الس،	6 6	daug	Main	EPty	المنحر	Lus	and the sale way and a sail
··/	السيسر،،	ا عن الم	•. ,	+00	· mea	loover	4.	per-	Derrad Congegues of
٠	•••	س ب	6. , u lu	por 1	your	colono	lennon	-	and a cer par part in the same provide
•• ~~ 181	•••	لىر،	•·,	300	care,	1 months	ane. mid	696	er sen son an or a land
«. <u></u>) ده	•••~~	مريد	6.,	C	re lu	Contraction of	male	71,545	Made In Holes where there
·····	•• ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	•., <u>-</u> -	6L	HANDS	Janua	+ les	10000	here	en la des la ser la ser la ser
··/18	•••••••	•. /	•• , 8	\square	••8	01	.8	Y	
• / الـــــــــــــــــــــــــــــــــــ	•	•.,	ن ســــر ۵۰	\square	••CV	<u></u>	۶Ü	~	Unplanned: Sogdia
··/	••	·.,	ىكى ىسىر	<u> </u>	••	<u>م</u>	• 8~	٩	
	السی المی المی المی المی المی المی المی الم	۱۰۰,۰۰۰-۹۰۰,۰۰۰ ۱۰,۰۰۰-۹۰,۰۰۰ یاویخ ۲۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰۰ ۱۰٫۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰	$1 \cdot \cdot , \cdot \cdot - q \cdot , \cdot \cdot \cdot $ $1 \cdot , \cdot - q \cdot , \cdot \cdot \cdot $ $1 \cdot \cdot - q \cdot \cdot \cdot \cdot $ $i < j < j < j < j < j < j < j < j < j < $	$1 \cdot \cdot, \cdots - q \cdot \cdot, \cdots$ $1 \cdot \cdot, \cdots - q \cdot \cdot, \cdots$ $1 \cdot \cdots - q \cdot \cdots$ $1 \cdot \cdots - q \cdot \cdot$ $j = j = j$ $i \cdot - \cdots - i = i$ $i \cdot - \cdots - i = i = i - i - i = i - i - i = i - i -$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	الما محمد على المعالي المحمد المحمد على المحمد المحمد على المحممد على المحمد على المحمد على المحمم على المحمد على المح	

service and a support of the second and the second and the second second

ورمدا وحدوي ودونهم بوتعامد عدوتهم

meser bring have ner was bring and and

Has building House per building

and a stand a stand

and had

100 10 500

BELLEVILLE &

State.

service and so and an income she par the set and the Identification ALL DESCRIPTION OF THE OWNER HE HOL YOU SHOLE PHONE DA GOODENES hasibasi (Baiteital) andiana and the base bracker bas Street Street No. of Concession, Name ب جنسی سیاقی(ترتیب دهدهی) Presence 0 الرواسية التحا الرا المحطة الدما 1.4 1 * * , * * * - 9 * * , * * 4 "Co. ballance a state a second BALLA ACLE The state state and the state of the state of the state and the state of the خروار Planned: Persian Siyaq - HAR CLASSINGUS AUNT OF ST YER D OF THIS HEAR HISS ARE SIGHT DELLES בוצא צניא צויא בא כד ליצום פרממאציה צארא בנייז יים פרוים ל בזראל אאנא גריבאל צו עניצ שירבי מזכי אנאר יאנטרא נוב 84 L800-0 1 95305 y we we was we was we we we 450005 14. 5770 tout . 10 75-519 75 7: 4753 2020 JOCHAC . 24 Zever 210 Robin isi to go have 205 4500 set sites 2434:2553 Perer متعاوج مجيبا مجتملتين 15715575 3 به الت 1 45 te 150 100 5) فخوبتمو عج فجاسا -2181 335 h 7250 1945 - 3 420 420 WHERE HE RE State States yess mere 1437 15 er . 105. 1500 142 44: me seed LONG IT ONSE 55 - 105500 le acer press men 1000 15 mg 4755 Actual Advantation and well TWINGE & AWE # 6. ... PSYND, DON SACH TOSE ON THINK IS TANK CE : msen 5.515,--11-ALYNS SHE 200 and and التصاوية الداول فتصديعه ONH. LON THEN & CE 12 24 MANY PLAN ייברצצ ניתל צבל ניקל ניאצ צאר אניצוני באיניים ניך ADMED . ک ر 1000 2012 2012 1200 200 200 200 2005 4:05 4:05 4:05 EN 105 BELL GALLE GEL THE **Unplanned: Sogdian** ·~~~_______() was by or salvery serve

