Syriac Unicode Standard

Peter (Jasim) BetBasoo
Nineveh Software Corporation

off-print from

SyrCOM-95

Proceedings of the First International Forum On Syriac Computing
(In Association wit Syriac Symposium II)

June 8, 1995
The Catholic University of America
Washington, D.C.

Edited by
George Anton Kiraz
University of Cambridge
(St. John's College)

Published by the Syriac Computing Institute

Syriac Unicode Standard

Peter Jasim
Nineveh Software Corporation

A data interchange code is a standard coding scheme applied to the letters, symbols and punctuation marks (collectively called characters) that comprise a language. The two prevailing standards for English are the American Standard Code for Information Interchange (ASCII), used by all microcomputers, and the Extended, Binary Coded, Data Interchange Code (EBCDIC), used by IBM mainframes.

ASCII and EBCDIC are both one byte standards. A byte is the smallest unit of memory a computer can operate on. Physically, a byte is a location in memory that has eight switches, each of which can be in one of two states, on or off. Therefore, one byte can represent $2^8 = 256$ combinations of states. In other words, one byte can store a number between 0 and 255.

In a one byte coding standard, up to 256 characters can be defined (Appendix A). For example, using ASCII the word Assyrian would be stored internally by a computer as follows

Character	A	s	s	y	r	i	a	n
ASCII code	65	115	115	121	114	105	97	110

On any computer that uses ASCII, the above sequence of numbers, when interpreted in a textual context[1], would yield the word Assyrian.

In the absence of such a standard, communication among computers (and the people who use them) would be difficult. As an example, when transferring a document created on a personal computer (which uses ASCII) to an IBM mainframe (which uses EBCDIC) it is necessary to translate the codes from ASCII to EBCDIC. This is not difficult to do, since there is a one-to-one mapping between ASCII and EBCDIC, but it would have been unnecessary had both computers used one coding scheme.

The one byte length of ASCII, EBCDIC and similar coding standards imposes severe limitations on coding non-Latin languages, particularly the oriental languages, which have millions of ideographs. In these cases, ASCII is abandoned and there exist many local standards (i.e., no standards). This makes it very difficult to share documents among computers. In today's global community, it is becoming increasingly important to communicate effectively and efficiently. Clearly, a worldwide coding standard is needed to facilitate global communications.

The Unicode Standard

Unicode is a new coding standard which encompasses all of the languages of the world. It has rapidly gained acceptance by the major computer vendors and has been merged with the International Standards Organization's worldwide coding standard (ISO 10646).

---

[1] Since a byte is just a number between 0 and 255, and all information stored by a computer is in bytes (i.e., numbers), the interpretation of what is stored in memory (i.e., what the numbers represent) depends on the context. In a word processing program, 65 means A, but in a spreadsheet program, 65 may mean just that, the number 65.

Unicode is a two byte standard. This means that it can accommodate $2^{16} = 65536$ characters. This is enough codespace for all the living languages in the world, and for archaic and extinct languages as well. It is pleasing to note that Syriac, the oldest extant spoken language, is part of this newest of standards.

The key features of Unicode are:

1) Each code point (character) represents an abstract semantic entity, and is independent of how that entity is rendered on an output device. For example, the following forms of *Hea*, ܗ ܗ ܗ are the same in Unicode and would be assigned the same codepoint; the only difference between them is their appearance -- their semantic identity is identical.

2) The Unicode standard is a unified coding scheme; this means that if two languages have a semantically or physically identical character, such as a period or questions mark in Latin languages, then that entity is defined only once and is shared among the different languages, except when this leads to semantic ambiguity, in which case different codepoints are assigned to each entity. It is for this reason that Unicode is able to represent all Korean, Japanese and Chinese ideographs in a Unified Han Character Set, which uses about twenty one thousand of Unicode's codepoints.

The Syriac Unicode Standard

The Syriac Unicode Standard (Appendix F) was drafted by the author and Sargon Hasso. The author had been working alone on a Syriac Unicode Standard when he learned from the Unicode representative that Sargon Hasso was also working on the standard. We joined forces and produced the final standard.

Unicode reserves 512 codepoints for Syriac; this is enough to encode all the characters in Syriac. There are three categories of characters in Syriac

1. The alphabet
2. Symbols and punctuation marks
3. Diacritical marks, which are called *Paroshe* in Syriac (ܦܪܘܫܐ)

The Alphabet

The Syriac Unicode Standard includes the twenty two Syriac letters. In addition, it reserves an additional codepoint for the twenty third letter in Mandaic (codes SSSS+0001 to SSSS+0023).

