Representing Myanmar in Unicode
Details and Examples Version 3

Martin Hosken[1]

Table of Contents

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

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.

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, within a writing system. We do this by specifying a valid string as being a sequence of slots. Each slot may be empty or contain a character (or sequence as specified by the slot). 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. It should also be born in mind that while the unity of the script as a whole may well be affected by the addition or changes in a single language, each language stands alone in its encoding and needs its own consistency. Care is taken that any changes that a difference in language may cause on the script as a whole (adding more legal sequences), do not cause any changes in other language encodings. This may result in some decisions made for a particular language, looking different from those for another language and the temptation to try to over unify languages should be avoided.

Following the one time change for Burmese in Unicode 5.1, there will be no more changes to Unicode for Burmese. The extra characters described here are additions for minority and historic languages. This version of UTN#11 brings the specification in line with Unicode 5.2.

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 implementers 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.

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.

Introduction to version 3

The first two editions of this document were almost exclusively concerned with the needs of the Burmese language. This edition drastically extends the set of allowable sequences and considers the needs of a number of minority languages. It also adds summary descriptions of a number of languages that have Myanmar based writing systems and gives indication on how they are encoded along with other computational issues that these writing systems raise.

Unicode 5.1 Model

Basic Myanmar

The basic consonants and vowels are relatively obvious in how they are encoded, by examining the character charts. Thus:

စာ	1005 102C	letter

Above 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, ငြ, ဒ, ဓ[2], ဝ, or ဝှ), 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 pronounced, 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 pronounced (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 implementers 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

[2] Some characters may take tall or short forms of -aa based on stylistic preference.

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
ခြင်း	1000 **103C** 1031 1038	grime
မွေး	1019 **103D** 1031 1038	give birth
မှ	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 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 ည် is encoded with U+1009 followed by U+103A even though the glyph shape closely resembles the independent vowel ဉ U+1025 MYANMAR LETTER U. Keyboard implementers 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 vowel 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 င 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
သင်္ဘော်	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, independent vowel or digit. Since all these diacritics are not spacing, how do we know in which order they should be stored? For example, ဝိုဲ can be stored as U+1004 U+102D U+102F or as U+1004 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. The specification of each slot is a sequence of characters. Where there is a list of characters enclosed in [], only one of them may occur in that position. x .. y implies an inclusive range of characters.

Name	Specification	Example	Constraints
Kinzi	[U+1004, U+101B, U+105A] U+103A U+1039	◌	
Consonant	[U+1000 .. U+102A, U+103F, U+1041 .. U+1049, U+104E, U+105A .. U+105D, U+1061, U+1065, U+1066, U+106E .. U+1070, U+1075 .. U+1081, U+108E, U+AA60 .. U+AA6F, U+AA71 .. U+AA76]	◌	required
Stacked	U+1039 [U+1000 .. U+1019, U+101C, U+101E, U+1020, U+1021, U+105A .. U+105D]	◌	
Stacked2	U+1039 [U+1000 .. U+1019, U+101C, U+101E, U+1020, U+1021, U+105A .. U+105D]	◌	
Asat	U+103A	◌	_ [^U+103E, U+1082, U+1037]
Medial Y	[U+103B, U+105E, U+105F]	◌	
Medial R	U+103C	◌	
Medial W	[U+103D, U+1082]	◌	
Medial H	[U+103E, U+1060]	◌	
Mon Asat	U+103A	◌	[U+103E, U+1082] _ [^U+1037]
E vowel	[U+1031, U+1084]	◌	
Shan E vowel	U+1031	◌	U+1031 _
Upper Vowel	[U+102D, U+102E, U+1032 .. U+1036, U+1071 .. U+1074, U+1085, U+109D]	◌	[U+1032, U+1036] [^U+102B, U+102F, U+1030]
Lower Vowel	[U+102F, U+1030]	◌	
Karen Vowel	[U+1062, U+1037]	◌	[^U+102F, U+1030] _
Shan Vowel	U+1086	◌	
A Vowel	[U+102B, U+102C, U+1062, U+1063, U+1067, U+1068, U+1083]	◌	
Anusvara	[U+1036, U+1032]	◌	[U+102D, *Lower Vowel, A Vowel*] _
Pwo Tone	[U+1064, U+1069 .. U+106D]	◌	
Lower Dot	U+1037	◌	[*Lower Vowel*, U+1086, *A Vowel*, *Anusvara, Pwo Tone*] _
Mon h	U+103E	◌	U+102C _ U+103A
Visible virama	U+103A	◌	[*A Vowel*, *Pwo Tone*, U+103E[3], U+1037] _
Visarga	[U+1038, U+1087 .. U+108D, U+108F, U+109A .. U+109C]	◌	
Reduplication	U+AA70	◌	

[3] Where this follows U+102C as in the Mon h slot.

There is a general order of: initial consonant cluster, vowels, finals, tones.

Each line in the order presents a possible slot into which a code may be placed. Only the Consonant slot is required all others are optional. It doesn't actually matter which particular slot a code goes in so long as the relative order of codes with each other is correct.

In the constraints, if a list `[]` begins with a `^` then it is a negative assertion (anything but the characters listed). This also means that an absence of any character will meet that part of the constraint. The _ in the constraint represents the character in the slot that is being constrained.

For example:

ပသျ့ိုး	1015 101E 103B 103E 1030 1038	Malay
မြ	1019 103C 103D 103E 102C	segmentalize
သျှင်	101E 103B 103E 1031 102C 1004 103A	top knot

Notice 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.

The precise slot a character is interpreted as being in, has no significance beyond specifying a relative storage order of characters in a string. Thus in the absence of intervening characters, it may be ambiguous as to which slot a particular character should be put. This is not a problem since with the intervening slots empty, it makes no difference which slot the character is in.

Kinzi

What makes a kinzi a kinzi is that it is actually part of the previous syllable. Thus it is stored first. Notice that there are a number of characters that can be used in this way. In Sanskrit there is a repha which is an r character (`U+101B`) rendered above the consonant: ် `U+101B U+103A U+1039`. In Mon, the Mon nga `U+105A` is used instead of the Burmese nga `U+1004`. These two nga characters render identically in this context, but for consistency within the Mon language, `U+105A` is always used and `U+1004` is never used.

Consonant

The consonant position slot may hold much more than just a pure consonant. Digits, in certain circumstances, may take diacritics and so are included in the list. Although, digit ○ `U+1040` is not included in the set since any diacritics attached to it would assume to be attached to a ဝ `U+101D`.

Stacked

Not every consonant can be stacked, and while theoretically any consonant can take a subjoined form, not all implementations will necessarily need to support all subjoined forms. The ones listed here are the ones known to exist.

Stacked2

In some languages, particularly Sanskrit in rare situations, complex conjuncts involving three non medial consonants may occur. For example: tsna in Sanskrit would be encoded `U+1010 U+1039 U+101E U+1039 U+1014`.

Asat

This slot position is used for all cases where an asat is rendered over a consonant, unless that consonant is followed by a `MEDIAL HA U+103E` which is used as a contraction in Mon, or a `SHAN MEDIAL WA U+1082` which is used with asat to form a vowel in Shan. See the section on Mon for more details.

When data is normalized, if `U+1037` directly follows `U+1039` or `U+103A` then it is reordered before it. So an additional constraint is that `U+103A` may not occur immediately before `U+1037`.

Medial Y

This slot also includes any specific medials that do not correspond to other medials. In this case we include the Mon -m and -n medials.

Medial R

Notice that medial ra is stored after the consonant even though it may be considered to be rendered before it.

Medial W

This slot also includes the Shan medial w. This results in an encoding of ◌ၜ as U+1082 U+103A with the asat being in the Mon Asat slot.

Medial H

This slot includes MON MEDIAL LA U+1060 since it is used as a medial h in Karen.

Mon Asat

This only occurs in Mon, where it immediately follows a MEDIAL HA U+103E or a SHAN MEDIAL WA U+1082.

