

Representing Myanmar in Unicode

Details and Examples Version 3

Martin Hosken¹

Introduction

This document aims to give guidance on the encoding of text using the Myanmar script. Since the script is used for a number of orthographies covering different languages, the development of this document is ongoing. It aims to bring together the results of consensus between experts in the encoding of the various orthographies using the script. In terms of the Unicode standard, this document is purely informative since it is concerned with issues not covered by that standard. But within the country, and by developers of the script, this document has been accorded a certain degree of authority. This provides further encouragement to maintain this document and update it as new issues arise.

Readers interested in following the history of the development of this script are recommended to read the different versions of this document, rather than expecting to find this document containing all versions of itself within it.

The Myanmar script is used for a number of languages. This means that when considering the script as a whole, care must be taken not to over specify constraints on what character sequences should be considered valid or in error. The temptation is to use script level sequence constraints as a form of spell checking. But spell checking is inherently language specific. The result is that script constraints need to be the lowest common denominator of all the orthographies supported by the script. The orthography list is not closed: we have not described all the existing orthographies yet; languages change and develop and their orthographies with them. As a result, script constraints cannot simply be the intersection of all known writing system constraints, but must take a more intentional approach. The basic principle used here is not to try to constrain what users can generate, but only to ensure that there are no two different valid sequences that look the same. This is achieved by specifying the relative order of characters in a sequence as a sequence of slots that can take any code from the set specified for that slot, but not to specify which slots may be filled with which of the codes possible for that slot, if any. Implementations may well add further, language specific, constraints to help their users.

A further concern when reading a developing document such as this is the stability criteria. What can we be sure about going into the future. The approach taken in this document follows the core principle of stability in Unicode: Any valid data today will always remain valid. This requires that any changes to the sequence order, for example, will always be to loosen it. Thus more sequences will be allowed rather than less. This means that invalid data today may not always remain invalid in future versions of this document.

Introduction to Version 2

The first edition of this technical note addressed the issue of how Myanmar text was encoded using the Unicode standard as it stood until version 5.1. With Unicode 5.1 various new characters were added to the Myanmar block which had the effect of simplifying the encoding model considerably. Such a change could only come about with agreement from all implementors and those with existing data because they will need to update and change to the new model. This is nearly impossible to achieve if existing implementations are already in widespread use, which was not the case at the time for the Myanmar block. In addition, such a change was necessary to facilitate the encoding of minority scripts. So with a necessity and a unique opportunity for change, the characters were added and the encoding model simplified.

¹ SIL International and Payap University, Chiang Mai, THAILAND

This technical note describes the simplified model and keeps the older model description as a later section for comparison. The information is structured to follow closely the previous edition of this technical note.

The author wishes to thank the Myanmar Language Commission, the Myanmar NLP Lab and the Myanmar Computer Federation for reviewing and providing input to this version of the document.

Unicode 5.1 Model

Basic Myanmar

The basic consonants and vowels are relatively obvious in how they are encoded. Thus:

တ	1005 102C	letter
---	-----------	--------

Here we show the Myanmar word, the underlying Unicode codes that would be stored to represent this and an English gloss of the word. As this example shows, characters are stored in the order in which they are read.

ခါ	1001 102B	to shake
သိက္ခာ	101E 102D 1000 1039 1001 102C	dignity
သဒ္ဓါ	101E 1012 1039 1013 102B	faith

In this example, we highlight the code of interest. Notice how the ့ (U+102B MYANMAR VOWEL SIGN TALL AA) has a different code to the ္ (U+102C MYANMAR VOWEL SIGN AA). The Myanmar character underlying the two codes is the same, and there are rendering rules that can give the correct form, so why has the tall -aa been given its own code? The primary reason is that Sgaw Karen, among other minority scripts, only has the tall form, and so a rendering system that works for the Myanmar language is not going to work for Sgaw Karen and vice versa. A Myanmar language specific keyboarding implementation could choose to enforce a particular variant of the -aa vowel in the context of certain consonants (in Burmese following ခ, ဝ, င, ဒ, ဝ, or ဝ), medial combinations and syllable chainings, but this is not required.

ညို	100A 102D 102F	brown
ထိုး	1011 102F 1036 1038	to tie

Notice how the two forms of ့ (U+102F MYANMAR VOWEL SIGN U) have the same code. It is up to the rendering system to choose which form should be shown and different fonts can have different rules depending on the designer's preference.

