

**Title: Proposal of Myanmar Script Extensions:
Mon, Shan, and Karen (Kayin)**

Source: Myanmar Unicode and NLP Research Center

Status: National Contribution

Action: For consideration by WG2

Date: 31-May-2004

This document requests additional characters to be added to the UCS and contains the proposal summary form.

A. Administrative

1. Title

Proposal of Myanmar Script Extensions: Mon, Shan, and Karen (Kayin)

2. Requester's name

Myanmar Unicode and NLP Research Center

3. Requester type (Member body/Liaison/Individual contribution)

National contribution.

4. Submission date

2004-05-31

5. Requester's reference (if applicable)

Thein OO, President, MCF <mcf@mail4u.com.mm>

Thein HTUT, Secretary, MCSA

Tun TINT, Member, Myanmar Language Commission

Zaw HTUT, Program Manager, Myanmar UNLP Research Center <zhtut@myanmars.net>

Ngwe TUN, Program Manager, Myanmar UNLP Research Center <ngwestar@etrademyanmar.com>

c/o:

MYANMAR UNICODE AND NLP RESEARCH CENTER

Myanmar Computer Federation

B1R1, Myanmar ICT Park, Universities' Hlaing Campus,

11052, Yangon, Myanmar

Tel: +95-1-652307

eFax: +1-707-988-0300

Email: myanmar-nlp@mail4u.com.mm , mcf@mail4u.com.mm

Internet: <http://www.mcf.org.mm> , <http://myanmars.net/unicode/>

6. Choose one of the following:

6a. This is a complete proposal

Yes.

6b. More information will be provided later

Yes.

B. Technical – General

1. Choose one of the following:

1a. This proposal is for a new script (set of characters)

No.

Proposed name of script

1b. The proposal is for addition of character(s) to an existing block

Yes.

1b. Name of the existing block

Myanmar (U+1000 to U+109F)

2. Number of characters in proposal

30

3. Proposed category (see section II, Character Categories)

Category A

4a. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000)

Level 3

4b. Is a rationale provided for the choice?

Yes.

4c. If YES, reference

- Existing Mon, Shan, and Karen Primers
- Documents on <http://www.mcf.org.mm> , <http://myanmars.net/unicode/>

5a. Is a repertoire including character names provided?

Yes

5b. If YES, are the names in accordance with the character naming guidelines in Annex L of ISO/IEC 10646-1: 2000?

Not confirmed.

5c. Are the character shapes attached in a legible form suitable for review?

Yes.

6a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?

Zaw HTUT.

TrueType.

6b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

Zaw HTUT, Program Manager, Myanmar Unicode & NLP Research Center.

email: myanmar-nlp@mail4u.com.mm, mcf@mail4u.com.mm,

internet: <http://www.mcf.org.mm>, <http://www.myanmars.net/unicode/>

FontLab.

7a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

Yes, see bibliography.

7b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

Yes.

8. 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, the basics only.

9. 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/Public/UNIDATA/UnicodeCharacterDatabase.html> and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

Yes, see proposal below.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.
No.

2a. 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.

2b. If YES, with whom?

- Myanmar ICT Standardization Steering Committee (national body)
- Mon, Shan, and Karen linguists who are working with the Myanmar Language Commission (national body)
- Myanmar Computer Federation [<http://www.mcf.org.mm>]
- Myanmar Computer Scientist Association [<http://www.mcsa.org.mm>]
- Myanmar Unicode and NLP Research Center [<http://myanmars.net/unicode>]

2c. If YES, available relevant documents

All available documents related to Myanmar NLP are listed at
<http://www.mcf.org.mm>, <http://myanmars.net/unicode/doc/>

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
Yes.

4a. The context of use for the proposed characters (type of use; common or rare)
Common everyday use.

4b. Reference
See bibliography.

5a. Are the proposed characters in current use by the user community?
Yes.

5b. If YES, where?
In Myanmar (formerly Burma), especially in Mon, Shan, and Karen States.

6a. After giving due considerations to the principles in Principles and Procedures document (a WG 2 standing document) must the proposed characters be entirely in the BMP?
Yes, since there is a reserved space for these.

6b. If YES, is a rationale provided?
Yes.

6c. If YES, reference
All Myanmar points are in the BMP.

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
Strongly suggest to be kept together with Myanmar.

8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
Yet to be confirmed later.

8b. If YES, is a rationale for its inclusion provided?
Not yet.

8c. If YES, reference

9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?
Still yet to be confirmed soon.