E Vowel

All pre-vowels go into this slot.

Shan E Vowel

The Common Shan script encodes the vowel that is more usually encoded as ◌ဲ U+1084 using two E vowels, following Thai. Since this only occurs in a historic script, the easiest solution is to allow two E vowels as in ◌ေ◌ U+1031 U+1031. This slot is for the second U+1031.

Upper Vowel

This slot contains anything that can go on top of a consonant. Notice that only one upper vowel can occur. U+1036 may only occur in this slot if there is nothing in the Lower Vowel slot or there is a following spacing component and the anusvara is to be rendered over the consonant. There is one exception to this and that is in Mon where the sequence U+102B U+1036 is rendered ◌ံ. The constraint listed in the chart only applies to the characters U+1032 and U+1036, hence there is no _ placeholder.

Lower Vowel

These are the standard Burmese lower vowels. This also specifies the order of ◌ု as being U+102D U+102F.

Karen Vowel

This slot does not occur with the previous Lower Vowel slot. It contains characters that are used as vowels in other languages. Notice that in Sgaw Karen one can have two occurrences of ◌ါ U+1062 as in ကါ် U+1000 U+1062 U+1062 U+103A.

Shan Vowel

This upper diacritic may either occur above a consonant, or above a following Shan a vowel ◌ါ U+1062. The position depends on which of the various Shan scripts is being written. As a result, this slot position is optimal since it occurs between two slots containing U+1062.

A Vowel

Unlike other slots which may or may not include spacing characters, the A vowel slot always contains a spacing character. This is not to say that the A Vowel slot always has to be filled.

Anusvara

In Mon ◌ံ U+1032 acts as a final character and so may occur over a ◌ ◌ U+102C. In the situation where it occurs after a ◌ U+102D, it is still rendered as a visual ligature with the U+1032 occurring first as in: ◌. Different languages use ◌ U+1036 in different ways. ◌ U+1036 here is acting as a final character, in contrast to the same character in the Upper Vowel slot where it is acting as a vowel.

There is one language in which this approach may result in a possible invisible ambiguity and that is Mon. Mon treats ansuvara ◌ U+1036 as a final nasal and as such it may follow a ◌ U+102C. 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 ◌ U+1036 U+102B and U+102B U+1036, according to the above table, rendering the same, hence the constraint on the Upper Vowel. Likewise for ◌ U+102F U+1032. The visually identical sequence U+1032 followed by a Lower Vowel (U+102F or U+1032) is illegal. For more details see the section on Mon.

Pwo Tone

These are all spacing and may take ◌ U+1037.

Lower Dot

This lower dot slot position may only be filled when either of the A Vowel or Pwo Tone, spacing slots are filled. It is possible for two ◌ U+1037 to occur. For example, in Pwo Karen: ◌ U+1000 U+1060 U+1037 U+106B U+1037. In addition, lower dot may occur after a lower vowel, since lower dot cannot occur in the Karen vowel slot in that context.

Mon h

Mon has the concept of contracting final consonants using diacritics. One such is using medial h followed by an asat to represent a final h. Since the medial h may occur under a U+102C it is listed here before the visible virama which will also occur. This slot is only filled if there is a U+102C and a following visible virama.

Visible Virama

This is only used if there is a spacing character after the consonant on which the asat is rendered (I.e. something in any of the A Vowel or Pwo Tone slots), or immediately following U+1037.

Visarga

The visarga slot not only includes visarga U+1038 but also Shan tone letters.

Reduplication

The reduplication character is found in Khamti Shan. In addition it may ligate with some other characters, but regardless of this ligation, it occurs at the end of the sequence.

Symbols

The following characters classes do not take part in the diacritic order.

Symbol	[U+104A, U+104B, U+104C, U+104D, U+104F, U+109E, U+109F]	◌
Digits	[U+1090 .. U+1099]	◌

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 set[4] 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.

There are other characters that might be expected to be canonically equivalent to sequences, but that are not. In the following, the two cells in a row are not canonically equivalent. Therefore, users should only use the left hand character (except where the right hand side looks different and you need that particular sequence).

သ U+103F	≠	သ U+101E U+1039 U+101E
ဩ U+1029	≠	ဩ U+101E U+103C
ဧ U+102A	≠	ဧ U+101E U+103C U+1031 U+102C U+103A
ဧ U+102A	≠	ဧ U+1029 U+1031 U+102C U+103A

Use of U+104E

One of the significant changes between Unicode 4 and Unicode 5.1 was the change in spelling of lagaun ၎င်း changed from U+104E to U+104E U+1004 U+103A U+1038. This is to facilitate an alternative spelling of lagaun of ၎ U+1004 U+103A U+1039 U+104E. This change results in a subtle change of behaviour for U+104E ၎ from being a complete punctuation symbol with corresponding predefined line breaking behaviour, to being just another character that needs algorithmic analysis both for segmentation and for sorting.

Fractions

A number of legacy fonts have special glyphs for particular fractions. Rather than encoding these with special codes, they can be marked using the U+2044 FRACTIONAL SLASH which is used to build fractions.

Keyboarding

As yet there are few standard keyboard layouts. What can clearly be seen with the complexity of the diacritic order is that expecting a user to type in the correct order is unreasonable. In designing a keyboard, therefore, it is unwise to attempt to produce a simple, non-reordering layout. Care needs to be taken to ensure that the diacritic order is correct, particularly given that users are liable to type in a different order.

[4] 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. In addition, there is no requirement that normalized order must mirror linguistic order.

Advanced Issues

So far we have covered what is explained in the Unicode Standard[5]. In this section we examine some of the more difficult areas of the Myanmar language including some implementation details regarding line breaking, sorting and rendering; further examination of the kinzi question; contractions and some issues with respect to Old Myanmar.

Line breaking

Burmese does not have inter-word 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 line breaking is required within a phrase then there are a number of possible approaches. What is presented here is a sliding scale of quality of line breaking approaches, starting with the simplest.

Insert Zero-Width Spaces

The simplest approach is to insert a U+200B ZERO WIDTH SPACE (ZWSP) between words in a phrase. This would allow line breaks between words in a phrase. The problem is, though, ensuring that ZWSP characters are only inserted between words. The standard approach is to insert ZWSP between syllables, since most words in the languages using the script are monosyllabic. But the problems occur when ZWSP is erroneously inserted into the middle of a polysyllabic word. Such insertions cause problems for searching and indexing. Thus ZWSP should only be used where there is certainty that there is a word break.

Automated Syllable Breaking

A better approach uses a purely algorithmic approach to line breaking based on syllable breaks. The outline algorithm described here should work for any of the languages using the Myanmar script. A syllable break may occur before any cluster (as described in the diacritic ordering section) so long as the kinzi, asat and stacked slots remain empty in the cluster following the possible break point[6]. In reality such an algorithm requires refinement, but it still requires no dictionary. For example, sequences of digits should be kept together and visible virama needs more complex analysis.

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 ကျောင်းကိုသွား**ကြ** တယ်။	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 : 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 polysyllabic 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:

[5] Version 5.1

[6] This is made more complicated when U+1037 is normalized before U+103A, but a syllable break should still not be allowed.

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

The problem with inserting Word Joiners is that it makes searching for polysyllabic words much harder since the searching engine must be able to recognise the Word Joiner characters and ignore them. This is unlikely to happen. Therefore it is advisable not to use Word Joiner characters if at all possible.

Dictionary Based Line Breaking

The next level of sophistication builds on the previous by adding the ability for the line breaker to identify polysyllabic words. Such words are usually held in a dictionary. Thankfully, such a dictionary only need contain polysyllabic words which are far fewer than a complete word list for a language. The main weakness of this approach is where new words are used that are not in the dictionary. For this, one may need to fall back to ZWSP or WJ approaches. The complexity of this approach is that users are not generally aware of the contents of such dictionaries and so cannot predict when they will have difficulties and when not.