Unplanned: Old Sogdian

Identification

ب جنسی سیاقی(ترتیب دهدهی)

Cannon sales and an investigat have not and the ALCON O SCIENCE Upe Has muchange House par burgers ونو ويسمعا حمما حرب بداء مسا ويشمعها سجعد States and the second second 100 10 10 10 10 10 las State. الهميدة الالتلية ومتية والمتقاصية فيلاله and as come and the time of the second of the

Planned: Persian Siyaq

נאד באס אבמצה באזיני בעי עולים אדונגים בציאה מוצמואני היעש אציא ני כהדע באנם אנוציי אבי אלא גר אני איני איני עוציי ני עבר הבמאים אנוגאני איני ב خروار you & your the site and the subst bridge mursysse a sets mans funs eyers iss stipt and as an percenter צארא בנייז יים פרוים ל בזראל אאנא גריבאל צו sores aree are show show LIL YSJOS you the see wesses was not we 455425 12 25-319 7: 4753 2020 DEMAE 124 Zne Robin in 210 Kon to have 725 4500 set sites systemetre 137105075 •••••••• 1 45 te 75 x x 1005 -11 3 5 250 420 420 yest mere 25. -64 Jun abse 22 7100 -6 4755 usen 5335-2201- -1201 028 24 LON MER & G ישרצצעראל צבל נצל נאצ בארי ant 2012 2014 4105 4105 000 MIL GERED 5-20 7200 was loted by second Se **Unplanned: Old Sogdian**

1.4

1 + + , + + + - 9 + + , + + +

Unplanned: Chorasmian

Unplanned: Sogdian

had be and the

10.20 St. 23.00

to said a las para ser " has been

COURSE AND AND AND ALL AND OF

State States

حصوا فروة تصنعوا بدلا مع الادية عد

مصاوع متهدينا محمصغ متوا

HOND BEINGS THE LEWISSING

State Ball and

COLO HOLE HADE D

No Children

NOT STATE OF STATE

Case Study: Old Sogdian

- Consonantal alphabet
- Used for writing Sogdian
 - now-extinct eastern Iranian language
- Derived from Imperial Aramaic
 - Variation of Achaemenid administrative script
- In use from 3rd 7th century CE



Old Sogdian: Identification

- Known to scholars
 - 19th century to present
 - active scholarship, publication trends
- Digital humanities content
 - TITUS project: transliterations
 - UW Silk Road Project: translations
- Manuscript digitization
 - International Dunhuang Project (IDP)



Situating the Sogdians



Old Sogdian: Identification



Old Sogdian: Identification

Sogdian Ancient Letter No. 3

[Verso] From (his) daughter Shayn to the noble lord Nanai-dhat.

[On another part of the verso] From (his) servant [left unfinished].

[*Recto*] To (my) noble lord (and) husband Nanai-dhat, blessing (and) homage on bended knee, as is offered to the gods. And (it would be) a good day for him who might see you healthy, happy (and) free from illness, together with everyone; and, sir, when I hear (news of) your (good) health, I consider myself immortal!

