SC2/WG2 N3083

L2/05-178

A Sgaw Karen Unicode Proposal
Extending Myanmar to Incorporate Sgaw Karen

Martin Hosken
SIL International and Payap University

Introduction

Sgaw Karen is spoken by approximately 1.5 million people in Myanmar and Thailand. The dominant script of the language group is Myanmar based. The script is in active use with a number of publications, fonts, websites, etc. available using the writing system.

Script Description

The Sgaw Karen writing system is relatively modern and therefore the language has not had time to drift far from its orthography roots. The result is a clean writing system with relatively little complexity when compared with its Burmese base. The following tables list the characters and their corresponding Unicode values, existing and proposed.

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ၡ	ည	တ	ထ	ဒ	န
k	k^h	ɣ	x	ŋ	s/c	s^h/c^h	ʃ	ɲ	t	t^h	d	n
1000	1001	1002	1003	1004	1005	1006	1050	100A	1010	1011	1012	1014

ပ	ဖ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ	အ	ဧ
p	p^h	b	m	ʝ,ʒ/z	r	l	w/v	θ/s	h	ʔ	
1015	1016	1018	1019	101A	101B	101C	101D	101E	101F	1021	1027

Vowels

◌ါ	◌ံ	◌ၢ	◌ု (ု)	◌ူ (ူ)	◌့	◌ဲ	◌ိ	◌ီ
a	i	ɤ	ɯ	u	e	æ	o	ɔ
102C FE00	1036	1051	102F	1030	1037	1032	102D	102E

Tones

ၢ်	ၣ်	း	ၥ်	ၤ
1052	1053	1038	1054	1055

Medials

ှ	ျ	ြ	ၠ	ွ
ဟ	ယ	ရ	လ	ဝ

Numbers

၀	၁	၂	၃	၄	၅	၆	၇	၈	၉
1040	1041	1042	1043	1044	1045	1046	1047	1048	1049

Misc

်
1039 200C

The phonology of Sgaw Karen greatly aids in the creation of a simple script. The basic syllable pattern is C (C)? V T? (C K)?. Where C is a consonant, V a vowel, T a tone mark and K the visible virama (`U+1039 MYANMAR SIGN VIRAMA U+200C ZERO WIDTH NON-JOINER`). Final consonants are rare in Sgaw Karen, only occurring in borrowed words. In the case that there is a tone mark and the vowel is -a, the vowel is removed. Sgaw Karen also has an inherent vowel which is a short -a. Sgaw Karen has no concept of chaining syllables.

Sorting

Sgaw Karen has four levels of sorting: consonant, medial, vowel, tone[1]. Thus a consonant followed immediately by another consonant sorts before the same consonant with a vowel sorts before the same consonant with a medial. The relative orders of the categories are as listed in the tables above.

Encoding

A straight encoding for Sgaw Karen is relatively straightforward. Allocate a codepoint to each of the letters listed above and you have a very workable encoding.

Rendering

Sgaw Karen renders much like Myanmar except for a few key stylistic differences:

- The medial -wa (`U+1039 U+101D`) is in the shape of a teardrop rather than a circle
- `U+1037 MYANMAR SIGN DOT BELOW` is always rendered to the left of any medial present (or inside in the case of -ra).

[1] Finals are very rare and it is unknown where they occur in the sorting.

Integration Into Unicode

Integrating Sgaw Karen into the existing Unicode model as an extension to the Myanmar script block is a different matter. Almost every character that is not already encoded has some issue associated with it. Here we examine each new character and changed character in turn.

Consonants

There is only one consonant that is different from those in the standard Myanmar block, and one that has different behaviour:

ၡ - sha

This can be constructed visually using the sequence: `U+101B MYANMAR LETTER RA U+1039 MYANMAR SIGN VIRAMA U+101F MYANMAR LETTER HA`. But this letter is part of the alphabetical order of Sgaw Karen and on the same basis that U+1008 `MYANMAR LETTER JHA` receives its own code, it is proposed that this character receive its own code.

ဧ

In Myanmar this is an independent vowel. In Sgaw Karen this character is a consonant much like `U+1021 MYANMER LETTER A` except that it has no sound. In Sgaw Karen glottal is a contrastive phoneme. Therefore this character, in Sgaw Karen, may take a diacritic vowel.

Vowels

ါ - a

The -a vowel in Sgaw Karen only has the tall form as shown here, it does not have a normal Myanmar form as shown for U+102C `MYANMAR VOWEL SIGN AA`, or its complex behaviour (regarding whether it should render short or tall). The choice is whether to give this character its own code or to use a variant selector. Using a specific code opens up the opportunity for considerable confusion in the main Myanmar encoding with people using whichever code according to a visual preference. Instead a variant selector should be used in all cases of Sgaw Karen. Thus this character is encoded as `U+102C MYANMAR VOWEL SIGN AA U+FE00 VARIANT SELECTOR-1`. If people start to use the variant selector in Myanmar language, which they should not, at least all processes can ignore it, as per the specification of a variant selector.