9b. If YES, is a rationale for its inclusion provided?
Not yet.

9c. If YES, reference

10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

Yes, some can be.

10b. If YES, is a rationale for its inclusion provided?

Yes.

10c. If YES, reference

Many of them derived from Pali and Sanskrit, but they have different functions and shapes.

11a. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)?

Yes.

11b. If YES, is a rationale for such use provided?

Yes.

11c. If YES, reference

Sample printed documents.

12a. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

Yes.

12b. If YES, reference

See Annex-3.

13a. Does the proposal contain characters with any special properties such as control function or similar semantics?

No.

13b. If YES, describe in detail (include attachment if necessary)

14a. Does the proposal contain any Ideographic compatibility character(s)?

No.

14b. If YES, is the equivalent corresponding unified ideographic character(s) identified?

14c. If YES, reference

D. Proposal

D.1. Background

D.1.1. Myanmar script was firstly proposed to Unicode Consortium in 1992, and was successfully encoded in the repertoire in 1998 in Unicode Standard 3.0. Other sub-scripts, such as Kayah, Karen, Chin, Mon, and Shan, based on the Myanmar script.

D.1.2. Out of 160 code points reserved for the Myanmar script (between U+1000 and U+109F), only 78 code points were encoded and used for Myanmar script, as of the Unicode 4.0.

D.1.3. Totally 82 code points in the Myanmar block in the Unicode BMP (Basic Multilingual Plane) are still being reserved for the rest of the sub-scripts to be encoded.

D.2. User Community

D.2.1. The Mon state is inhabited by approximately 2,000,000 people, and its size is 4748 square miles. A lot of Mons still live in western Thailand, near Myanmar border. The use of IT is moderate, and the number of Mon books published a year is not available.

D.2.2. The Shan state is the largest state in the country, and the estimated number of inhabitants is as large as 4,400,000. The size of Shan state is 60155 square miles. Due to the nature of high Shan plateau, most parts are remote areas, and IT usage is very limited. Shan culture is close to the Siamese, now called Thai. A lot of Shan people also live in Thailand and China. Publications printed in Shan script may be the largest after Myanmar. Its demographics shows that the Shans have 33 minor ethnic groups.

D.2.3. The Karen state is bordering Thailand and has 1,200,000 inhabitants approximately. The size 11,731 square miles is not the smallest state in Myanmar. Use of IT in schools and offices are quite limited. It was noted that the Christian bible was even translated and then printed in Karen language. Regarding demographics, there are 11 minor ethnic groups in Karen group, and among them Sakaw Karen and Poe Karen are dominant and their scripts have a few minor differences.

D.3. Proposed Characters

D.3.1. Mon Sub-Script

D.3.1.1. Actually it was the Mons who brought in the Indict script first and applied it for their own language. Myanmar follows later. Mons use all the alphabets of the Myanmar, plus 2 more consonants.

D.3.1.2. Number of new proposed Mon characters: 6 (3 consonants, 2 medials (combining consonants), and 1 special character)

D.3.1.3. Graphical presentation of proposed Mon character set: See [Annex-1](#).

D.3.1.4. List of proposed Mon character names: See [Annex-2](#).

D.3.2. Shan Sub-Script

D.3.2.1. Shan script seems to be a lot more simplified than Myanmar and Thai, having only 20 consonants.

D.3.2.2. Number of new proposed Shan characters: 14 (2 medials (combining consonants), and 5 vowel sign and 7 tone mark sign)

D.3.2.3. Graphical presentation of proposed Shan character set: See [Annex-3](#).

D.3.2.4. List of proposed Shan character names: See [Annex-4](#).

D.3.3. Karen Sub-Script

D.3.3.1. This presented version of Karen script is a combined set of characters both from Sakaw Karen and Poe Karen scripts.

D.3.3.2. Number of new proposed Karen characters: 10 (2 consonants, 2 medials (combining consonants), and 6 devowelisers)

D.3.3.3. Graphical presentation of proposed Karen character set: See [Annex-5](#).

D.3.3.4. List of proposed Karen character names: See **Annex-6**.

D.4. Character Properties of Proposed Characters

D.4.1. Line-breaking Behavior

The line-breaking rules applies same as the base script – Myanmar. It is true that Myanmar script does not use "white space" like Latin to break up words. But Myanmar line breaking is not as complicated as Thai.

According to Myanmar script, syllables have be formed up, and all line breakstake place ONLY outside syllable boundaries (always before or after syllables). To clarify, the line-breaks takes place always before pre-base and base characters, and always after a set of post-base characters.

