Representing Myanmar in Unicode
Details and Examples

Martin Hosken[1]

Introduction

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.

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.

[1] SIL International and Payap University, Chiang Mai, THAILAND

Unicode 5.1 Model

Basic Myanmar

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

စာ	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, ဂ, င, ဒ, ပ, or ဝ), medial combinations and syllable chainings, but this is not required.

ညို	100A 102D **102F**	brown
ထုံး	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
မွေး	1019 **103D** 1031 1038	give birth
မှု	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.

ပတ္တာ	1015 1010 **1039** 1010 102C	hinge

Devoweliser

There are two ways of representing the devowelising process. The first is by creating a medial or 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
ကြဉ်	1000 103C 1009 **103A**	avoid
ကော်	1000 1031 102C **103A**	glue

The second example also illustrates that ဉ် 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.

စင်္ကြံ	1005 **1004 103A 1039** 1000 103C 1036	path
သင်္ဘော	101E **1004 103A 1039** 1018 1031 102C	ship

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.

The following list gives the relative order that each diacritic should be stored, if it occurs, following a base consonant.

Name	Specification	Example
Consonant	`[U+1000 .. U+102A, U+103F, U+104E]`[2]	က
Asat[3]	`U+103A`	◌်
Stacked	`U+1039 [U+1000 .. U+1019, U+101C, U+101E, U+1020, U+1021]`	◌္ဒ
Medial Y	`U+103B`	◌ျ
Medial R	`U+103C`	◌ြ
Medial W	`U+103D`	◌ွ
Medial H	`U+103E`	◌ှ
E vowel	`U+1031`	◌ေ
Upper Vowel	`[U+102D, U+102E, U+1032]`	◌ီ
Lower Vowel	`[U+102F, U+1030]`	◌ု
A Vowel	`[U+102B, U+102C]`	◌ာ
Anusvara	`U+1036`	◌ံ
Visible virama	`U+103A`	◌်
Lower Dot	`U+1037`	◌့
Visarga	`U+1038`	◌း

Notice the general order of: initial consonant cluster, vowels, tones.

For example:

ပသျှူး	`1015 101E 103B 103E 1030 1038`	Malay
မြွှာ	`1019 103C 103D 103E 102C`	segmentalize
သျှောင်	`101E 103B 103E 1031 102C 1004 103A`	top knot

Notice also that the diacritic storage order does not define a phonetic syllable. If `Asat` or `Stacked` are present, then a syllable break occurs in the middle of the order, following them. In addition, A syllable containing a devowelised consonant will follow the order twice, once for the main consonant and once for the devowelised one.

Advanced Issues

So far we have covered what is explained in the Unicode Standard[4]. 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` က) while others can be quite long (8.5em or 13 times the width of `U+1000` က). 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.

[2] Notice the extension of the list here to include independent vowels. The Unicode Standard V4.0 only lists values up to U+102A. U+104E has changed glyph and can function in consonant position as in ၎င်း (`104E 1004 103A 1038`)

[3] Only for use with Kinzi and contractions

[4] Version 5.1

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
- 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.

ကောင်လေးတွေကျောင်း ကိုသွားကြတယ်။	`1000 1031 102C 1004 103A | 101C 1031 1038 | 1010 103D 1031 | 1000 103B 1031 102C 1004 103A 1038 | 1000 102D 102F | 101E 103D 102C 1038 | 1000 103C | 1010 101A 103A 104B`	the kids are going to school
အိပ်ခန်း**တံခါး**ကို	`1021 102D 1015 103A | 1001 1014 103A 1038 |` **`1010 1036`** `|` **`1001 102B 1038`** `| 1000 102D 102F`	to the bedroom 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 that was added for version 4.0: `U+2060 WORD JOINER`. Previous to this the character `U+FEFF ZERO WIDTH NON-BREAKING SPACE` was used. Since `U+FEFF` is most commonly used at the start of a Unicode text file to both identify it as being Unicode data and to indicate the encoding form of the data, `U+2060` was added to the standard to take over the function of zero width non-breaking space. 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:

အိပ်ခန်း**တံခါး**ကို	`1021 102D 1015 103A | 1001 1014 103A 1038 | 1010 1036` **`2060`** `1001 102B 1038 | 1000 102D 102F`	to the bedroom 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:

အင်္ဝေ	`1021 1004` **`103A 1039`** `101D 1031`
အငွေ	`1021 1004` **`103D`** `1031`

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

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
ဂိမှာန်	1002 102D 1019 103E 102C 1014 103A	momentum

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:

က္လ	1000 1039 101C	ကျ	drop
ကျ္လပ်	1000 103B 1039 101C 1015 103A	ကျပ်	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.