Symbols and Punctuation Marks

Ten Syriac symbols and punctuation marks are defined (S+0001 to S+0010).

Paroshe

Paroshe are subdivided into two groups, vowels (ܐܒܓ݀) and accents (SS+0001 to SS+00020 and SSS+0001 to SSS+0017).

The Unification of Syriac

In keeping with the design philosophy of Unicode, the Syriac Unicode Standard unifies the characters of Syriac into one codeset. The alphabet and *Paroshe* are unified.

The unification of the alphabet is straightforward, as there are no semantic ambiguities. The only special consideration is the addition of a twenty third character to support Mandaic.

The Syriac *Paroshe* are varied and complex, reflecting their two thousand year history. At first it seemed a daunting task to unify the *Paroshe* from different periods, but it turned out to be surprisingly easy. The result is a standard that covers the past two thousand years of Syriac writing.

Following Segal, we initially group *Paroshe* as follows

1. Before 7th century (Appendix B)
2. 7th to 10th century, Western (Appendix C)
3. 7th to 10th century, Eastern (Appendix D)

A comparative analysis of these *Paroshe* of differing periods and locales shows the similarity between them. The *Paroshe* of Appendix B and Appendix C are a subset of the *Paroshe* in Appendix D. The unified *Paroshe* are shown in Appendix E.

When two or more semantically distinct *Paroshe* have the same appearance, they are unified and given one codepoint (e.g., C5 and D6). Their semantic identity must be inferred from their context, in which case it does not matter if they have distinct codes (this is analogous to a period being used to mark the end of a sentence or to denote the decimal portion of a number). For this reason, we are able to unify the *Paroshe* from differing periods and locales into one superset.

Each *Parosha* in Appendix E is given a unique name so that it can be identified unambiguously. Appendix E also shows which *Paroshe* were unified.

Appendix F shows the complete Syriac Unicode Standard.

Conclusions

The Syriac Unicode Standard is a comprehensive coding standard for the Syriac language. Once adopted, the standard will facilitate the computerization of Syriac across varied hardware and software platforms. With this standard, any Syriac manuscripts can be reproduced electronically, whether they be the hymns of *Mar Aprim* from the 5th century or the novels of the 20th century Assyrian Michael Lazar 'eesa. This opens up the vast and exciting realm of software analysis of Syriac manuscripts.

References

DeKelaita, Joseph. *Grammar of the Aramaic Language*. Assyrian Church of the East Press. 1929

Segal, J. B. *Diacritical Points and the Accents in Syriac*. Oxford University Press. 1953.

Unicode Consortium. *The Unicode Standard, Version 1.1, Volume 1*. Addison-Wesley.

Unicode Consortium. *The Unicode Standard, Version 1.1, Volume 2*. Addison-Wesley.

Appendix A
American Standard Code for Information Interchange

Code	Character
000	NUL
001	SOH
002	STX
003	ETX
004	EOT
005	ENQ
006	ACK
007	BEL
008	BS
009	HT
010	LF
011	VT
012	FF
013	CR
014	SO
015	SI
016	SLE
017	CS1
018	DC2
019	DC3
020	DC4
021	NAK
022	SYN
023	ETB
024	CAN
025	EM
026	SIB
027	ESC
028	FS
029	GS
030	RS
031	US
032	SPACE
033	!
034	"
035	#
036	$
037	%
038	&
039	'
040	(
041	)
042	*

Code	Character
043	+
044	,
045	-
046	.
047	/
048	0
049	1
050	2
051	3
052	4
053	5
054	6
055	7
056	8
057	9
058	:
059	;
060	<
061	=
062	>
063	?
064	@
065	A
066	B
067	C
068	D
069	E
070	F
071	G
072	H
073	I
074	J
075	K
076	L
077	M
078	N
079	O
080	P
081	Q
082	R
083	S
084	T
085	U

Code	Character
086	V
087	W
088	X
089	Y
090	Z
091	[
092	\
093	]
094	^
095	
096	`
097	a
098	b
099	c
100	d
101	e
102	f
103	g
104	h
105	i
106	j
107	k
108	l
109	m
110	n
111	o
112	p
113	q
114	r
115	s
116	t
117	u
118	v
119	w
120	x
121	y
122	z
123	{
124	|
125	}
126	~
127	DEL