U+1031 –e vowel

We will see later why the vowels are stored in this relative order. But for now it is important to note that the Unicode standard states that vowels are stored after the consonant, according to how they are read, regardless of where they are rendered. This introduces one of the complexities of implementing Myanmar script:

နေ	1014 1031	the sun
ပေါ	1015 1031 102B	plentiful

The ့ vowel is rendered in front of the consonant that it is read (and so stored) following. Notice that this says nothing about the relative order for typing, but it does mean that anyone implementing the Myanmar script needs to take special care of this character. In general people are used to and want to type the ့ vowel in front of the consonant, and so implementors need to address issues of keyboarding as well as rendering.

Medials

The medial characters have their own codes and are always stored after the base consonant and before any vowels. Although the character ့ has traditionally been typed in non-Unicode fonts before the consonant, it is consistent with normal spelling to store U+103C MYANMAR CONSONANT SIGN MEDIAL RA after the consonant.

ဖျား	1016 103B 102C 1038	fever
ကြေး	1000 103C 1031 1038	grime

မွ့	1019	103D	1031	1038	give birth
မူ ူ	1019	103E	102F		regard important

Syllable Chaining

In the case of syllable chaining, subjoined characters are not given their own codes. Instead a virama character is used to indicate that the following character is subjoined and should take a subjoined form.

ပတ္တ	1015	1010	1039	1010	102C	hinge
------	------	------	-------------	------	------	-------

Devoweliser

There are two ways of representing the devowelising process. The first is by creating a syllable chained form, using U+1039 to mark the devowelising (as shown above). The second is to use the visible virama character ◌် (U+103A MYANMAR SIGN ASAT) in conjunction with a base consonant.

ထင်	1011	1004	103A		think
ကြည်	1000	103C	1009	103A	avoid
ကော်	1000	1031	102C	103A	glue

The second example also illustrates that င် is encoded with U+1009 followed by U+103A even though the glyph shape closely resembles the independent vowel ဥ U+1025 MYANMAR LETTER U. Keyboard implementors may wish to enforce this.

The third example is not a true devowelising, but it shows that U+103A can also be used as a tone mark in combination with U+102B and U+102C.

Kinzi

The remaining issue regarding representation needed for the modern Myanmar language is how kinzi is represented in Unicode. Glyph based encodings give the kinzi its own code. But linguistically, the kinzi is merely a special form of a devowelised nga (U+1004 MYANMAR LETTER NGA). We encode kinzi as a devowelised nga with the following letter underneath, subjoined. But the difference is that when rendered, the devowelised nga changes shape and the subjoined base character remains a full character. Thus we use U+1004 U+103A U+1039.

စကြ်	1005	1004	103A	1039	1000	103C	path
သင်္ဘော်	101E	1004	103A	1039	1018	1031	ship
					102C		

Like the –e vowel, kinzi is particularly problematic to implement since people want to type it following the base consonant and it also needs careful handling during rendering.

Diacritic storage order

It is possible for a Myanmar syllable to have a number of diacritics surrounding a base consonant. Since all these diacritics are not spacing, how do we know in which order they should be stored? For example, ညို can be stored as U+100A U+102D U+102F or as U+100A U+102F U+102D. But what happens if one person stores it one way and then someone searches for that word spelled the other way? It is important that there is a consistent way of storing strings so that applications can work consistently.

care has been taken that if the wrong slot is used in a significant way (there is an A vowel, for example) that there will be a visible difference that will indicate the misspelling.

There is one language in which there is a possible invisible ambiguity and that is Mon. Mon treats anusvara (U+1036) as a final nasal and as such it may follow a U+102C. Apart from its linguistic function, this is the same position as in Burmese. In Mon, though, anusvara may also follow U+102B. But when that happens, it is rendered above the preceding consonant. This may result in two valid sequences, according to the above table, rendering the same. This requires us to add a further constraint that is not captured by the chart above: U+1036 may not directly precede U+102B. We can say this because there are no known situations in which U+1036, acting as a vowel, is used in conjunction with the vowel U+102B. For more details on Mon see the section on Mon further down this document.

In Pwo Karen, one can have occurrences of two lower dots as in: ၵၵ U+1000 U+1060 U+1036 U+106B U+1036. Likewise in Sgaw Karen one can have two occurrences of U+1062 together as in: ၵၵ U+1000 U+1062 U+1062 U+103A. Both of these examples are covered by the sequence chart above.

Normalization

The chart shown in this document differs from what one might expect with regard to the relative order of visible virama and lower dot. The normal typing order of these two characters is the visible virama first as part of the final and then the tone mark. But due to an oversight in the standard checking, the combining orders of visible virama and lower dot were set³ such that any normalization process will order them with the lower dot first, but only when they are stored directly after each other. Thus U+103A U+1037 will always be normalized to U+1037 U+103A.

This makes no difference to keyboard entry and people should still be able to type visible virama before lower dot. But it impacts rendering, searching and sorting. It is best if such processes can handle both orders of encoding U+103A U+1037 and U+1037 U+103A, recognising that after normalization the order will be U+1037 U+103A regardless of the order text was entered.