Notice that at each level of sophistication, it is necessary for the line breaking approach to be able to handle data that has been generated for a less sophisticated line breaking approach and to handle that appropriately. For example, if a text contains ZWSP characters, they should be honoured.

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.

The basic principle used in sorting most Myanmar based languages, in the script, is to treat a syllable as consisting of one or more of the following components in order:

***C**onsonant **M**edials **V**owels **F**inals **T**one*

There are two primary approaches to sorting. The thinbongyi approach is the current national standard and reorders the components so that the Finals occur before the Vowel:

***C**onsonant **M**edials **F**inals **V**owels **T**one*

The Pali sort uses a different reordering:

***C**onsonant **M**edials **V**owels **T**one **F**inals*

Then sorting proceeds simply, taking each component as having a primary sort relationship to the other components. It should be noted that where there is more than one medial character, they may interact to produce a single sort key. This is also true for sequences of vowels.

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.

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

This storage approach also affects syllable breaking 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
ဂိမ္ဟာန်	1002 102D **1019** 103E 102C 1014 103A	summer

Contextual Shaping

There are a number of situations in which characters change shape to accommodate diacritics and to avoid glyphs clashing.

နှ + lower diacritic or medial ra → နှ

ည + lower diacritic → ည

ရ + lower vowel or medial other than medial h → ရ (short tail)

ရ + medial h → ရ (long tail with no hook)

ြ changes width according to the base character being wrapped. It also truncates its top arm if an upper diacritic would clash with it.

ြ + ွ → ြွ

ု and ူ if they would clash with anything under the base character or a tail → ု and ူ

ဉ + medial or asat → ဉ. Which means that you never use U+1025 for U+1009

Languages

This section gives summary descriptions of a number of writing systems that are based on the Myanmar Unicode block. Each description consists of:

A language tag identifying the particular writing system

- Summary of characters in the alphabet, given in alphabetical order
- Unicode encoding for all characters
- Rendering information including standard ligatures and shaping

Where information is omitted about a particular feature of a writing system, it is assumed that the writing system follows Burmese in that respect.

Since there are no standardised keyboard layouts, none are included.

Burmese

The Burmese language is the primary language that uses the Myanmar script. All other languages are described in terms of it. So where another language does not describe something, it should be assumed to be the same as the Burmese language in that respect.

Language Tag

my

Alphabet

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	စျ	ဉ	ည	ဋ	ဌ	ဍ	ဎ	ဏ
1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	100A	100B	100C	100D	100E	100F

တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ
1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	101A	101B	101C	101D	101E	101F

ဠ	အ
1020	1021

Independent Vowels

These sort as if they are အ followed by the corresponding dependent vowel.

ဣ	ဤ	ဥ	ဦ	ဧ	ဩ	ဪ
1023	1024	1025	1026	1027	1029	102A
အိ	အီ	အု	အူ	အေ	အော	အော်
1021 102D	1021 102E	1021 102F	1021 1030	1021 1031	1021 1031 102C	1021 1031 102C 103A

Medials

ျ	ြ	ွ	ှ
103B	103C	103D	103E

In addition to the basic medials, the following is the relative sort order for medial sequences:

ျွ	ြွ	ျှ	ြှ	ွှ	ျွှ	ြွှ
103B 103D	103C 103D	103B 103E	103C 103E	103D 103E	103B 103D 103E	103C 103D 103E

Dependent Vowels

ာ, ါ	ိ	ီ	ု	ူ	ေ	ဲ	ော	ို	ော်
102C, 102B	102D	102E	102F	1030	1031	1032	1031 102C	102D 102F	1031 102C 103A

ော်
1031 102B 103A

The relative sort order for ော is ော, ော့, ော်

Tones

့	း
1037	1038

Final Consonants

Final consonants are those that are marked as having their inherent vowel killed. That is they are consonants that are either followed by a U+103A MYANMAR SIGN ASAT ် or they are in a stacking relationship with a following subjoined full consonant, in which case they are followed by U+1039 MYANMAR LETTER VIRAMA. The kinzi character U+1004 MYANMAR LETTER NGA U+103A MYANMAR LETTER ASAT U+1039 MYANMAR LETTER VIRAMA ်် is stored before the base character it occurs over and is treated as a final consonant of the previous syllable to that base character.

Note that the final ဉ် is encoded U+1009 U+1039 and not using U+1025 MYANMAR LETTER U.

Symbols

The charts show various symbols, how they are encoded and their corresponding sort equivalent sequences.

သ	ဌ	၍	၎င်း	၏	ံ
103F	104C	104D	104E 1004 103A 1038	104F	1036
သ္သ	န္နက်	ရွေ	လည်းကောင်း	အို	မ်
101E 1039 101E	1014 103E 102D 102F 1000 103A	101B 103D 1031	101C 100A 103A 1038 1000 1031 102C 1004 103A 1038	1021 102D	1019 103A

Note that U+1036 only acts as a final consonant for sorting purposes in combination with another vowel: ိံ U+102D U+1036 or ုံ U+102F U+1036.

Sequences

There are a few words involving contractions which ideally sort differently from how they are spelled. A complete list is not included here and processes may sort such words using default character sorting as though they were not special.

ကောက်ျ	နှင်ပ်	လက်ျ	သို့	ထွင်း	လွှက်
1031 102C 1000 103A 103B	1014 103A 102F 1015 103A	101C 1000 103A 103B	101E 1039 1019 102E	1011 1039 1019 1004 103A 1038	101C 1039 1018 1000 103A
ကောက်ကျ	နှံနှုပ်	လက်ယာ	သမီ	ထမင်း	လက်ဘက်
1031 102C 1000 103A 1000 103B	1014 103A 1014 102F 1015 103A	101C 1000 103A 101A 102C	101E 1019 102E	1011 1019 1004 103A 1038	101C 1000 103A 1018 1000 103A

Punctuation

၊	။
104A	104B

Rendering

Subjoined Consonants

Not all consonants have a corresponding subjoined form. In some cases the corresponding medial character is used since a subjoined consonant indicates a new syllable.

◌္က	◌္ခ	◌္ဂ	◌္ဃ	◌္င	◌္စ	◌္ဆ	◌္ဇ	◌္ဈ	◌္ဉ	◌္ဋ	◌္ဌ	◌္ဍ	◌္ဎ	◌္ဏ	◌္တ	◌္ထ
1039 1000	1039 1001	1039 1002	1039 1003	1039 1004	1039 1005	1039 1006	1039 1007	1039 1008	1039 100A	1039 100C	1039 100D	1039 100E	1039 100F	1039 1010	1039 1011	1039 1012

◌္ဒ	◌္ဓ	◌္ပ	◌္ဖ	◌္ဗ	◌္ဘ	◌္မ	◌္ရ	◌္လ	◌္သ	◌္ဟ
1039 1013	1039 1014	1039 1015	1039 1016	1039 1017	1039 1018	1039 1019	1039 101B	1039 101C	1039 101E	1039 1021

Ligatures

Burmese uses a number of standard ligatures.

ဦ	ဿ	ဏ္ဍ	ဏ္ဍာ	◌ါ	၎
100D 1039 100E	103F	100F 1039 100B	100F 1039 100D	102B 103A	1020 1039 1020

ဥ္	◌ံ
100B 1039 100C	1004 103A 1039

Variants

Alternate forms of some characters exist:

ဥ
100B

Old Burmese

Language Tag

obr-Mymr

Alphabet

The Old Burmese alphabet is identical to that of Burmese for the most part. The only difference occurs in some ligatures and what characters can be subjoined.

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ဉ	ည	ဋ	ဌ	ဍ	ဎ	ဏ
1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	100A	100B	100C	100D	100E	100F

တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ
1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	101A	101B	101C	101D	101E	101F

ဠ	အ
1020	1021

Independent Vowels

These sort as if they are အ followed by the corresponding dependent vowel.