နိယ္အ်	1014 102D 101A **1039 1021** 103A	နေ့	day

Minority Language Extensions

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: ၡ (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.

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.

Pre Unicode 5.1 Encoding Model

Introduction

This section contains most of the text of the original edition of UTN#11 and was written in conjunction with Maung Tuntunlwin[5].

Basic Myanmar

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

စာ	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 **102C**	to shake

In this example, we highlight the code of interest. Notice how the ါ has the same code as the ာ (`U+102C MYANMAR VOWEL SIGN AA`) and that it is up to the rendering system to decide which form of the character is to be displayed. The same goes for diacritics. There is only one code for a particular character and it is up to the rendering system to ensure that the diacritic is appropriately placed.

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

Here the two forms of ု (`U+102F MYANMAR VOWEL SIGN U`) are decided by the rendering system.

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	1014 **1031** 102C	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 and Syllable Chaining

So much for what is clearly visible on the Unicode chart. What about all those glyphs that are not there? How are words including medials or involve syllable chaining stored?

ပတ္တာ	1015 1010 **1039** 1010 102C	hinge
ဖျား	1016 **1039** 101A 102C 1038	fever
ကြေး	1000 **1039** 101B 1031 1038	grime
မွေး	1019 **1039** 101D 1031 1038	give birth
မှု	1019 **1039** 101F 102F	regard important

In the linguistic model, a medial is formed by devowelising the inherent vowel of the preceding consonant. Likewise, for a syllable chained letter, the inherent vowel at the end of the previous syllable is devowelised.

[5] Myanmar World Distribution

In Unicode this devowelising process is marked using the virama code (`U+1039 MYANMAR SIGN VIRAMA`). Thus we store a consonant followed by the virama and then follow it with the consonant of interest.

Devoweliser

There are two ways of representing the devowelising process. The first is by creating a medial or syllable chained form, using `U+1039` to mark the devowelising. The second is to use the visible virama character (◌်) in conjunction with a base consonant. But if `U+1039` is being used to mark medials and syllable chaining, how is the visible character to be represented? The Unicode standard gives the answer. The sequence `U+1039 MYANMAR SIGN VIRAMA` followed by `U+200C ZERO WIDTH NON-JOINER` is used to represent a visual virama (◌်).[6]

ထင်	1011 1004 **1039 200C**	I think

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)`. Thus we encode kinzi as `U+1004 U+1039`.

စင်္ကြံ	1005 **1004 1039** 1000 1039 101B 1036	path

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.

The following list gives the relative order that each diacritic should be stored, if it occurs, following a base consonant.

Name	Specification	Example
kinzi	`U+1004 U+1039`	◌င်္
Consonant	`[U+1000 .. U+102A]`[7]	ဃ
Stacked	`U+1039 [U+1000 .. U+1019, U+101C, U+101E, U+1020, U+1021]`	◌္ဒ
Medial Y	`U+1039 U+101A`	◌ျ
Medial R	`U+1039 U+101B`	◌ြ
Medial W	`U+1039 U+101D`	◌ွ
Medial H	`U+1039 U+101F`	◌ှ
E vowel	`U+1031`	ေ◌
Lower Vowel	`[U+102F, U+1030]`	◌ု
Upper Vowel	`[U+102D, U+102E, U+1032]`	◌ီ
A Vowel	`U+102C`	◌ာ
Anusvara	`U+1036`	◌ံ
Visible virama	`U+1039 U+200C`	◌်
Lower Dot	`U+1037`	◌့
Visarga	`U+1038`	◌း

[6] For fallback purposes, `U+1039` also displays a ◌် if not followed by a consonant. This is an implementation detail and is not used in spelling words. I.e. all such occurrences should be considered wrong spellings.

[7] Notice the addition to the list of independent vowels. The Unicode Standard v4.0 only lists values up to U+1021.

Notice the general order of: initial consonant cluster, vowels, tones.