ၢ - ɤ

This is a new character and requires a new code.

Tones

To aid in analysis, we take the tones out of order:

ာ် - tone 2

This glyph occurs in Burmese, in the vowel sequence: ော် which is a tone variant of ော. Notice that in both cases the ာ vowel is significant as a vowel and is integral to the compound vowel. In Sgaw, the vowel part of the glyph is not considered. The vowel + killer combination is just a productive mechanism to create more shapes for use as tone marks. Therefore, encoding the ာ် as a sequence is wrong. Instead the tone should be given its own code. Care should be taken in the encoding description that this code is only for use as a tone mark and not as a presentation form of the sequence in Myanmar. In addition, using the Myanmar sequence

(U+102C U+1039 U+200C) would be liable to mis-rendering since the U+102C may take a tall form, which this character never does.

ၢ် - tone 1

A glyph based encoding would encode this as ၢ+် . For the same reasons as listed for ာ်, this character should receive its own code.

း - tone 3

This tone uses the code U+1038 MYANMAR SIGN VISARGA since in Burmese this is a sign, which happens also to be used in a similar way, for marking tone. The character's behaviour in terms of rendering, clustering, etc. is identical in the two writing systems.

ၣ် - tone 4

This is a new character and should receive its own code.

ၤ - tone 5

This is a new character and should receive its own code..

Medials

This is where the interaction between the Sgaw Karen writing system and the Burmese writing system gets messy. Since the Unicode encoding for Myanmar associates medials with their corresponding base consonants, we examine those same correspondences in Sgaw Karen and compare.

Medial	***Myanmar/Unicode base***	***Sgaw Karen base***
ှ	ဟ (U+101F)	ဟ (U+101F)
ၠ		**ယ (U+101A)**
ြ	ရ (U+101B)	ရ (U+101B)
ျ	**ယ (U+101A)**	လ (U+101C)
ွ	ဝ (U+101D)	ဝ (U+101D)

The associations between medials and base consonants that hold for the Burmese language do not hold for the Sgaw Karen language.

There are various options open to us in resolving the issue:

- Encode new versions of U+101A and U+101C for Sgaw Karen with the correct associations.
- Assume they are font variants.
- Use a variant selector.
- Dissociate all the medials from their base characters in the Myanmar script. Thus we create new codes for the medials and provide no encoded correspondences between medial and base character. If a language needs to make an association, that is part of the language specific processing and not part of the encoded character.

We examine each in turn.

New Base Characters

Since the association between a base character and its subjoined form is part of the definition of the base character, if the association changes it is necessary to encode a new base character. Thus for Sgaw Karen we would need to encode two new characters to correspond to U+101A MYANMAR LETTER YA and U+101C MYANMAR LETTER LA. The problem with this solution is that it introduces two new characters with identical base form to two existing characters. This is liable to cause considerable confusion with people using the wrong character.

Font Variants

The easiest approach from an encoding point of view is to just say that these different forms of U+101A and U+101C are simply font variants. If you want to view text as Sgaw Karen you have to use a different font. But Unicode is a text encoding and is designed to display text legibly in all languages without knowledge of the language of the text. Legibility certainly does not include beauty or culturally appropriate shape variation, but it does include not displaying a glyph that is used for something else. Thus the sequence (U+1000 MYANMAR LETTER KA U+1039 MYANMAR SIGN VIRAMA U+101A MYANMAR LETTER YA) would be displayed as: ကျ in Burmese, but this would be wrong since it would imply a completely different medial in Sgaw Karen.

Variant Selector

Another alternative is to encode the alternative forms of the medials using a variant selector. Thus we would encode the alternative form of the medial -ya as U+1039 MYANMAR SIGN VIRAMA U+101A MYANMAR LETTER YA U+FE00 VARIANT SELECTOR-1. Medial -la would need greater care since normally U+1039 MYANMAR SIGN VIRAMA U+101C MYANMER LETTER LA is considered to be the start of a new syllable rather than a medial. Instead, using the approach in UTN#11, we would need to encode medial -la as U+200D ZERO WIDTH JOINER U+1039 MYANMAR SIGN VIRAMA U+101C MYANMAR LETTER LA U+FE00 VARIANT SELECTOR-1.