It is important to note that there are pre-base characters before the base consonants, although they have be stored behind the base character in the backing store.

D.4.2. Combing Behavior and Spacing Behaviors

D.4.2.1. All consonants and special vowel characters have spacing same as any other consonants.

D.4.2.2. Above-base, below-base, and stacked characters (as modifiers of the base character) doesnot take up any spacing.

D.4.2.3. Some post-base characters takes up a little bit of spacing.

D.4.2.4. Mon characters

Mon: (base)			
Mon: (above-base)			
Mon: (below-base)			
Mon: (stacked)			

D.4.2.5. Mon Consonants

1004	1008	1060	1061

D.4.2.6 Mon Medial

1063	1064	1065

D.4.2.7. Mon Vowel Signs

1024	1027	1062

D.4.2.8. Shan characters

Shan: (base)



Shan: (pre-base)



Shan: (post-base)



Shan: (below-base)



D.4.2.9. Shan Consonants



1000



1001



1005



1006



100A



1014



1016



101F



1021

D.4.2.10. Shan Medial Characters



1068



1069

D.4.2.11. Shan Vowel Signs



106A



106B



106C



106D



106E

D.4.2.12. Shan Tonal Marks



106F



1070-



1071



1072



1073

D.4.2.13. Shan Paoh Tonal Marks



1066



1067

D.4.2.14. Karen characters

Karen: (base)



Karen: (post-base)



Karen: (below-base)



D.4.2.15. Karen Consonants



1074



1075

D.4.2.16. Karen Medial Characters



1065



1076

D.4.2.17. Karen Devowelisers (killers)



1077



1078



1079



107A



107B



107C

D.4.4. Directional Behavior

All Myanmar scripts and sub-scripts are written from left to right, then top to bottom, exactly as Latin does. Needless to say that the numerals are written from right to left, as same as Arabic numerals used in Latin encoding.

D.4.5. Default Collation Behavior:

n.a.

D.4.6. There is nothing or no available data at the moment to submit on casing, numeric, default collation behaviors, relevance at markup context, compatibility equivalence, and other Unicode normalization.

D.4.7. Currency information is submitted in a separate proposal.

E. Bibliography

1. Myanmar-Myanmar-English Dictionary
by Myanmar Language Commission
2. Myanmar Orthography
by Myanmar Language Commission, 1999 Edition
3. Mon-English Dictionary
by Halliday
4. English-Mon Dictionary
by Reverend Stevens, ...
5. Shan-English Dictionary
by Cushing
6. The Unicode Standard 3.0
7. The Unicode Standard 4.0
by Unicode Consortium, 2002
8. Karen Primer
9. Mon Primer
10. Shan Primer
11. A Comparative Study of Shan, Poe Karen, Sakaw Karen, and Paoh Scripts
by
12. Collection of Myanmar NLP Research Work and Papers
<http://myanmars.net/unicode/>

1000

Myanmar(Mon)

	100	101	102	103	104	105	106	107	108	109
0	က 1000	တ 1010	င 1020	ူ 1030	ဝ 1040	ဂ 1050	ဇ 1060			
1	ခ 1001	ထ 1011	အ 1021	ေ 1031	၁ 1041	ဗ 1051	မှ 1061			
2	ဂ 1002	ဒ 1012		ဲ 1002	၂ 1042	င 1052	် 1062			
3	ယ 1003	မ 1013	ဒ 1023		၃ 1043	ဗ 1053	ွ 1063			
4	င 1004	န 1014	ဒ 1024		၄ 1044	ဇ 1054	ွ 1064			
5	စ 1005	ပ 1015	၂ 1025		၅ 1045	ဇ 1055	ွ 1065			
6	ဆ 1006	ဖ 1016	၂ 1026	ံ 1036	၆ 1046	ယ 1056				
7	ဇ 1007	တ 1017	ဇ 1027	ံ 1037	၇ 1047	ယ 1057				
8	ဇ 1008	ဘ 1018		ံ 1038	၈ 1048	ွ 1058				
9	၂ 1009	မ 1019	သ 1029		၉ 1049	ွ 1059				
A	၂ 100A	ယ 101A	သ 102A		၁ 104A					
B	င 100B	ရ 101B			၂ 104B					
C	င 100C	လ 101C	ာ 102C		၃ 104C					
D	ဗ 100D	ဝ 101D	ံ 102D		၄ 104D					
E	ပ 100E	သ 101E	ံ 102E		၅ 104E					
F	က 100F	တ 101F	ူ 102F		ဝ 104F					