Behold, I am living ..., badly, not well, wretchedly, and I consider myself dead. Again and again I send you a letter, (but) I do not receive a (single) letter from you, and I have become without hope towards you. My misfortune is this, (that) I have been in Dunhuang for three years thanks(?) to you, and there was a way out a first, a second, even a fifth time, (but) he(!) refused to bring me out. I requested the leaders that support (should be given) to Farnkhund for me, so that he may take me to (my) husband and I would not be stuck in Dunhuang, (for) Farnkhund says: I am not Nanai-dhat's servant, nor do I hold his capital. I also requested thus: If he refuses to take me to (my) husband, then ... such support for me that he may take me to (my) mother. The leaders say: Here in Dunhuang there is no other relative closer than Artivan, (but) Artivan [say]s: Farnkhund ... whatever ... to do for you. If(?) I(?) (had) no guarantee, no protection, my father ... I have become ... not ... How much more would I have ... by my father if ... a servant of the Chinese! A free man ... who found ... and ... keeps (his) clothing in good condition(?). And you write (your) bidding to me about everything in ... so that I should ... you and I should know how to think, and if I do not ... you, then you write to me so that I should know how to serve the Chinese. In my paternal abode I did not have such a restricted ... as with(?) you. I obeyed your command (lit. took your command upon my head) and came to Dunhuang and I did not observe (my) mother's bidding nor (my) brothers'. Surely(?) the gods were angry with me on the day when I did your bidding! I would rather be a dog's or a pig's wife than yours! And for me ...

Sent by (your) servant Miwnay. This letter was written in the third month on the tenth day.

[*Added in the margin*] From (his) daughter Shayn to the noble lord Nanai-dhat, blessing (and) homage. And (it would be) a good [day] for him [who] might see [you] healthy, rested (and) happy. ... I have become ... and I watch over a flock of domestic animals. Differently to you, I had a ..., and ... went out. I am ... and I know that you do not lack twenty staters(?) to send. It is necessary to consider the whole (matter). Farnkhund has run away; the Chinese seek him but do not find him. Because of Farnkhund's debts we have become the servants of the Chinese, I together with (my) mother.

Script Encoding Process: Identification







Old Sogdian: Identification



Aramaic	Sogdian Ancient Letters	Sogdian sutra script	Manichean Sogdian	Christian Sogdian	Principal Phonetic Values (Sogdian)
>	2	ه , ه	ч	2 🕶	a, ā
b	5	ه , ه	4	-	b , β
(β)			ü		β
g	**	*	4	2	g, γ
(γ)			Y	>	γ
d	لا		7. 7	,	d, δ
h (<u>h</u>)		E	7	G 77	a, Ø
w	0	4, 0	•	0	w, ŏ, ŭ
z		٠	5	۲	z
(j)			2		ž
(ž)			Š	v	ž
ḥ (h)	на	h , n	بد		γ, x, h
ţ			e	x	t
У	4	ه. ه	•	•	y, ĕ, ĭ
k	2	و,ما	-	9	k
(x)			ف	ડ	х
1 (δ)	2	1,1	<u>a L</u>	17	δ
m	ولا	*, *	et ev	7 3	m
n	د	L., .	5 .	1	n
s	ود	34, 3	39	<u>.</u>	S
c	5	6	7	>	Ø
р	و	ى		٩	р
(f)			خــ	æ	f
ș (c)	و	*	Se	S	č, j
q			בל בט	35	k
r	"	7, 5	÷ .÷	ì	r
š	مهو	2,2	ω	X V	š
t	ور ول	6, 5	k	1	t, θ

Sogdian script

In the Sogdian script used in the "Ancient Letters" (TABLE 48.2), most of the letters are distinct and do not change shape when joined. In the "formal" and "Uyghur" Sogdian scripts, most of the letters are joined and, owing to the use of a broad pen, are frequently difficult to distinguish. In the earlier form, 'i s still distinguished from n; but in the later, '= n, 'n = n'. Some scribes distinguish z from n by not connecting z to the preceding letter, but others make no distinction. In the later, increasingly cursive, form, other letters tend to become indistinguishable as well: $\gamma/x/s/s$, $r/\beta/y$. Some letters are distinguished only in final position (by some scribes), e.g., $n \sim z$, $x \sim \gamma$.

z is sometimes distinguished from **n** or **z** from \check{z} by a diacritical point, and the foreign sound b was noted as \check{j} **p**.