ဣ	ဤ	ဥ	ဦ	ဧ	ဩ	ဪ	ဿ
1022	1023	1024	1025	1026	1027	1028	1029	102A

Medials

ျ	ြ	ွ	ှ
103B	103C	103D	103E

Notice that ြျ U+103B U+103C exists as a sequence, as does ြွ U+103B U+103C U+103D.

Dependent Vowels

ာ	ိ	ီ	ု	ူ	ေ	ဲ	ော	ို
102C	102D	102E	102F	1030	1031	1032	1031 102C	102D 102F

Tones

့	း
1037	1038

Rendering

Old Burmese has a few unique ligatures:

၌	၍	◌ၞ	◌ၟ
1051 1039 100C	1051 1039 100D	101B 103A 1039	1039 100B

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.

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

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. U+2060 WORD JOINER cannot be used since it is functionally identical to U+FEFF ZERO WIDTH NON-BREAKING SPACE and so acts as a space character. This would cause a rendering problem with the following diacritic.

Sanskrit/Pali

The writing system described here is used for both Sanskrit and Pali.

Language Tag

san-Mymr, pli-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ဉ	ဋ	ဌ	ဍ	ဎ	ဏ
1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	100B	100C	100D	100E	100F

တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	ဓ	ဓ
1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	101A	101B	101C	101D	1050	1051

သ	ဟ	အံ	အး
101E	101F	1021 1036	1021 1038

Independent Vowels

အ	အာ	ဣ	ဤ	ဥ	ဦ	ဧ	ဪ	ဩ	ဪ	ဪ	အဲ	ဩ	ဪ
1021	1021 102C	1023	1024	1025	1026	1052	1053	1054	1055	1027	1021 1032	1029	102A

Dependent Vowels

◌ာ	◌ိ	◌ီ	◌ု	◌ူ	◌ၢ	◌ၣ	◌ၤ	◌ၥ	◌ေ	◌ဲ	◌ော	◌ော်
102C	102D	102E	102F	1030	1056	1057	1058	1059	1031	1032	1031 102C	1031 102C 103A

◌ံ	◌း
1036	1038

Conjuncts

Rather than a true medial mechanism, Sanskrit follows the Indic script tradition with the use of conjuncts. Here we list some of them, showing some of the complexities of rendering, but also showing how such conjuncts fit the encoding model naturally.

ကြွေ	က္ခ	ကြွေ	စ္စ	တြံ
1000 1039 1010 103C 102D	1000 1039 1010 103D	1004 103A 1039 1000 1039 1010 102D	1005 1039 1005 1032	101B 103A 1039 1010 1039 1010

ဟြေ	ဂ္ဓဲ	ဈ္ဈ
101B 103A 1039 101F 103C 102E	1000 1039 1051 1032	101B 103A 1039 1050 102D

Finals

In addition to normal final consonants, Sanskrit has final conjuncts including those that are made up of kinzi or rapha.

မ္ဗ်	န္တ်ြ	င်္ဂ်	ရ်္ပ်ျ
1019 1039 1017 103A	1014 1039 1010 103A 103C	1004 103A 1039 1002 103A	101B 103A 1039 1015 103A 103B

Mon

Language Tag

mnw-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	ငꩻ	စ	ဆ	ဇ	ဈ	ဉ	ည	ဋ	ဌ	ဍ	ဎ	ဏ
1000	1001	1002	1003	105A	1005	1006	1007	105B	1009	100A	100B	100C	100D	100E	100F

တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ
1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	101A	101B	101C	101D	101E	101F

ဠ	အ	ဓ	ꩻ
1020	1021	105C	105D

The nga letter in Mon is encoded U+105A ငꩻ and not U+1004 ဂ 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: ◌ꩻ (U+1039 U+105A).

Mon has a character 'great nya' which is encoded ညြ U+100A U+1039 U+100A. But this is stylistic and the same sequence may also be rendered ဩ U+100A U+1039 U+100A.

Medials

Mon has a number of medial forms even where the characters are not linguistic medials. The specific forms in Mon are:

◌ꩻ	◌ꩾ	◌ꩿ	◌ျ	◌ြ	◌ွ	◌ှ	◌ꪾ
1039 105A	105E	105F	103B	103C	1060	103D	103E
ငꩻ	ဏ	မ	ယ	ရ	လ	ဝ	ဟ

Independent Vowels

အ	အာ	ဣ	ဣီ	ဥ	ဥꩻ	ဩ	ဪ	ဪာ
1021	1021 102C	1023	1023 1033	1025	1025 102F	1028	1029	102A

Dependent Vowels

◌ာ	◌ိ	◌ီ	◌ု	◌ူ	◌ေ	◌ဲ	◌ော	◌ံ	◌ိုꩻ	◌ꩻ	◌ိꩻ	◌း
102C	102D	1033	102F	1030	1031	1032	1031 102C	1034	102D 102F	1036	1035 102F	1038

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

ကံ်	U+1000 **U+1036** U+102C U+103A
ကံ	U+1000 U+102C **U+1036**
ဂါ	U+1002 U+102B **U+1036**

The ordering of U+1036 U+102C U+103A follows the default encoding order and keeps consistency across the script.

Contractions

Mon has the concept of final character contractions. One of these is where ဟ် becomes ှ် on the final character of the syllable. Thus one can have ◌ှ်. The natural order for these would be U+102C U+103E U+103A following the order of the characters being contracted. But a contraction may also occur before 102C. Thus the following examples are all possible: စှော် စှော စော့်.

နှ်	1005 **103E 103A**
နှံ်	1005 **103E 103A** 1031
နှော်	1005 **103E** 1031 102C **103A**
နှော်	1005 **103E 103A** 1031 102C
နော့်	1005 1031 102C **103E 103A**

Likewise with the sequence ◌ိုယ် which can contract to ◌ို. For example:

မ်ို	105D 102D 102F **1032**
ဓဲ့	1013 101D 102F **1032**

Rendering

Mon has some extra complex stacking:

◌္ဒ	1039 1010 103D

Sgaw Karen

The Sgaw Karen language is the primary language in the Karen language group. Other languages in the group often base their writing system on this Sgaw Karen writing system. Karen languages have no final consonants. Thus while they may be sorted as any other Myanmar script based language, there is actually no reordering required.

Language Tag

ksw-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ၐ	ဥ	တ	ထ	ဒ	န	ပ	ဖ	ဘ
1000	1001	1002	1003	1004	1005	1006	1061	100A	1010	1011	1012	1014	1015	1016	1018

မ	ယ	ရ	လ	ဝ	သ	ဟ	အ	ဧ
1019	101A	101B	101C	101D	101E	101F	1021	1027

U+1061 looks as if it could be encoded as U+101B ရ U+103E ◌ှ. But since this character occurs as an independent consonant in Sgaw Karen, it has its own code. The two spellings are not equivalent.

Medials

Sgaw Karen medials have different linguistic values and styling to Burmese. The third row gives the base consonant that the medial represents.

◌ှ	◌ွ	◌ျ	◌ြ	◌ွှ
103E	1060	103B	103C	103D
ဂ	ယ	လ	ရ	ဝ

Vowels

◌ါ	◌ံ	◌ာ	◌ု	◌ူ	◌ိ	◌ဲ	◌ီ	◌ဳ
102B	1036	1062	102F	1030	1037	1032	102D	102E

Tones

◌်	◌်ာ	◌း	◌်ါ	◌ာ
1062 103A	102C 103A	1038	1063 103A	1064

Ligatures

The following contractions expand as listed and sort according to their expansion:

ဒ်	မ်
1012 103A	1019 103A
ဒံ	မီာ
1012 1036	1019 102E 1064

Rendering

Sgaw Karen has no subjoined consonants.

One stylistic positioning preference is that U+1037 renders to the left of any lower diacritic. Thus ဥ◌ံ renders as ဥံ

U+103E is styled differently to Burmese in that the main stem is angled and the foot is horizontal ◌ှ.

Some older readers of Sgaw Karen like to always use the full height forms of U+102F and U+1030.