For example:

ပသျှူး	1015 101E 1039 101A 1039 101F 1030 1038	Malay
မြွှာ	1019 1039 101B 1039 101D 1039 101F 102C	segmentalize
သျှောင်	101E 1039 101A 1039 101F 1031 102C 1004 1039 200C	top knot

Advanced Issues

So far we have covered what is explained in the Unicode Standard[8]. 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 က`) while others can be quite long (8.5em or 13 times the width of `U+1000 က`). 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 base consonant so long as the consonant:

- is not devowelised with a visible virama and
- has no stacked consonant below it (ignoring true medials: –y –r –w –h) 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 occuring between the two characters in a syllable chain. For example these phrases show possible inter-syllable line breaks.

ကောင်လေးတွေကျောင်း သို့သွားကြသည်။	1000 1031 102C 1004 1039 200C | 101C 1031 1038 | 1010 1039 101D 1031 | 1000 1039 101A 1031 102C 1004 1039 200C 1038 | 101E 102F 102D 1037 | 101E 1039 101D 102C 1038 | 1000 1039 101B | 101E 100A 1039 200C 104B	the kids are going to school
အိပ်ခန်း**တံခါး**ကို	1021 102D 1015 1039 200C | 1001 1014 1039 200C 1038 | **1010 1036** | **1001 102C 1038** | 1000 102F 102D	to the bedroom door

Notice how in the second example the word `1010 1036 | 1001 102C 1038` is a single word with multiple syllables. Is there some way without a dictionary, that the we can ensure that the word is not line broken? There is a Unicode character that was added for version 4.0: `U+2060 WORD JOINER`. Previous to this the character `U+FEFF ZERO WIDTH NON-BREAKING SPACE` was used. Since `U+FEFF` is most commonly used at the start of a Unicode text file to both identify it as being Unicode data and to indicate the encoding form of the data, `U+2060` was added to the standard to take over the function of zero width non-breaking space. The role of this character is to indicate a non-breaking point in a text. Lines should not be broken at that point.

[8] Version 4.0, 2003

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:

အိပ်ခန်း**တံခါး**ကို	1021 102D 1015 1039 200C | 1001 1014 1039 200C 1038 | 1010 1036 **2060** 1001 102C 1038 | 1000 102F 102D	to the bedroom 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

Consider the word အင်္ဝေ. How can it be represented? The normal encoding we would expect would be `1021 1004 1039 101D 1031`. But there are two ways of interpreting this string:

အင်္ဝေ	1021 (1004 1039) 101D 1031
အင္ဝေ	1021 1004 (1039 101D) 1031

The question is how different systems will interpret the string. One approach is to say that kinzi above a consonant is rare and that kinzi above one of the for *medial* consonants (`U+101A U+101B U+101D U+101F`) is very rare, and so we can say that for the sequence `U+1004 U+1039 U+10xx` we interpret as `(U+1004 U+1039) U+10xx` if `U+10xx` is a normal non medial consonant and that we interpret the sequence as `U+1004 (U+1039 U+10xx)` if `U+10xx` is one of the medial consonants.

But `1004 1039` represents a kinzi:

အင်္ဝေ	1021 1004 1039 101D 1031
အင်‌ဝေ	1021 1004 1039 **200C** 101D 1031

But neither of these representations are the word we want, so how can we represent this word? The main issue is whether we interpret `1021 1004 1039 101D 1031` as `1021 (1004 1039) 101D 1031` or as `1021 1004 (1039 101D) 1031`?

There are different approaches we can take, the approach we propose here is to introduce either a `U+200C ZERO WIDTH NON-JOINER` or a `U+200D ZERO WIDTH JOINER` between the `U+1004` and the following `U+1039`. The effect of this is to break the kinzi string and to make the `U+101D` look more like a medial than a main consonant. The sub-sequence `1004 200C 1039 101D` results in the rendering we want. But it also marks that there is a syllable break between the `U+1004` and the `U+101D`, while we want the syllable break to occur before the `U+1004`. So a better solution is: `1004 200D 1039 101D`. Here the `ZERO WIDTH JOINER` indicates that the syllable should be held together and therefore that there is a break before the syllable. Notice that all this talk of syllable breaks makes no difference for rendering (even line breaking). It is only needed for sorting purposes.