SAMPLES OF SOGDIAN

ANCIENT LETTERS

			ىلاو	צאצרר	nelezzeo	צנצרנס	ercer 50	ony	55	\rightarrow
			PLI	kk'n'k	wr [»] βδynn	kk'rβ	w'twx	wγβ		DO←
		012244	y 4	עצונאו	4.745	وعطراعدوه	سوده ور	× 142	ممورة	37533
		wn ^{››} γ	βwM	XyKZ	YZKYA	ykwn'zt'	ps wych	mn	MLŠ	rwyrβ
					بودوريه	ענפרה ווא	Less -	دور ودر	4040	b754
					ktnβyı	nn ktnβ	δрух Ν	IM	tšyp	tryβ
Ι.	Translitera	tion:	OD	βγw	xwt'w	ßr'kk	nnvδß	^{oo} rw	k'n'	°kk
2.	Normalizat	tion:	at	βαγμ	xutāw	Barak	nanē-	θβār	kan	ak
3.	Gloss:		to	lord.AC	c master	Barak	Nana	s-gift	Kar	nak
	- I D	0		ČI M			24 2 2			WW
Ι.	ILP	prywr		SLM	nm'c	yw	spitzi	iwky	AY	KZY
2.	(ēw-)zār	βrēwa	r	*āfrīw	van namā	icyu	spātzā	inūk	kað	-uti
3.	thousand	ten.the	ousand	d greetin	ng(?) rever	ence.AC	c bende	d.knee	e whe	en-that.and
Ι.	ZKyXMw	βγ"n	w	βyrt	pyšt	MN 2	κγρθ βι	ntk	nnyβ	ntk
2.	wēšanu	βaγā	n(u)	Byart	pišt	con y	xēpθ β	antē	nanē	-Bantē
3.	them.OBL	lords	.OBL	received	d written	from o	own se	ervant	Nana	's-servant
5										
	To the Di	wine M	laster I	Barak(?)	Nanethvar	Kanak a	thousand	d, ten t	nousa	nd
	greetings,	revere	ntly w	ith bende	a knees wr	ien recer	ved by th	ieir div	inities	s. writ-
	ten by his	own se	ervant	Nanevan	te.	· · ·			~	
	-From th	e Old S	ogdia	n "Ancie	nt Letters"	found in	a mailbe	ag in th	ie Gre	at Wall
						(AL II, I	Reichelt	1931: 1	12 ana	l pl. 2).

Script Encoding Process: Identification

- Challenges
 - Charts or specimens not in published in books on writing systems
 - Studies may be outdated or very limited
- Resolution
 - Conduct research: primary sources!

the set of the state of the set o

Script Encoding Process

1200

22220

MANT.

Research

2.22

730

Old Sogdian: Research (2015)

Roadmap to the SMP

Revision	8.0.1
Authors	Michael Everson, Rick McGowan, Ken Whistler, V.S. Umamaheswaran
Date	2015-08-17
This Version	http://www.unicode.org/roadmaps/smp/smp-8-0-1.html
Previous Version	http://www.unicode.org/roadmaps/smp/smp-8-0-0.html
Latest Version	http://www.unicode.org/roadmaps/smp/

0 1 2 3 4 5 6 7 8 9 A B C D	E	F
-----------------------------	---	---

108	Сур	Imp.Aramaic Palmyrer			yrene	Nabataea	an ???	¿Numidian?	Hatran		
109	Phoenician	???	???	???	???	Meroitic H.		Meroitic Cur	Aeroitic Cursive		
10A		Kharoshthi		O.S.Arabian			O.N.Arabian	(Balti)	Ma	Aanichaean	
10B	Aves	stan	Part	rthian Insc. Phlv.		Psalt. Ph	Iv. (E	Book Pahlavi)	(Old Sogdian)		
10C	O	d Turkic		(Baburi) ???				Old	Hungarian		
10D	Pontr	ngya)		(Garay	(Wolof))	???		¿Byblos?		
10E		¿Sogdian?			Rumi	Rumi Symb. ¿Uyghur?				¿Elymaic?	
10F			1				???				
110		Bra	ahmi					Kaithi	S	ora Sompeng	