1000

Myanmar(Shan)

	100	101	102	103	104	105	106	107	108	109
0	၈ 1000	၉ 1010	၁၀ 1020	၁၁ 1030	၁၂ 1040	၁၃ 1050		၁၄ 1070		
1	၁၅ 1001	၁၆ 1011	၁၇ 1021	၁၈ 1031	၁၉ 1041	၂၀ 1051		၂၁ 1071		
2	၂၂ 1002	၂၃ 1012		၂၄ 1032	၂၅ 1042	၂၆ 1052		၂၇ 1072		
3	၂၈ 1003	၂၉ 1013	၃၀ 1023		၃၁ 1043	၃၂ 1053		၃၃ 1073		
4	၃၄ 1004	၃၅ 1014	၃၆ 1024		၃၇ 1044	၃၈ 1054				
5	၃၉ 1005	၄၀ 1015	၄၁ 1025		၄၂ 1045	၄၃ 1055				
6	၄၄ 1006	၄၅ 1016	၄၆ 1026	၄၇ 1036	၄၈ 1046	၄၉ 1056	၅၀ 1066			
7	၅၁ 1007	၅၂ 1017	၅၃ 1027	၅၄ 1037	၅၅ 1047	၅၆ 1057	၅၇ 1067			
8	၅၈ 1008	၅၉ 1018		၆၀ 1038	၆၁ 1048	၆၂ 1058	၆၃ 1068			
9	၆၄ 1009	၆၅ 1019	၆၆ 1029		၆၇ 1049	၆၈ 1059	၆၉ 1069			
A	၇၀ 100A	၇၁ 101A	၇၂ 102A		၇၃ 104A			၇၄ 106A		
B	၇၅ 100B	၇၆ 101B			၇၇ 104B			၇၈ 106B		
C	၇၉ 100C	၈၀ 101C	၈၁ 102C		၈၂ 104C			၈၃ 106C		
D	၈၄ 100D	၈၅ 101D	၈၆ 102D		၈၇ 104D			၈၈ 106D		
E	၈၉ 100E	၉၀ 101E	၉၁ 102E		၉၂ 104E			၉၃ 106E		
F	၉၄ 100F	၉၅ 101F	၉၆ 102F		၉၇ 104F			၉၈ 106F		

1000

Myanmar(Karen)

	100	101	102	103	104	105	106	107	108	109
0	၁ 1000	၂ 1010	၃ 1020	၄ 1030	၅ 1040	၆ 1050				
1	၇ 1001	၈ 1011	၉ 1021	၁၀ 1031	၁၁ 1041	၁၂ 1051				
2	၁၃ 1002	၁၄ 1012		၁၅ 1002	၁၆ 1042	၁၇ 1052				
3	၁၈ 1003	၁၉ 1013	၂၀ 1023		၂၁ 1043	၂၂ 1053				
4	၂၃ 1004	၂၄ 1014	၂၅ 1024		၂၆ 1044	၂၇ 1054		၂၈ 1074		
5	၂၉ 1005	၃၀ 1015	၃၁ 1025		၃၂ 1045	၃၃ 1055	၃၄ 1065	၃၅ 1075		
6	၃၆ 1006	၃၇ 1016	၃၈ 1026	၃၉ 1036	၄၀ 1046	၄၁ 1056		၄၂ 1076		
7	၄၃ 1007	၄၄ 1017	၄၅ 1027	၄၆ 1037	၄၇ 1047	၄၈ 1057		၄၉ 1077		
8	၅၀ 1008	၅၁ 1018		၅၂ 1038	၅၃ 1048	၅၄ 1058		၅၅ 1078		
9	၅၆ 1009	၅၇ 1019	၅၈ 1029		၅၉ 1049	၆၀ 1059		၆၁ 1079		
A	၆၂ 100A	၆၃ 101A	၆၄ 102A		၆၅ 104A			၆၆ 107A		
B	၆၇ 100B	၆၈ 101B			၆၉ 104B			၇၀ 107B		
C	၇၁ 100C	၇၂ 101C	၇၃ 102C		၇၄ 104C			၇၅ 107C		
D	၇၆ 100D	၇၇ 101D	၇၈ 102D		၇၉ 104D					
E	၈၀ 100E	၈၁ 101E	၈၂ 102E		၈၃ 104E					
F	၈၄ 100F	၈၅ 101F	၈၆ 102F		၈၇ 104F					