Western Pwo Karen

Pwo Karen is based on Sgaw Karen and has many similar behaviours.

Language Tag

pwo-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ည	၎	တ	ထ	ဒ	န	ပ	ဖ
1000	1001	1002	100E	1004	1005	1006	1007	100A	1061	1010	1011	1012	1014	1015	1016

ဘ	မ	ယ	ရ	လ	ဝ	၀	သ	အ	ဧ	၌
1018	1019	101A	101B	101C	101D	1065	101F	1021	1027	1066

Notice that ၌ has its own code U+1066 and the sequence U+1015 U+103E is not used and constitutes a spelling error.

Medials

◌ွ	◌ျ	◌ြ	◌ွ	◌ှ
1060	103B	103C	103D	103E

Vowels

◌ါ	◌ံ	◌့	◌ဲ	◌ာ	◌ံ	◌ု	◌ူ	◌ိ	◌ီ
102B	1036	1037	1032	1067	1068	102F	1030	102D	102E

Tones

◌်	◌း	◌်	◌်	◌်	◌့်	◌့်	◌း့	◌း
1069	106A	106B	106C	106D	1069 1037	106B 1037	106A 1037	1038

Eastern Pwo Karen

This is known as the monastic script and is based on the Mon script. Tone is not marked.

Language Tag

kjp-Mymr

Alphabet

The sort order for this writing system is unknown.

Consonants

The consonants are listed in the corresponding order to pwo-Mymr.

က	ခ	ဂ	စ	ဆ	ည	တ	ထ	ဒ	န	ကာ	ပ	ဖ	ဓ	မ	ယ
1000	1001	1004	1005	1006	100A	1010	1011	100D	1014	106E	1015	1016	105C	1019	101A

ရ	လ	ဝ	ဟ	အ
101B	101C	101D	101F	1021

Medials

◌ွ	◌ျ	◌ြ	◌ှ	◌ၢ	◌ၣ	◌ၤ
103D	1060	103C	103B	103E	1039 1012	1039 101A
ဝ	လ	ရ	ဃ	ဟ	ဒ	ယ

Vowels

◌ေ	◌း	◌ဲ	◌ာ	◌ိ	◌ီ	◌ု	◌ူ	◌ို	◌ေါ	◌ံ
1031	1038	1032	102C	103A 102F	105C	102D	102E	1030	102D 102F	102F	1036

Finals

The Monastic script follows Mon in supporting some final contractions. Details of how these are encoded is the same as in Mon.

Pa'o Karen

Language Tag

blk-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ည	ဋ	ဌ	ဍ	ဎ	ဏ	တ
1000	1001	1002	1003	1004	1005	1006	1007	1008	100A	100B	100C	100D	100E	100F	1010

ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ	ဠ
1011	1012	1013	1014	1015	1016	1017	1018	1019	101A	101B	101C	101D	101E	101F	1020

အ
1021

Pa'o also has stacking consonants and kinzi as in Burmese.

Medials

◌ျ	◌ွ	◌ှ
103B	103C	103D

Vowels

◌ာ, ◌ါ	◌ိ	◌ီ	◌ု	◌ူ	◌ေ◌	◌ော	◌ံ့	◌ံ
102C, 102B	102D	102E	102F	1030	1031 0137	1031	1032 1037	1032

◌ံ့	◌ံ◌	◌ော့	◌ော်	◌ိ	◌ိ
102F 1032 1004 1037 103A	102F 1032 1004 103A	1031 102C 1037	1031 102C 103A	102D 102F 1037	102D 102F

◌ံ့	◌ံ	◌ိံ	◌ုံ	◌ံ
1036 1037	1036	102F 1036 1037	102F 1036	102F 1032

U+102F U+1032 and U+102F U+1036 have their orderings because this in order to be consistent across writing systems, we need to follow Mon here. Notice that due to normalization, the order of ◌ု 1037 and ◌် 103A following the final c U+1004 is counter intuitive.

Tones

◌့	◌း	◌꩜
AA7B	1038	108F

Kayah

There are no final consonants in Kayah.

Language Tag

kyu-Mymr

Alphabet

Each of the sets of characters are in alphabetic order.

Consonants

က	ခ	ဃ	င	စ	ဆ	ဇ	ည	တ	ထ	ဒ	န	ပ	ဖ	ဗ	ဘ
1000	1001	1003	1004	1005	1006	1007	100A	1010	1011	1012	1014	1015	1016	1017	1018

မ	ယ	ရ	လ	ဝ	သ	ဟ	အ
1019	101A	101B	101C	101D	101E	101F	1021

Medials

Kayah uses the two lower vowel characters (U+102F, U+1030) as medials. Following the script order, these two characters are stored following the vowel (excepting U+1032 and U+1036). While this is linguistically inaccurate, it only causes problems during keying and sorting.

◌ꤰ	◌ꤱ	◌ꤼ	◌ꤻ	◌ꤽ	◌ꤾ
102F	1030	103C	103B	103D	103E

Vowels

The sequence order for the two vowels: U+1032 and U+1036 are that they follow U+102F and U+1030.

◌ꤵ	◌ꤷ	◌ꤶ	◌ꤺ	◌ꤹ	◌ꤸ	◌ꤴ	◌ꤳ
1072	102E	102D	1036	1032	1073	1074	1034

Tones

◌꤮	◌꤯
1064	1038

Asho Chin

Language Tag

csh-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ည	တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ
1000	1001	1002	1004	1005	1006	1007	100A	1010	1011	1012	1013	1014	1015	1016	1017

ဘ	မ	ယ	ရ	ၗ	လ	ဝ	ဟ	အ	ဧ
1018	1019	101A	101B	1061	101C	101D	101F	1021	1027

Medials

◌ျ	◌ွ	◌ှ	◌ြ
103E	1060	103D	103B

Vowels

◌ံ	◌ု	◌ော	◌ေ	◌ာ	◌ါ	◌ိ	◌ီ	◌ို	◌ဲ	◌ဳ
1036	1037	1034	1032	1067	1068	102F	1030	102D	102E	1033

Tones

◌း	◌့	◌္	◌်	◌့်	◌္
1069	106A	106D	106C	1069 1037	106A 1037

Notice that ◌ U+1037 may occur twice. It may occur as a vowel and also as a tone modifier.

Digraphs

◌ိ်	◌ီ်	◌ဃ
102D 102B	102E 102B	1038

Shan

Language Tag

shn-Mymr (modern Shan script)

Alphabet

Consonants

ၵ	ၶ	ၷ	င	ၸ	သ	ၺ	ၻ	တ	ထ	ၼ	ပ	ၽ	ၾ	ၿ	ဝ
1075	1076	1077	1004	1078	101E	107A	1079	1010	1011	107B	107C	1015	107D	107E	107F

မ	ယ	ရ	လ	ဝ	၍	၎	ဢ
1019	101A	101B	101C	101D	1080	1081	1022

The character ၎ 1081 is not represented using the sequence U+1002 U+103E and often takes a different visual form. Likewise ၾ 107E is not represented by U+107D U+103E. U+103E does not occur in Shan.

The consonants are listed in alphabetical order. But typically characters: U+1077, U+1079, U+107B, U+107F, U+1080 are not included in the alphabet when it is taught since they are only used for loan words.

Medials

ျ	ြ	ွ
103B	103C	1082

Vowels

The following are used in open syllables and are listed in alphabetical order.