Old Sogdian: Mistaken identity

y a set and a set as sugar as a for and the set and red state of andread and star and state of andread and siesial and conserved can and any ind present ment وولعه بعا لجد موجد عد ومد و وعما معا معد عما معا وجريده as you and and the second deside and a second and موج مع ومديع معدومة مسطة ومة و محمد مع و و جمع معويدم وهد معدد مع و مود وي موجه مديد مديد مديد عد محد way as a base soft feat soft and and and and وجمعد عويد مديد مديد مديد مديد معد موجد عويد معد In was for any and for was for may and part and any you want and the base or and a part from the too and the sound and and and the second and the lawyers معج موجورا موجود معرومة ومسوما وجمعة مدمدمعار موجود ومما ومد ومنعو المد وما مع الم عود المحد مع المحد مع

Old Sogdian: Mistaken identity



Old Sogdian: Unicode Status

14 *Aramaic* forms a rather complex family of scripts, with a number of descendants. Certainly there is a basic Aramaic, but it has many descendents (including Mongolian and possibly Brahmi) which are unique enough to merit their own encoding (see table 5.5). More research is required. However, Aramaic is expected to encompass at least: Aramaic proper

Middle Persian Parthian Sogdian

Everson, "Roadmapping early Semitic scripts", 2001.

8	Old North Arabic (see table 5.6), which	TABL	ь 5.6: Nor	th Arabic Scri	pts (Garbini	979, fig. 9)
	encompasses:	_	XXI	XXII	XXIII	XXIV
	Dedanite	,	ň	17	ň	I
	Deutinite	ь	Π	17		nc
	Lihyanite	g d	2	2	a	4
	Thamudic	đ	щ	Ť	H	ψ¥
	Safaitic	h	1111	6	00	θœ
	Sujunio	z	H	Υ.	ΨΨ	I H Y A
		b b	Ω Σ	32	• •	xx
AC	uneiform script:	1	Ē	7	,	S.
		Ţ.			*	÷
9	Ugaritic	У	Ŷ	9	r H	1
-		k 1	LT	2	1	2
		m	1 2 2	ภ	88	20
No	rthern Linear scripts:	n	54	2	2	ŧ١
		s	th .	ŲĢ4	ή	ń -c
10	Nahataean (see table 5 5)	<u>.</u>	0	0	00	3
10		g f		174	0 0	ĉ
		s	4	â	810	Я
11	<i>Palmyrene</i> (see table 5.5)	đ	8	m v	8	дĦ
		q	¢	\$ 4	ķ	4 4
12	Hatran/Armazi (used in Armonia and	r	>>	>>) 7	2 5
14	manania (used in Armenia and	š	3	3		+
	Georgia)		x +	â	* *	9

13 Elymaic

Col. XXI, Dedanite; col. XXII, Late Lihyanite; cols. XXIII-XXV, Thamudic (XXIII, (XIV. Heiaz: XXV. Tabuk): col. XXVI. Safaitic