While variant selectors are a wonderful way to dig oneself out of an encoding hole, they were not envisioned to be used for common letters (which both of these characters are in Sgaw Karen). In addition, the Sgaw Karen form of medial -ya is also used in other scripts, including Mon, where it is used for a medial -la. Once we start down the road of encoding by variant selector we introduce a parallel encoding mechanism, greatly increasing the complexity of Unicode. But this does not mean that the variant selector approach will not work.

Dissociating Medials

At first thought, dissociating the Myanmar medial characters by creating new codes for them and deprecating the current mechanism of encoding them using a U+1039 MYANMAR SIGN VIRAMA, seems excessive. A more thorough discussion of this issue is covered in a related paper: “Dissociating Myanmar Medials”. If such an approach is taken, the encoding of Sgaw Karen medials becomes straightforward, a new medial is added for -ya and since there is no longer a direct association between the medials and their base characters, the other medials can be reappropriated according to their use in the language.

Misc

် - killer

This character is encoded the same way as in the Myanmar Unicode encoding, as U+1039 MYANMAR SIGN VIRAMA U+200C ZERO WIDTH NON JOINER.

Conclusion

This proposal proposes the addition of 1 new consonant code, 1 new vowel code and 4 new tone mark codes. In addition it proposes the dissociation of medials from base consonants in the Myanmar block and the addition of 1 extra medial character. The net result is the addition 7 new codes for Sgaw Karen and 4 new codes for Myanmar, if accepted. There is no particular recommendations regarding code allocation for these new codes.

```
103E;MYANMAR SGAW MEDIAL LETTER YA;Mn;0;L;;;;;N;;;;;
1050;MYANMAR LETTER SHA;Lo;0;L;;;;;N;;;;;
1051;MYANMAR VOWEL SIGN SGAW U;Lo;0;L;;;;;N;;;;;
1052;MYANMAR SGAW TONE LETTER 1;Lo;0;L;;;;;N;;;;;
1053;MYANMAR SGAW TONE LETTER 2;Lo;0;L;;;;;N;;;;;
1054;MYANMAR SGAW TONE LETTER 4;Lo;0;L;;;;;N;;;;;
1055;MYANMAR SGAW TONE LETTER 5;Lo;0;L;;;;;N;;;;;
```

The Myanmar additions are covered in their own proposal, which takes precedence over these codes listed.

```
103A;MYANMAR SEMIVOWEL SIGN YA;Mn;0;L;;;;;N;;;;;
103B;MYANMAR SEMIVOWEL SIGN RA;Mn;0;L;;;;;N;;;;;
103C;MYANMAR SEMIVOWEL SIGN WA;Mn;0;L;;;;;N;;;;;
103D;MYANMAR SEMIVOWEL SIGN HA;Mn;0;L;;;;;N;;;;;
```

Bibliography

http://www.drumpublications.org	publisher of materials in Sgaw Karen
http://www.kwekalu.net	indigenous nationalist newspaper
http://www.ktwg.org	Karen Teacher's Working Group website
Hosken, Martin and Maung TunTunLwin	“Representing Myanmar in Unicode” (Unicode Technical Note #11)
Hosken, Martin	“Dissociating Myanmar Medials” (Unicode Document L2/05-???)

Proposal Form

A. Administrative

1. **Title: ___Sgaw Karen____**

2. Requester's name: ___Martin Hosken____

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

4. Submission date: ___28 July 2005____

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): _no__________

Proposed name of script: ______________________________________

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

Name of the existing block: ________Myanmar___________________________

2. Number of characters in proposal: __6____________

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

4. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000): ____2__

Is a rationale provided for the choice? ____no_________

If Yes, reference: ____

5. 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 ISO/IEC 10646-1: 2000? ____yes_______

b. Are the character shapes attached in a legible form suitable for review? __yes___

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

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

7. 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? _see bibliography_____________

8. 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)?

______presentation, sorting____

9. 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 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.

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? ___Thai user community_____________________________

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: ____

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

Reference: ____

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

If YES, where? Reference: ___Thailand, Myanmar_________________________

6. 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______

If YES, is a rationale provided? __additions to existing BMP block__

If YES, reference: ____

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
__yes, since one script__

8. Can any of the proposed characters be considered a presentation form of an existing
character or character sequence? ___yes______

If YES, is a rationale for its inclusion provided? ____yes_______

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? ____yes_______

If YES, is a rationale for its inclusion provided? ____yes_______

If YES, reference: ____

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

If YES, is a rationale for its inclusion provided? ____yes_______

If YES, reference: ____

11. 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______

If YES, is a rationale for such use provided? _____yes______

If YES, reference: ____

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

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 character(s)? ______no______

If YES, is the equivalent corresponding unified ideographic character(s) identified? ____________

If YES, reference: ____