ာ	ိ	ီ	ေ်	ေ	ဧ	ု	ူ	ုဝ်	ူဝ်
1083	102D	102E	103A 1031	1031	1084	102F	1030	102F 101D 103A	1030 101D 103A

ေါ်	ော	ိုဝ်	ိူဝ်	ံ	်ရ
1031 1083 103A	1031 1083	102D 102F 101D 103A	102D 1030 101D 103A	1086	107A 103A

ါံ	ာ်ရ	ုရ်	ူရ်	ံ	ွရ်
1062 1086	1062 107A 103A	103D 107A 103A	1030 107A 103A	103D 1086	103D 107A 103A

ိုရ်	ိူရ်	ျ	ဝ်	ာဝ်
102D 102F 107A 103A	102D 1030 107A 103A	109F	101D 103A	1062 101D 103A

ိဝ်	ီဝ်	ဳဝ်	ိုဝ်ဝ်
102D 101D 103A	1035 101D 103A	1085 101D 103A	102D 102F 101D 103A 101D 103A

ိုင်ဝ်	်
102D 1030 101D 103A 101D 103A	1082 103A

The sequence U+1082 U+103A uses the visible virama U+103A to mark the medial wa U+1082 as a final.

ၵ်ေ has the encoding U+103A U+1031 which is counterintuitive, but is necessary because of the complexities surrounding such a sequence in other languages. In Burmese there is the issue of contractions and in Mon there is the issue of final contractions. The conclusion is that Shan integrates best by having this order. Note that U+103A comes before the medial. For example, ၵြၢ်ႇ U+1075 U+103A U+103C U+1031.

The following are used in closed syllables:

◌ၢ	◌ိ	◌ဵ	◌ဳ	◌ု	◌ို	◌ူ	◌ိူ	◌ိုဝ်
1062	102D	1035	1085	102F	1030	103D	102D 102F	102D 1030

Notice that different codes are used for the -a vowel when in an open syllable (U+1083) and a closed syllable (U+1062). If a single key is desired for the one vowel then it is up to the keyboard to make the contextual change rather than storing a single code and using rendering to change shape.

Finals

All initial consonants (except U+101D and U+107A) may be used as conventional finals if followed by U+103A. In addition, the following consonants may also take a medial U+103B with U+103A. In such cases they follow their unadorned counterparts in the sort order.

ၵျ်	ၶျ်	ၷျ်	သျ်
1075 103A 103B	1076 103A 103B	1077 103A 103B	101E 103A 103B

Notice that while the linguistic order would imply the sequence U+103B U+103A, in order to resolve the ambiguity in the diacritic sequence order that such an order would introduce, it is necessary to store these sequences as U+103A U+103B. This is not a problem so long as keyboard implementations can handle the typing order being reversed and analytical processes make suitable allowance.

Tones

◌့	◌း	◌ဲ	◌ံ	◌ဴ
1087	1088	1038	1089	108A

Digits

Shan has its own set of digits, although they are rarely used.

၀	၁	၂	၃	၄	၅	၆	၇	၈	၉
1090	1091	1092	1093	1094	1095	1096	1097	1098	1099

Symbols

There are two symbols in Shan that are used for standalone words.

၎	၍
109E	109F

Old Shan

There are a number of old Shan orthographies, but the traditional script adds one extra character: U+1036 as a final -m. In the modern script this has been replaced by a full final ဉ် U+1019 U+103A. The use of the character also introduces one extra sequence:

◌်
103A 1036

which corresponds to ◌ိုဉ် U+102D U+1019 U+103A in the modern script.

Khamti Shan

Support for Khamti Shan is added in Unicode 5.2. The most noticeable feature of the writing system is that many characters have a stylistic dot added to them. This dot does not necessarily make them a different character since the dot is only considered to be stylistic rather than a normative part of the character.

Language Tag

kht-Mymr

Alphabet

Consonants

ကꩻ	ခꩻ	ဂꩻ	ဃꩻ	ငꩻ	စꩻ	ဆꩻ	ဇꩻ	ဈꩻ	ဉꩻ	ညꩻ	ဋꩻ	ဌꩻ	ဍꩻ	ဎꩻ
1000	1075	AA71	1002	1004	AA61	AA62	AA63	AA64	AA65	AA66	AA67	AA68	AA69	107C

တꩻ	ထꩻ	ဒꩻ	ဓꩻ	နꩻ	ပꩻ	ဘꩻ	မꩻ	ယꩻ	ရꩻ	လꩻ	ဝꩻ	သꩻ
1010	1011	107B	AA6A	AA6B	1015	1078	107F	1079	1019	101A	101B	101C	101D	AA6C

၍	ဿ	ာꩻ	ဟꩻ	ဩ
AA6D	AA6E	1022	AA6F	1080

Medials

◌ျ	◌ြ	◌ွ
103B	103C	103D

Vowels

There are no independent vowels in Khamti Shan

◌ꩻ	◌ꩻ	◌ိ	◌ီ	◌ဵ	◌ံ	◌ꩻ	◌ꩻ	◌ꩻ	◌ꩻ	◌ꩻ	◌ꩻ
1062	1083	102D	102E	1085	1032	102F	1030	1031	1084	1082 103A	1036

Tones

◌့	◌့	◌း	◌,	◌;	◌း	◌း
109A	1089	109B	1087	1088	1038	108A

Notice that the unmarked tone is in fact tone 7 and is sorted before tone 8 U+108A.

Digits

Khamti Shan uses the Shan digits.

Logograms

Khamti Shan has 3 characters which can each take tone but that represent complete syllables

ဢꩻ	ဤꩻ	ဥꩻ
AA74	AA75	AA76

Finals

The following consonants may occur with an asat at the end of a syllable. Khamti does not chain syllables.

ꩃ	ꩄ	ꩅ	ꩈ	ꩆ	ꩉ	ꩋ	ꩌ
1000	1004	AA65	1010	AA6B	1015	1019	101D

Reduplication

The reduplication character is functionally similar to the corresponding character in Thai ๆ U+0E46 THAI CHARACTER MAIMAYOK. It is a spacing character.

ꩻ
AA70

The reduplication character ligates with two diacritics. This ligation may also occur across a tone mark.

◌ꩻ◌	◌ꩻ◌
1032 AA70	103A AA70

Historic Khamti Shan

The Khamti Shan script has undergone script development and as such there are some transition characters that are no longer used. But due to the existence of documents using these characters, they are included in Unicode and sort at the end of the list of consonants in this order.

ꩍ	ꩊ	ꩇ
AA60	AA72	AA73

Aiton & Phake

Aiton and Phake are closely related languages with nearly identical orthographies. The differences are purely stylistic. Aiton and Phake have their own font styles that are related to Khamti Shan but different. The style used here is Khamti Shan and not Aiton or Phake.

Language Tag

aio-Mymr, phk-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	ငှ	တ	ထ	ဒ	ပ	ဘ	မ	ယ	ရြ	လ	ဝ
1000	1075	1004	AA61	107A	1010	1011	AA6B	1015	1078	1019	101A	AA7A	101C	101D

ꩻ	ဿ
AA6D	1022

Medials

◌ျ	◌ြ	◌ၞ
103B	103C	105E

Subjoined Consonants

Aiton follows Burmese in using subjoined consonants to chain syllables in a polysyllabic word. The following subjoined characters exist:

◌္က	◌္ဃ	◌္တ	◌္ထ	◌္ပ	◌္ယ	◌္လ
1039 1000	1039 AA60	1039 1010	1039 1011	1039 1015	1039 101A	1039 101C

Vowels

These are final vowels that have no following consonant.

◌ါ	◌ိ	◌ီ	◌ိူ	◌ေ	◌ော	◌ုံ	◌ူံ
1083	109C	102E	1030	1031	1031 1083	102F 101D 103A	102D 102F 101D 103A

These vowels are followed by a final consonant.

◌ိ	◌ု	◌ူ	◌ူ
102D	102F	103D	102D 102F

Diphthongs

◌ံ	◌်ံ	◌ဲ	◌ဲ့	◌ော	◌ိုး	◌ွံ	◌ွၞ
1036	103A 1036	109D	103D 109D	103D 1031	102D 102F 109C	103A 103D	103A 105E

Ligatures

The following ligatures do not take diacritics, but are considered as words.