5

80

XXVI

ĭΪ

20

XXV

пс

×

FY

2

Л ns-

CTV

	scripts, with a number of descendants.					
	Certainly there is a basic Aramaic, but it has					
	many descendents (including Mongolian					
	and possibly Brahmi) which are unique					
	enough to merit their own encoding (see	TARI	F 5 5 Sering	Derived from a	vanale Seriet (Carbini 107	0 60 71
	table 5.5) More recearch is required		XVII	XVIII	XIX	9. Jig. 71 XX
	table 5.5). More research is required.	,	~ *	* *	** 4 4 4	1
	However, Aramaic is expected to encom-	h	ร้วว	בב	CTL RE	ć
	pass at least:	g	x	7 4	۲۲ ۲	7
	Aramaic proper	d	4 1	4 4	44	>
	Middle Persian	h	ti	XX	17770	-16
	Parthian	w	71	~ 2	999	٩
	Sogdian	z	1	1 1	11	_ ጉ
	eoguiun	b	нп	жж	им у	>
	Edesson is likely to be either Aremaic or	ţ	0	00	6	Ъ.
	Edessal is likely to be either Aramaic of	y	1'		444166	2.5
	Syriac. More research is required.		, 27	2.5	155557	د ک
			5	n n	1100	
15	Phoenician is the catch-all for the largest	 n	11	1 4	115	L L
	group of related scripts including its	8		5 4	97 97 n	_
	ancestors, Proto-Sinaitic/Proto-Canaanite.		v v	y y	1714X	ж
	Looking at tables 5.1, 5.3, and 5.4 (below)	р	12	3 3	و و و	9
	most of the scripts are so similar that there	ş	ر ر	۲	م م	مر
	doesn't coom to be any point in trying to	q	L b	ממ	1 239	e
	ubesit is seen to be any point in trying to	r	5	47	3 7 7	ر
	encode them separately.	š	6 6	22	* * *	

14 Aramaic forms a rather complex family of scripts with a number of descendants

> hл h + a. Col. XVII. Hebrew square script: col. XVIII, Palmyrene script; col. XIX, Nabatea col. XX. Ancient Arabic scri

Sogdian: Unicode Status – Unification?

Roadmap to the SMP

Revision	8.0.1
Authors	Michael Everson, Rick McGowan, Ken Whistler, V.S. Umamaheswaran
Date	2015-08-17
This Version	http://www.unicode.org/roadmaps/smp/smp-8-0-1.html
Previous Version	http://www.unicode.org/roadmaps/smp/smp-8-0-0.html
Latest Version	http://www.unicode.org/roadmaps/smp/

	0		1	2	3	4	5	6	7	8	9	Α	В	С	D	E	F
108	Cypriot In				Imp.A	Aramaic Paln yrene			Nabataean ??? ¿N			¿Nun	umidian? Hatran				
109	Phoe	*		Lyo	dian	???	???	27.1	???	Meroitic H. Meroitic Cursive						sive	
10A	Kharoshthi						O.S.Arabian			O.N.A	rabian	(E	Balti)	Manichaean			
10B	Avestan Parthian Insc. Phlv.						Phlv.	Psalt. Phlv. (Book Pahlavi) (Old Sog					ogdian)				
10C			Olo	d Turk	ric .		(Ba	ouri)	???	Old Hungarian							
10D			Cohin	gya)			(Garay	(Wolof)))	??? ¿Byblos?							
10E	¿Sogdian? Rumi Symb.							Symb.	: Uyghur? ¿Elymaic)						naic?		
10F					?	??				???							
110	Brahmi									Kaithi Sora Sompeng						eng	

Script Encoding Process: Research

- Analysis of sources to document repertoire and script grammar
- Compare grammar and orthography of related scripts
- Outreach to user community to understand current usage

Sogdian sources

- Kultobe: the oldest Sogdian records are stone inscriptions found at Kultobe in modern Kazakhstan
- 'Ancient Letters': the earliest attested Sogdian manuscripts are known as the 'Ancient Letters' (c.312 CE). These paper documents were found in 1907 by Aurel Stein in Dunhuang, western China
- 'Upper Indus Graffiti': appears on more than 600 rock carvings at Shatial and other sites in the Gilgit region of Pakistan (4th-7th c. CE)