Appendix B
Diacritical Points Before the 7th Century

B1	◌̇	[illegible]
B2	◌̇	[illegible]
B3	◌݃	[illegible]
B4	◌̈	[illegible]
B5	◌̣	[illegible]
B6	◌̣	[illegible]
B7	.◌	[illegible]
B8	˙◌	[illegible]
B9	.◌̣	[illegible]
B10	:◌	[illegible]
B12	◌̤	[illegible]

Appendix C
Western Diacritical Points 7th to 10th Centuries

C1	˙◌	[illegible]
C2	◌̇	[illegible]
C3	◌˙	[illegible]
C4	◌̇˙	[illegible]
C5	˸◌ | ◌݃	[illegible]
C6	·◌̈ | ◌̈	[illegible]
C7	◌̣	[illegible]
C8	◌̣	[illegible]
C9	.◌̣	[illegible]
C10	.◌	[illegible]
C11	◌̤	[illegible]

Appendix D
Eastern Diacritical Points
7th to 10th Centuries

D1	◌̇	ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ
D2	◌̇	ܟܝ ، ܟܬ ، ܟܬ
D3	◌̇	ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ
D4	◌̣	ܗܘ ܕܟܝ
D5	◌̇	ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ
D6	◌̣	ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ ، ܟܬ ، ܟܬ ، ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ
D7	◌̇	ܟܬ
D8	◌̈ | ◌̈	ܟܬ
D9	◌̊	ܟܬ ܕܟܝ ، ܟܬ ܕܟܝ
D10	◌̤	ܟܬ ܕܟܝ
D11	◌̣	ܟܬ
D12	◌̣	ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ ، ܟܬ
D13	◌̣ | ◌̣	ܟܝ ، ܟܝ
D14	◌̣ | ◌̣	ܟܬ
D15	◌̤	ܟܬ
D16	◌̤	ܟܬ ܕܟܝ ، ܟܬ ܕܟܝ
D17	◌̣	ܗܘ ܕܟܝ ، ܗܘ ܕܟܝ ، ܟܬ
D18	◌̤	ܟܝ ، ،
D19	◌̣	ܗܘ ܕܟܝ ، ،
D20	◌̤	ܗܘ ܕܟܝ ، ،

Appendix E
Unified Diacritical Points

		Eastern	Western
E1	◌̣	ܗܘ ܕܟܝ	C1
E2	◌̣	ܗܘ ܕܟܝ	C2
E3	◌̣	ܗܘ ܕܟܝ	C3
E4	◌̣	ܗܘ ܕܟܝ	
E5	◌̣	ܗܘ ܕܟܝ	
E6	◌̣	ܗܘ ܕܟܝ	C5
E7	◌̣	ܗܘ ܕܟܝ	C4
E8	◌̈ | ◌̈	ܗܘ ܕܟܝ	C6
E9	◌̊	ܗܘ ܕܟܝ ܕܟܝ ، ܗܘ ܕܟܝ ܕܟܝ	
E10	◌̤	ܗܘ ܕܟܝ ܕܟܝ	
E11	◌̣	ܗܘ ܕܟܝ	C7
E12	◌̣	ܗܘ ܕܟܝ	C8
E13	◌̣	ܗܘ ܕܟܝ	
E14	◌̣ | ◌̣	ܗܘ ܕܟܝ	
E15	◌̤	ܗܘ ܕܟܝ	C9
E16	◌̤	ܗܘ ܕܟܝ ܕܟܝ ، ܗܘ ܕܟܝ ܕܟܝ	
E17	◌̣	ܗܘ ܕܟܝ	C10
E18	◌̤	ܟܝ ، ،	
E19	◌̣	ܗܘ ܕܟܝ ، ،	
E20	◌̤	ܗܘ ܕܟܝ ، ،	C11

Appendix F
Syriac Unicode Standard

Assyrian[1] S[2]+0000 -- SSSS+0000

The Assyrian script (*Syriac*), which is used for writing the Assyrian language, includes the Eastern Assyrian script (*Nestorian*), the Western Assyrian script (*Serto* or *Jacobite*), and the *Estrangelo*[3] script. It is also used for writing Mandaic[4].

Assyrian script which is predominantly cursive is written from right to left even in its printed form. Few letters are written in different forms depending on how they join to their neighbors. Vowels (*zaw'e)* are placed above or below the consonantal base letters (radical).

Mandaic. Mandaic is written with the same script, with an additional, 23rd letter; this extra letter is given the independent code SSSS+0023.

Punctuation. Most punctuation marks used in Assyrian are not given independent codes (they are unified with the Latin, Arabic, and Hebrew punctuation) except for the few cases where the mark has a unique form and function in Assyrian.