၎	၏	၍
AA77	AA78	AA79

Rumai Palaung

Language Tag

rbb-Mymr

Alphabet

Consonants

က	ခ	ဂ	ဃ	စ	ဆ	ဇ	ည	တ	ထ	ဒ	န	ပ	ဖ	ဘ	မ
1000	1001	1002	1004	1005	1006	1007	100A	1010	1011	1012	1014	1015	1016	1018	1019

ယ	ရ	လ	ဧ	ဝ	ဟ	အ
101A	101B	101C	108E	101D	101F	1021

Medials

◌ျ	◌ြ	◌ွ	◌ျွ	◌ှ	◌ျှ	◌ွှ
103B	103C	103D	103B 103D	103E	103B 103E	1039 101C

Vowels

◌ာ	◌ိ	◌ီ	◌ု	◌ူ	◌ေ	◌ဲ	◌ေဲ	◌ော	◌ို	◌ံ
102C	102D	102E	102F	1030	1031	1032	1031 1032	1031 102C	102D 102F	102C 103A

Tones

◌း	◌့	◌္	◌်
1038	1089	1088	108F

Charts

1000 **Myanmar** **109F**

	100	101	102	103	104	105	106	107	108	109
0	က	တ	ဠ	◌ူ	၀	ၐ	◌ၠ	ၰ	ႀ	႐
1	ခ	ထ	အ	◌ေ	၁	ၑ	ၡ	◌ၱ	ႁ	႑
2	ဂ	ဒ	ဢ	◌ဲ	၂	ၒ	◌ၢ	◌ၲ	◌ႂ	႒
3	ဃ	ဓ	ဣ	◌ဳ	၃	ၓ	◌ၣ	◌ၳ	◌ႃ	႓
4	င	န	ဤ	◌ဴ	၄	ၔ	◌ၤ	◌ၴ	◌ႄ	႔
5	စ	ပ	ဥ	◌ဵ	၅	ၕ	ၥ	ၵ	◌ႅ	႕
6	ဆ	ဖ	ဦ	◌ံ	၆	ၖ	ၦ	ၶ	◌ႆ	႖
7	ဇ	ဗ	ဧ	◌့	၇	ၗ	◌ၧ	ၷ	◌ႇ	႗
8	ဈ	ဘ	ဨ	◌း	၈	◌ၘ	◌ၨ	ၸ	◌ႈ	႘
9	ဉ	မ	ဩ	◌္	၉	◌ၙ	◌ၩ	ၹ	◌ႉ	႙
A	ည	ယ	ဪ	◌်	၊	ၚ	◌ၪ	ၺ	◌ႊ	◌ႚ
B	ဋ	ရ	◌ါ	◌ျ	။	ၛ	◌ၫ	ၻ	◌ႋ	◌ႛ
C	ဌ	လ	◌ာ	◌ြ	၌	ၜ	◌ၬ	ၼ	◌ႌ	◌ႜ
D	ဍ	ဝ	◌ိ	◌ွ	၍	ၝ	◌ၭ	ၽ	◌ႍ	◌ႝ
E	ဎ	သ	◌ီ	◌ှ	၎	◌ၞ	ၮ	ၾ	ႎ	႞
F	ဏ	ဟ	◌ု	ဿ	၏	◌ၟ	ၯ	ၿ	◌ႏ	႟

Consonants

1000 MYANMAR LETTER KA
1001 MYANMAR LETTER KHA
1002 MYANMAR LETTER GA
1003 MYANMAR LETTER GHA
1004 MYANMAR LETTER NGA
1005 MYANMAR LETTER CA
1006 MYANMAR LETTER CHA
1007 MYANMAR LETTER JA
1008 MYANMAR LETTER JHA
1009 MYANMAR LETTER NYA
100A MYANMAR LETTER NNYA
100B MYANMAR LETTER TTA
100C MYANMAR LETTER TTHA
100D MYANMAR LETTER DDA
100E MYANMAR LETTER DDHA
100F MYANMAR LETTER NNA
1010 MYANMAR LETTER TA
1011 MYANMAR LETTER THA
1012 MYANMAR LETTER DA
1013 MYANMAR LETTER DHA
1014 MYANMAR LETTER NA
1015 MYANMAR LETTER PA
1016 MYANMAR LETTER PHA
1017 MYANMAR LETTER BA
1018 MYANMAR LETTER BHA
1019 MYANMAR LETTER MA
101A MYANMAR LETTER YA
101B MYANMAR LETTER RA
101C MYANMAR LETTER LA
101D MYANMAR LETTER WA
101E MYANMAR LETTER SA
101F MYANMAR LETTER HA
1020 MYANMAR LETTER LLA

Independent vowels

1021 MYANMAR LETTER A
• also represents the glottal stop as a consonant
1022 MYANMAR LETTER SHAN A
1023 MYANMAR LETTER I
1024 MYANMAR LETTER II
1025 MYANMAR LETTER U
1026 MYANMAR LETTER UU
≡ 1025 ◌ 102E ◌ူ
1027 MYANMAR LETTER E
1028 MYANMAR LETTER MON E
1029 MYANMAR LETTER O
102A MYANMAR LETTER AU

Dependent vowel signs

102B MYANMAR VOWEL SIGN TALL AA
102C MYANMAR VOWEL SIGN AA
102D MYANMAR VOWEL SIGN I
102E MYANMAR VOWEL SIGN II
102F MYANMAR VOWEL SIGN U
1030 MYANMAR VOWEL SIGN UU
1031 MYANMAR VOWEL SIGN E
• stands to the left of the consonant
1032 MYANMAR VOWEL SIGN AI
1033 MYANMAR VOWEL SIGN MON II
1034 MYANMAR VOWEL SIGN MON O
1035 MYANMAR VOWEL SIGN E ABOVE

Various signs

1036 MYANMAR SIGN ANUSVARA
1037 MYANMAR SIGN DOT BELOW
= aukmyit
• a tone mark
1038 MYANMAR SIGN VISARGA
1039 MYANMAR SIGN VIRAMA
= killer (when rendered visibly)
103A MYANMAR SIGN ASAT
= killer (always rendered visibly)

Dependent consonant signs

103B MYANMAR CONSONANT SIGN MEDIAL YA
103C MYANMAR CONSONANT SIGN MEDIAL RA
103D MYANMAR CONSONANT SIGN MEDIAL WA
103E MYANMAR CONSONANT SIGN MEDIAL HA

Consonant

103F MYANMAR LETTER GREAT SA

Digits

1040 MYANMAR DIGIT ZERO
1041 MYANMAR DIGIT ONE
1042 MYANMAR DIGIT TWO
1043 MYANMAR DIGIT THREE
1044 MYANMAR DIGIT FOUR
1045 MYANMAR DIGIT FIVE
1046 MYANMAR DIGIT SIX
1047 MYANMAR DIGIT SEVEN
1048 MYANMAR DIGIT EIGHT
1049 MYANMAR DIGIT NINE

Punctuation

104A MYANMAR SIGN LITTLE SECTION
→ (devanagari danda - 0964)
104B MYANMAR SIGN SECTION
→ (devanagari double danda - 0965)

Various signs

104C MYANMAR SYMBOL LOCATIVE
104D MYANMAR SYMBOL COMPLETED
104E MYANMAR SYMBOL AFOREMENTIONED
104F MYANMAR SYMBOL GENITIVE

Pali and Sanskrit extensions

1050 MYANMAR LETTER SHA
1051 MYANMAR LETTER SSA
1052 MYANMAR LETTER VOCALIC R
1053 MYANMAR LETTER VOCALIC RR
1054 MYANMAR LETTER VOCALIC L
1055 MYANMAR LETTER VOCALIC LL
1056 MYANMAR VOWEL SIGN VOCALIC R
1057 MYANMAR VOWEL SIGN VOCALIC RR
1058 MYANMAR VOWEL SIGN VOCALIC L
1059 MYANMAR VOWEL SIGN VOCALIC LL