Old Sogdian: Research



Old Sogdian: Research

באזדיני בעי עציעי אדעניני עלדינאי איניגעוואני איני איני איני איני איני 23 570 2 200 50 yes to asur 35 71 705 750 כמינים בופגיאר ער איני לי לנקשים במאר אראו אינים אלגאביו שאיניא REATHER ביישל אוני גער איישיאל אונאגעיי איישי איישי איישי איישיא אונא איישי איישי איישי איישי איישי איישי איישיא איישיא לציב פבממדיי שאניה כבי א אי לעום ל בזיצא אונה וניבואני דוים معلوم وعدول وريد المدور وروع الله المحافظ الم المحافظ الم الم ورو والمحافظ المحافظ المحاف נוצ-שע כיי עד פרוכף באני מייר באי 2059 march In same 44.00 754035 47 Robin 775 4502 algeria 330 syx bss roza march 1 335 M yest. 2045 - 1 ree 958 me de 54371 (5- 00 Jusy - 134523 264 Juni 55 Lane - so abso men exerce. yestise - 522 4755 In section abis - 4015 allows of > ~ ~ ~ 9.= tostato othered ENGLAND CALL STOP AJON DIX JUNIC J THURE it assents of THE SHE were droc for an איל ייבוצ צעות צבל ניול ניאצ בארץ ארוצנים באיניים ני נפואניל אי ני אוצע פיני at some not be prover & E ciering was lover or salution some

Old Sogdian: Research





Script Encoding Process: 'Script'

Define the 'script':

- Identify repertoire of distinctive signs and variants
- Establish representative glyphs
- Generate script grammar
- Develop encoding model
 - convert conventions for 'writing' into 'typing'
 - translate qualitative data into technical metadata

Script Encoding Process: 'Script'

Latin script for English: Distinctive signs

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Latin script for English: Variant signs

ABCDEFGHIJKLMMOPQRSTUVWXYZ

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

ABCDCFGHIJKLMNOPQRSTUVWXYZ

ABCDEFGHIJKLMNOPQRSTUVWXYZ





U+0041 LATIN CAPITAL LETTER A

Script Encoding: Unicode





Not-plain text Plain text

Script Encoding Process: 'Script'

- Is the 'script' a distinctive system or a stylistic variant?
- Character-glyph model:
 - encodes a 'normative' or representation form for a sign as a 'character'
 - defines other forms as 'glyphic' variants
 - plain text vs. not-plain text / 'rich text' (font selection)
- "Unicode encodes characters, not glyphs"

Script Encoding Process: 'Script'

- A script may have historical, scribal, and stylistic variants
- Considerable differences may exist between early and late forms
 - Old Sogdian and Sogdian
- Script boundaries unclear
 - Form of script in early phase may resemble an ancestor
 - Late form of script may resemble a descendant
- Graphically similar scripts may have structural differences

Developmental 'Variant': Vertical Old Sogdian



Developmental 'Variant': Formal Sogdian

service and a set as might be have the and a and set and and the set and and and and and and sies and set operated and any way and proved second ويلعه بعا لجد مديد مد وماس ويعما معا معد عما معا وجريد as to and and the safe many server and the los of as موج عد وماري عدومة مسطة ومة و وحمصة موج و حمده معهود العد معدو مدا و وجدد واع موجد مدو دهم عده عد way as was agen and they say and and and و و معد عو به عمالمه مدر و بدا مدر مدو بد و و معد عمد مدد Con may an any and an mag and part and معولا عولا المتعط معا المتد و ومعد معدما و و المت المود ومعد erenal as any and the as a forest are not and sold service and and and and a state of the service and the and the second second some as based second to a second cool and carse and all all act a series and an and and the loss

Developmental 'Variant'?: Cursive Sogdian



'Script' or 'Variant'?: Chorasmian

21 22 23 24 25 26 27 28 29 3 0 31 annel ters laget fange fener fannelinge feerfrer Fig. 2. Inscription No. 25.

'Script' or 'Variant'?: Old Uyghur



'Script' or 'Variant': Mongolian

Old Sogdian: How to encode?