A common question is whether the uu independent vowel is spelled U+1026 or U+1025 U+102E. According to the Unicode standard, the answer to this question is simple: either. Since the two sequences are canonically equivalent, a process needs to treat them identically.

Advanced Issues

So far we have covered what is explained in the Unicode Standard⁴. In this section we examine some of the more difficult areas of the Myanmar language including some implementation details regarding line breaking and sorting; further examination of the kinzi question; contractions and some issues with respect to Old Myanmar.

Line breaking

Myanmar does not have interword spaces like English. Instead spaces are used to mark phrases. Some phrases are relatively short (two or three syllables, 1.5em, or 2.3 times the width of U+1000 ၵ) while others can be quite long (8.5em or 13 times the width of U+1000 ၵ). A common approach to addressing line breaking issues is to adjust the phrase spacing so that a line breaks at a phrase break. If this approach fails and a phrase must be continued onto a second line, U+200B ZERO WIDTH SPACE may be used to indicate a possible line break point in the text.

The problem with this approach is that when phrases are quite long or a lot of text is to be typeset, the manual adjustment of phrasing or the introduction of zero width spaces can be onerous. A further option is to break lines automatically within phrases when needed. The clearest solution is to have a line break occurring at a word boundary, but since there are no word breaks in Myanmar this is not immediately possible. Most words, though, are mono-syllabic and so a mechanism of breaking lines at syllable boundaries is usually sufficient. From this we can say that a syllable break may occur before a Myanmar digit, an independent vowel, one of the various signs or a base consonant so long as the consonant:

- is not devowelised with an asat and

³ Due to the stability criteria of the Unicode standard, once a combining order is set in the standard, it is impossible to change it for that character.

⁴ Version 5.1

- has no stacked consonant below it and
- is not a kinzi.

These same syllable breaking rules are used for sorting purposes, with the addition of non-line breaking syllable breaks, such as those occurring between the two characters in a syllable chain. For example these phrases show possible inter-syllable line breaks.

ကောင်လေးတွေ	1000 1031 102C 1004 103A 101C 1031 1038	
ကျောင်းကိုသွားကြ	1010 103D 1031 1000 103B 1031 102C	the kids are
တယ်။	1004 103A 1038 1000 102D 102F 101E	going to
	103D 102C 1038 1000 103C 1010 101A	school
	103A 104B	
အိပ်ခန်းတံခါးကို	1021 102D 1015 103A 1001 1014 103A 1038	to the
	1010 1036 1001 102B 1038 1000 102D	bedroom
	102F	door

Notice how in the second example the word 1010 1036 | 1001 102B 1038 is a single word with multiple syllables. Is there some way, without a dictionary, that we can ensure that the word is not line broken? There is a Unicode character : U+2060 WORD JOINER. The role of this character is to indicate a non-breaking point in a text. Lines should not be broken at that point. Therefore, if we want to ensure that no line-break occurs at the syllable boundary within our poly-syllabic word, we can insert a U+2060 into our data stream between the two syllables and a rendering engine should not break a line at that point. Thus:

အိပ်ခန်းတံခါးကို	1021 102D 1015 103A 1001 1014 103A 1038	to the
	1010 1036 2060 1001 102B 1038 1000	bedroom
	102D 102F	door

In summary, therefore, we propose three levels of line breaking support: breaking at phrase spaces; breaking at syllable breaks and support for polysyllabic words. A rendering engine may choose the sophistication of line breaking support it provides.

Sorting

Sorting Myanmar strings is a complex process involving significant string transformation and four levels of comparison. The string transformation is a syllable based operation for which the identification of syllable boundaries (but not word boundaries) are required. The same techniques that are used for line-breaking, therefore, may be used for sorting.

Kinzi revisited

One of the significant improvements brought about by the addition of the asat character is that kinzi is now unambiguously encoded. Thus:

အေဝံ	1021 1004 103A 1039 101D 1031
အေဝု	1021 1004 103D 1031

There are two other characters that follow this same encoding model. Kinzi in Mon is represented using the Mon letter nga U+105A. Sanskrit has a kinzi type character based on the letter ra U+101B: ᳚ which is encoded following the same model as: U+101B U+103A U+1039.

Contractions

The Myanmar language has a system of double acting consonants, where a consonant acts as both the final of a syllable and the initial of a following syllable. These are significant for sorting purposes. Double acting consonants are rare, but occur in two common words.