Encoding Principles. The alphabet of Assyrian is well defined. Each letter receives only one Unicode character value regardless of the number of contextual shapes it may exhibit in text (this, indeed, is the only difference between Eastern Assyrian, Western Assyrian, Estrangelo, and Mandaic). The graphic form (glyph) shown in the Unicode character chart is primarily that of free-form Estrangelo.

Diacritical Points[5] (paroshe). These are marks (more commonly, these are points of large,

---

[1] We use *Assyrian* and *Syriac* interchangeably in this working proposal. However, we would strongly suggest the use of *Assyrian* as a proper name for this script.

[2] This notation is for this working proposal only, and it follows the same convention as used in Unicode Version 1.0, i.e., U+nnnn. *S* stands for Syriac.

[3] It is correctly spelled with an *o* the end.

[4] Two issues were raised in the exploratory proposal: the order of letters and Mandaic. The order of letters is correct as it appears -- *Waw* is in its correct place. *Waw* should not be placed at the end because *Waw* is, by virtue of its position in the Assyrian alphabet (the sixth letter) also the number 6. Assyrian letters are also used as numbers and have ordinal values. More information will be provided if need be. We have left the last space in the character set as reserved for the extra letter in Mandaic.

[5] Generally we refer to all objects that are placed around the base letter in various positions, as *Diacritical Points*. However, they fall into four well-defined categories: distinction points, e.g., SSS+0005, the plural sign, e.g., SSS+0001, the actual diacritical point in its various forms e.g., SSS+0008, and, finally, the accents, e.g., SS+0008. For an in depth treatment of this subject, please cf. Segal's *The Diacritical Point and The Accents in Syriac*, Oxford University Press, 1953.

medium, and small sizes) that indicate vowels (*zaw'e*), cantillation marks, accents, and other modifications of consonantal letters. The occurrence of a character in the *Paroshe* range and its depiction in relation to a dashed circle constitute an assertion that this character is intended to be applied via some process to the consonantal letter, phrase and/or clause that precedes it in the text stream. General rules for applying non-spacing marks are given in the Generic Diacritical Mark block description section in the Unicode Standard, version 1.0. The Unicode standard does not specify a sequence order in case of multiple marks applied to the same Assyrian base character since there is no possible ambiguity of interpretation. The Assyrian script contains a rich set of diacritical marks which reflects it development over the course of its long history.

Encoding Structure. The Assyrian character block is divided into the following[6]:

S+0001-- S+0010	Assyrian punctuation and number marks
SS+0001 -- SS+0020[7]	Assyrian diacritical marks (*paroshe*) -- I
SSS+0001 -- SSS+0017	Assyrian diacritical marks (*paroshe*) -- II
SSSS+0001 -- SSSS+0023	Assyrian letters

[6] This subdivision is for future addition and expansion. Cf. following note.

[7] This list by no means is complete. Although, the majority of all Assyrian documents and manuscripts can be reproduced with the aid of only this list -- whether they are from 500 A.D. or from 1993 A.D. Therefore, we would like to reserve additional space following this category for future addition and amendment.

SEQUENCE	GLYPH	DEFINITION
S+0001	܀	SYRIAC QUOTATION MARK = *sahrane*
S+0002	܁	SYRIAC ABBREVIATION MARK = *gadmaia*
S+0003	܂	SYRIAC END OF PARAGRAPH SEPARATOR
S+0004	܃	SYRIAC LONG PAUSAL MARK PASUQA = *period*
S+0005	܄	SYRIAC SHORT PAUSAL MARK ZAUGA = *comma*
S+0006	؟	SYRIAC QUESTION MARK
S+0007	܇	SYRIAC NUMERAL SIGN FOR TEN
S+0008	܈	SYRIAC NUMERAL SIGN FOR THOUSAND
S+0009	܉	SYRIAC NUMERAL SIGN FOR TEN THOUSAND
S+0010	܊	SYRIAC NUMERAL SIGN FOR MILLION