If we consider all the sequences that can occur, we can see what the rendering will be and where the syllable break will occur when sorting.

အင်္ဝေ့	1021 1004 1039 101D 1031 1037	(good)
အင္ဝေ့	1021 1004 **200C** | 1039 101D 1031 1037	(bad)
အင္ဝေ့	1021 | 1004 **200D** 1039 101D 1031 1037	(good)
အင်ဝေ့	1021 1004 1039 **200C** | 101D 1031 1037	(good)

အင္‍ဝေ့	`1021	1004 1039` **`200D`** `101D 1031 1037`	(bad)

The final column lists whether the syllable breaking is appropriate. Notice that there is no sequence that explicitly marks a kinzi (only the final sequence does this), and that also is appropriate for indicating that the kinzi is associated with the previous syllable. Therefore we have to make the unmarked initial sequence always represent a kinzi (with the syllable break following the kinzi).

The principles presented here form the basis for the solutions for other encoding problems in Myanmar.

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.

ယောက္‌္ယား	`101A 1031 102C 1000` **`1039 200C`** `1039 101A 102C 1038`	man, husband
က္ယ္ဝန္‌ုပ္‌	`1000 1039 101A 1039 101D 1014` **`1039 200C`** `102F 1015 1039 200C`	I (1st person singular)

Notice how the visible virama (`1039 200C`) 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.

ကလ္ယဏ	`1000 101C 1039 101A 100F`	womanly virtues
အဝ္ဟန္‌	`1021 101D 1039 101F 1014 1039 200C`	momentum

Old Myanmar

There are a few issues that storing old Myanmar text introduce.

Stacked Ya

There are occasions where a medial ya (`U+101A`) representation is used for a stacking ya. What is needed is a syllable break between the base consonant and the ya. Thus we propose:

ဥယ‌္ယာန	`1025 101A` **`200C`** `1039 101A 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. We propose this change of sorting be marked using `U+200D ZERO WIDTH JOINER` between the base consonant and the medial:

က‍္လ	`1000` **`200D`** `1039 101C`	ကျ	drop
က္ယ‍္လပ္‌	`1000 1039 101A` **`200D`** `1039 101C 1015 1039 200C`	ကျပ်	tight

In the second example, while the `U+200D` is not necessary (since from context it is possible to identify that the `U+101C` is a medial), it is included for consistency, making implementation easier.

AMyint (Medial A)

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.

နိယ္အ္‌	1014 102D 101A **1039 1021** 1039 200C	နေ့	day

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`.

`U+200C ZERO WIDTH NON-JOINER` and `U+200D ZERO WIDTH JOINER` are more problematic. At the strictest level of comparison, they should be included but since in many cases they are only used to control syllable breaking for sorting a helpful approach to searching and comparison would only consider these character when they affect rendering. `U+200C` affects rendering only when it follows `U+1039 MYANMAR SIGN VIRAMA`. `U+200D` only affects rendering when it occurs in the sequence `U+1004 U+200D U+1039`.Conclusion

With the change to the Myanmar encoding model comes a much greater simplicity while not changing the original character of the model which is both linguistic and practical. The model may come as a surprise to those who are used to a glyph based encoding in which each glyph shape and position receives its own code, or more radically each cluster receives its own code.

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.

Appendix

This is a subset of the complete chart of the Myanmar block in the Unicode standard. The other letters are not used.

	100	101	102	103	104
0	က	တ	ဠ	◌ူ	၀
1	ခ	ထ	အ	◌ေ	၁
2	ဂ	ဒ		◌ဲ	၂
3	ဃ	ဓ	ဣ		၃
4	င	န	ဤ		၄
5	စ	ပ	ဥ		၅
6	ဆ	ဖ	ဦ	◌ံ	၆
7	ဇ	ဗ	ဧ	◌့	၇
8	ဈ	ဘ		◌း	၈
9	ဉ	မ	ဩ	◌္	၉
A	ည	ယ	ဪ	◌်	၊
B	ဋ	ရ	◌ါ	◌ျ	။
C	ဌ	လ	◌ာ	◌ြ	၌
D	ဍ	ဝ	◌ိ	◌ွ	၍
E	ဎ	သ	◌ီ	◌ှ	၎
F	ဏ	ဟ	◌ု	ဿ	၏