- Unification with Aramaic?
 - uses characters / code points from Imperial Aramaic block?
 - Palaeographically valid, but should Sogdian be considered a 'style' of Aramaic?
- Single encoding:
 - same character set for representing different forms of the script
 - usage of fonts to display different script styles?
 - but, which form is to be representative for plain text?
- Separate encoding:
 - two different character sets for historical stages of Sogdian
 - enables both scripts to be displayed simultaneously, without font changes

Aramaic & Old Sogdian

לארף אנתן בדרי נושל היורים לעות אירים אירים אירים און אריך ריש רנשך כראב אינתן בדרי נחכקלה אחורותם משמת שיין אריך רש רנשך כרי בעינות עון אות מחרית אמתקשין אודים וין שני רשוי כתריון רקב כבר שת אאוקין שתן כבאי ווך אי שלט אינת ורטה רלעתן

750 YOUCHESON 955 CAN unen yes KK KATHER 5 7500 453 417HS 2565 Carrol ELESYNEW 40 me cress se you of 4505 amph P assynall fil 5 prach " \$40000 BUCHER SA SO CEN C 015 4,60 1 × 45305 15 : CON 25:470 79 9550 75 10 25-311 2m the sources MS manus 2 10

Imperial Aramaic

Central Asian Scripts Manichaean Marchen Mongolian Mongolian Supplement Old Sogdian Old Turkic Phags-Pa Sogdian Soyombo Tibetan Zanabazar Square

4 7 7 7 yudh gāmal bēth kāph tēth hēth zain hē dālath ālaph waw $[k/x] [j/i:/e:] [t^{S}]$ [ħ] [z] [w/o:/u:] [h] $[d/\delta]$ [g/y] [b/v] [2/a:/e:]

V J Р p\Y υ Ŋ shin rēsh qoph şādhē 'ē semkath mim lāmadh tau pē nun [ʃ] [²a] [p] [t/θ] [p/f] [2] [1] [r] [s] [n][m]

Middle Eastern Scripts Anatolian Hieroglyphs Arabic Arabic Supplement Arabic Extended-A Arabic Presentation Forms-A Arabic Presentation Forms-B Aramaic, Imperial Avestan Cuneiform (1MB) **Cuneiform Numbers and Punctuation** Early Dynastic Cuneiform Old Persian Ugaritic Hatran Hebrew Hebrew Presentation Forms Mandaic Nabataean **Old North Arabian Old South Arabian** Pahlavi, Inscriptional Pahlavi, Psalter Palmyrene Parthian, Inscriptional Phoenician Samaritan Syriac Syriac Supplement

Sogdian Scripts: Repertoire

	Aramaic	Old Sogdian	Sogdian		Aramaic	Old Sogdian	Sogdian
aleph	¥	ک ر بک	<u>_</u>	lamedh	L	7	٢
beth	>	ت, ع	4	тет	ゥ	Ун	さ
gimel	1	ч	~	nun	\$	ا ,د. ,ر	ل
daleth	7	У	—	samekh	,	n	~
he	7)	کم , ک	c	ayin	V	ע, 50, ע	© ∿, y
waw	,	2	9	pe	ر	و	لا
zayin	1	J	▲	sadhe	ŗ	۲, ۲, ۲	لد
heth	"	Я	h	qoph	マ	—	—
teth	G	—	_	resh	7	У	У
yodh	4	5	5	shin	V	74	s, ls
kaph	y	У	Ŀ	taw	r	ם , מ_ , מ	له

Encoding: Sogdian and Aramaic

- Differences in repertoire: Sogdian
 - *Heth, teth, qoph* dropped
 - Daleth, ayin, resh merged
 - Special ayin used for saluations, eg. `D 'to'
 - Scribal innovations, eg. tail extensions for final characters
 - Special numerical sign for 100
- Cultural and scribal divergences

Old Sogdian: final letters: aleph



Old Sogdian: ayin

- حمد دم موطله لا المحلود ورو ورو الم المعلو ولا و و الم الم المع المع المعلمة الم

UN ENO HORDO ENTROP ECT UNDER TRUES ALLOS AUSTAN PULL TALLE EN EL

2 min der son ange and indigene our of the

Conclusions

- Script encoding process
 - not art, not science, but a bit of both
- Skills to encode scripts:
 - interest in writing systems and willingness to learn
 - don't need to be a linguist, typographer, or computer scientist
 - research and analytical skills
 - search, search, and search more, then dig deeper, and follow leads