ယောက်ျား	101A 1031 102C 1000 103A 103B	man, husband
	102C 1038	
ကျွန်ုပ်	1000 103B 103D 1014 103A 102F	I (1 st person singular)
	1015 103A	

Notice how the visible virama (103A) occurs immediately after the double acting consonant. This position is not listed in the standard diacritic order, but is most appropriate for a double acting consonant. In order to identify such contractions, we propose that the visible virama be stored immediately following the

consonant. This storage approach will also affect the syllable definition since a devowelised consonant with a vowel acts like a normal base consonant with its preceding syllable break.

There are also words with double acting consonants which are unmarked. Since these are unmarked, it has been decided that despite their etymology, these words should be sorted as if there were no double acting consonant.

ဝါကျ	101D 102B 1000 103B	sentence
ဂိမ္မာန်	1002 102D 1019 103E 102C 1014 103A	summer

Old Myanmar

There are a few issues that storing old Myanmar text introduce, although again, most of these are resolved due to the simplified encoding model.

Stacked Ya

There are occasions where a medial ya (U+103B) representation is used for a stacking ya. What is needed is a syllable break between the base consonant and the ya. Thus we propose:

ဥယျာန	1025 101A 200C 103B 102C 1014	ဥယျာဉ်	garden/orchard
-------	---	--------	----------------

The extra column gives the modern spelling of the word. The use of U+200C ZERO WIDTH NON-JOINER indicates the break in the syllable. It makes no difference to rendering and is only used in Pali sorting.

LaSwe (Medial la)

In some words, a subscript la (U+101C) acts as a medial rather than as the start of a new syllable. This is simply encoded using a virama:

ဣ	1000 1039 101C	ကျ	drop
ဣ်	1000 103B 1039 101C 1015 103A	ကျ်	tight

AMyint (Archaic tone mark)

Old Myanmar includes a medial form of U+1021 which causes no particular problems since it is just treated as any other medial, and occurs very rarely.

နိယံ	1014 102D 101A 1039 1021 103A	နေ့	day
------	--------------------------------------	-----	-----

Sgaw Karen

With the addition of support for a number of minority languages which are based on the Myanmar script, it is possible that some characters that look like presentation forms of sequences found in Myanmar will be encoded. For example in Unicode 5.1 there is the character: ရှ (U+1061 MYANMAR LETTER SGAW KAREN SHA). This looks like the sequence ရ (U+101B MYANMAR LETTER RA) followed by ှ (U+103E MYANMAR CONSONANT SIGN MEDIAL HA). But the two are completely unrelated and the sequence must always be used in Myanmar language and the unit in the Sgaw Karen language.

Minority languages often have different ways of rendering certain sequences. Where the difference does not result in illegibility, there is no need to encode the difference. For example, in Sgaw Karen, the word: ဝ့ is rendered ဝ့ with the lower dot (U+1037) being pushed to the left, rather than to the right as it is rendered in the Burmese word .

Mon

Mon introduces some interesting extensions to the script.

Mon has a sequence U+102C U+1036 ဝံ and correspondingly U+102B U+1036, but here the dot is rendered over the previous consonant: ဝံ and for consistency this is encoded with the dot after the vowel.

ကံ် U+1000 **U+1036** U+102C U+103A

ကံ် U+1000 U+102C **U+1036**

The nga letter in Mon is encoded U+105A င and not U+1004 c as in Burmese. Independently, these characters look very different. But in the context of something occurring below the character, the Mon nga (U+105A) loses its tail. Thus a Mon kinzi is encoded using U+105A U+103A U+1039. In addition, the medial form of Mon nga is simply the tail: င (U+1039 U+105A).

Mon has a contraction involving medial ha (U+103E) and killer (U+103A) which needs further research, before this version of the document is released, as to the sequence ordering involved. The current order is this, but does that make sense linguistically?

၄ 1005 103E 103A

၄ 1005 103E 1031 103A

၄ 1005 103E 1031 102C 103A

Searching and Comparison

The approach we have taken here to add markers to handle syllable breaking can introduce problems when searching and during string comparison. Since the codes used do not make any change to rendering, it is likely that in many cases the codes will be left out of a text. Therefore there is a need for searching and comparison code to take into account and ignore the extra codes inserted.

The following codes should be ignored when searching and comparing: U+200B ZERO WIDTH SPACE and U+2060 WORD JOINER. In addition, with the simplified encoding, U+200C ZERO WIDTH NON-JOINER may also be ignored.

References

Bechert, et al 1979, *Burmese Manuscripts, Part 1* Wiesbaden.

Department of the Myanmar Language Commission 1993, *Myanmar – English Dictionary* Ministry of Education, Union of Myanmar.

Okell, John 1994, *Burmese: An Introduction to the Script* SOAS, London.

The Unicode Consortium 2003, *The Unicode Standard, Version 4.0* Addison-Wesley, Massachusetts.