SS+0001	◌̇	SYRIAC ONE DOT ABOVE RIGHT = *mshalana*
SS+0002	◌̈◌̇	SYRIAC TWO DOTS HORIZONTAL ABOVE = *rahta*
SS+0003	◌̇◌	SYRIAC TWO DOTS LEFT-SLANTED ABOVE = *mshalmanota*
SS+0004	◌̇	SYRIAC ONE DOT ABOVE CENTER = *ritma*
SS+0005	◌̈	SYRIAC TWO DOTS VERTICAL CENTER ABOVE = *zauga elaya*
SS+0006	◌̈◌̇	SYRIAC THREE DOTS ABOVE = *rahta d kane*
SS+0007	◌̇	SYRIAC ONE DOT ABOVE LEFT = *esyana*
SS+0008	◌̇	SYRIAC TWO DOTS VERTICAL RIGHT ABOVE = *mdamrana*
SS+0009	◌̇	SYRIAC ONE DOT IN-LINE LEFT = *pasoqa*
SS+0010	◌̈	SYRIAC TWO DOTS VERTICAL IN-LINE LEFT = *zauga*
SS+0011	◌̇◌̇	SYRIAC TWO DOTS IN-LINE ABOVE AND BELOW = *mqinana*
SS+0012	◌̣	SYRIAC ONE DOT RIGHT BELOW = *mnakhta*

SS+0013		SYRIAC ONE DOT LEFT BELOW = *samka*
SS+0014		SYRIAC TWO DOTS VERTICAL LEFT BELOW = *mikaslipana*
SS+0015		SYRIAC ONE LARGE DOT AND ONE SMALL DOT SLANTED LEFT BELOW = *esasa*
SS+0016		SYRIAC ONE LARGE DOT AND ONE SMALL DOT BELOW = *napsha*
SS+0017		SYRIAC TWO DOTS RIGHT-SLANTED ABOVE LEFT = *elaya*
SS+0018		SYRIAC TWO DOTS ABOVE AND IN-LINE DOT LEFT = *rohta d paseq*
SS+0019		SYRIAC ONE SMALL DOT AND ONE LARGE DOT LEFT-SLANTED BELOW LEFT = *takhtaya*
SS+0020		SYRIAC THREE DOTS BELOW = *takhtaya d talata*
SSS+0001		SYRIAC PLURAL MARK SYAME
SSS+0002		SYRIAC FRICATION MARK RUKAKHA
SSS+0003		SYRIAC AFFRICATION MARK MALIANA

SSS+0004		SYRIAC HALF CIRCLE BELOW A LETTER = *qishta*
SSS+0005		SYRIAC DIACRITICAL LARGE POINT OVER A LETTER
SSS+0006		SYRIAC DIACRITICAL LARGE POINT BELOW A LETTER
SSS+0007		SYRIAC DIACRITICAL POINT QUSHAYA
SSS+0008		SYRIAC VOWEL MARK ZQAPA
SSS+0009		SYRIAC VOWEL MARK PTAKHA
SSS+0010		SYRIAC VOWEL MARK ZLAME PSHIQE
SSS+0011		SYRIAC VOWEL MARK ZLAME QASHYE
SSS+0012		SYRIAC VOWEL MARK RWAKHA
SSS+0013		SYRIAC VOWEL MARK RWASA
SSS+0014		SYRIAC VOWEL MARK KHWASA
SSS+0015		SYRIAC ACCENT MTALQANA

SSS+0016	◌̱	SYRIAC HALF-VOWEL MHAGYANA
SSS+0017	◌̲	SYRIAC HALF-VOWEL MARHTANA
SSSS+0001	ܐ	SYRIAC LETTER ALLAP
SSSS+0002	ܒ	SYRIAC LETTER BET
SSSS+0003	ܓ	SYRIAC LETTER GAMMAL
SSSS+0004	ܕ	SYRIAC LETTER DALLAT
SSSS+0005	ܗ	SYRIAC LETTER HE
SSSS+0006	ܘ	SYRIAC LETTER WAW
SSSS+0007	ܙ	SYRIAC LETTER ZAIN
SSSS+0008	ܚ	SYRIAC LETTER KHET
SSSS+0009	ܛ	SYRIAC LETTER TET
SSSS+0010	ܝ	SYRIAC LETTER YUDH
SSSS+0011	ܟ	SYRIAC LETTER KAP
SSSS+0012	ܠ	SYRIAC LETTER LAMMADH
SSSS+0013	ܡ	SYRIAC LETTER MEEM
SSSS+0014	ܢ	SYRIAC LETTER NUN
SSSS+0015	ܣ	SYRIAC LETTER SIMKAT
SSSS+0016	ܥ	SYRIAC LETTE'

SSSS+0017	ܦ	SYRIAC LETTER PE
SSSS+0018	ܨ	SYRIAC LETTER SADEH
SSSS+0019	ܩ	SYRIAC LETTER QOP
SSSS+0020	ܪ	SYRIAC LETTER RESH
SSSS+0021	ܫ	SYRIAC LETTER SHEEN
SSSS+0022	ܬ	SYRIAC LETTER TAW
SSSS+0023		RESERVED