Extensions for Mon

105A MYANMAR LETTER MON NGA
105B MYANMAR LETTER MON JHA
105C MYANMAR LETTER MON BBA
105D MYANMAR LETTER MON BBE

105E MYANMAR CONSONANT SIGN MON MEDIAL NA
105F MYANMAR CONSONANT SIGN MON MEDIAL MA
1060 MYANMAR CONSONANT SIGN MON MEDIAL LA

Extensions for S'gaw Karen

1061 MYANMAR LETTER SGAW KAREN SHA
1062 MYANMAR VOWEL SIGN SGAW KAREN EU
1063 MYANMAR TONE MARK SGAW KAREN HATHI
1064 MYANMAR TONE MARK SGAW KAREN KE PHO

Extensions for Western Pwo Karen

1065 MYANMAR LETTER WESTERN PWO KAREN THA
1066 MYANMAR LETTER WESTERN PWO KAREN PWA
1067 MYANMAR VOWEL SIGN WESTERN PWO KAREN EU
1068 MYANMAR VOWEL SIGN WESTERN PWO KAREN UE
1069 MYANMAR SIGN WESTERN PWO KAREN TONE-1
106A MYANMAR SIGN WESTERN PWO KAREN TONE-2
106B MYANMAR SIGN WESTERN PWO KAREN TONE-3
106C MYANMAR SIGN WESTERN PWO KAREN TONE-4
106D MYANMAR SIGN WESTERN PWO KAREN TONE-5

Extensions for Eastern Pwo Karen

106E MYANMAR LETTER EASTERN PWO KAREN NNA
106F MYANMAR LETTER EASTERN PWO KAREN YWA
1070 MYANMAR LETTER EASTERN PWO KAREN GHWA

Extension for Geba Karen

1071 MYANMAR VOWEL SIGN GEBA KAREN I

Extensions for Kayah

1072 MYANMAR VOWEL SIGN KAYAH OE
1073 MYANMAR VOWEL SIGN KAYAH U
1074 MYANMAR VOWEL SIGN KAYAH EE

Extensions for Shan

1075 MYANMAR LETTER SHAN KA
1076 MYANMAR LETTER SHAN KHA
1077 MYANMAR LETTER SHAN GA
1078 MYANMAR LETTER SHAN CA
1079 MYANMAR LETTER SHAN ZA
107A MYANMAR LETTER SHAN NYA
107B MYANMAR LETTER SHAN DA
107C MYANMAR LETTER SHAN NA
107D MYANMAR LETTER SHAN PHA
107E MYANMAR LETTER SHAN FA
107F MYANMAR LETTER SHAN BA
1080 MYANMAR LETTER SHAN THA
1081 MYANMAR LETTER SHAN HA
1082 MYANMAR CONSONANT SIGN SHAN MEDIAL WA
1083 MYANMAR VOWEL SIGN SHAN AA
1084 MYANMAR VOWEL SIGN SHAN E
1085 MYANMAR VOWEL SIGN SHAN E ABOVE
1086 MYANMAR VOWEL SIGN SHAN FINAL Y
1087 MYANMAR SIGN SHAN TONE-2
1088 MYANMAR SIGN SHAN TONE-3
1089 MYANMAR SIGN SHAN TONE-5
108A MYANMAR SIGN SHAN TONE-6
108B MYANMAR SIGN SHAN COUNCIL TONE-2
108C MYANMAR SIGN SHAN COUNCIL TONE-3
108D MYANMAR SIGN SHAN COUNCIL EMPHATIC TONE

Extensions for Rumai Palaung

108E MYANMAR LETTER RUMAI PALAUNG FA
108F MYANMAR SIGN RUMAI PALAUNG TONE-5

Shan digits

1090 MYANMAR SHAN DIGIT ZERO
1091 MYANMAR SHAN DIGIT ONE
1092 MYANMAR SHAN DIGIT TWO
1093 MYANMAR SHAN DIGIT THREE
1094 MYANMAR SHAN DIGIT FOUR
1095 MYANMAR SHAN DIGIT FIVE
1096 MYANMAR SHAN DIGIT SIX
1097 MYANMAR SHAN DIGIT SEVEN
1098 MYANMAR SHAN DIGIT EIGHT
1099 MYANMAR SHAN DIGIT NINE

Extensions for Khamti Shan

109A MYANMAR SIGN KHAMTI TONE-1
109B MYANMAR SIGN KHAMTI TONE-3

Extensions for Aiton and Phake

109C MYANMAR VOWEL SIGN AITON A
109D MYANMAR VOWEL SIGN AITON AI

Shan symbols

109E MYANMAR SYMBOL SHAN ONE
109F MYANMAR SYMBOL SHAN EXCLAMATION

	AA6	AA7
0	ꩠ	ꩰ
1	ꩡ	ꩱ
2	ꩢ	ꩲ
3	ꩣ	ꩳ
4	ꩤ	ꩴ
5	ꩥ	ꩵ
6	ꩦ	ꩶ
7	ꩧ	꩷
8	ꩨ	꩸
9	ꩩ	꩹
A	ꩪ	ꩺ
B	ꩫ	◌ꩻ
C	ꩬ	
D	ꩭ	
E	ꩮ	
F	ꩯ	

Khamti Consonants

AA60 MYANMAR LETTER KHAMTI GA
AA61 MYANMAR LETTER KHAMTI CA
AA62 MYANMAR LETTER KHAMTI CHA
AA63 MYANMAR LETTER KHAMTI JA
AA64 MYANMAR LETTER KHAMTI JHA
AA65 MYANMAR LETTER KHAMTI NYA
AA66 MYANMAR LETTER KHAMTI TTA
AA67 MYANMAR LETTER KHAMTI TTHA
AA68 MYANMAR LETTER KHAMTI DDA
AA69 MYANMAR LETTER KHAMTI DDHA
AA6A MYANMAR LETTER KHAMTI DHA
AA6B MYANMAR LETTER KHAMTI NA
AA6C MYANMAR LETTER KHAMTI SA
AA6D MYANMAR LETTER KHAMTI HA
AA6E MYANMAR LETTER KHAMTI HHA
AA6F MYANMAR LETTER KHAMTI FA
AA70 MYANMAR LETTER KHAMTI REDUPLICATION
AA71 MYANMAR LETTER KHAMTI XA
AA72 MYANMAR LETTER KHAMTI ZA
AA73 MYANMAR LETTER KHAMTI RA

Khamti Extensions

AA74 MYANMAR LOGOGRAM KHAMTI OAY
AA75 MYANMAR LOGOGRAM KHAMTI QN
AA76 MYANMAR LOGOGRAM KHAMTI HM

Aiton Extensions

AA77 MYANMAR SYMBOL AITON EXCLAMATION
AA78 MYANMAR SYMBOL AITON ONE
AA79 MYANMAR SYMBOL AITON TWO
AA7A MYANMAR SYMBOL AITON RA

Pa'o Karen Tone Mark

AA7B MYANMAR SIGN PAO KAREN TONE

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.

Sai Kam Mong 2004, *The History and Development of The Shan Scripts* Silkworm Books, ChiangMai Thailand

Stribley, Keith "Collation of Myanmar (Burmese) in Unicode" (unpublished manuscript, 2009) http://www.thanlwinsoft.org/ThanLwinSoft/MyanmarUnicode/Sorting/

Stern, Theodore "Three Pwo Karen Scripts: A Study of Alphabet Formation" (Anthropological Linguistics, Vol 10, No. 1, 1968)

The Unicode Consortium 2006, *The Unicode Standard, Version 5.0* Addison-Wesley, Massachusetts.

Afterward

As a researcher, it is impossible to create a document such as this without the help of many people. There are too many people to name them all, but two people, Keith Stribley and Ngwe Tun, stand out as those who have worked to encourage me to make this document as accurate and useful as possible and have provided valuable information and insight.

This research has been undertaken with the support of Payap University and now as part of the work of the Payap University Linguistics Institute.

This document consists entirely of text conformant to Unicode 5.2 and was typeset using a version of OpenOffice with Graphite support. This has enabled me to use only one font for all the Myanmar script samples: Padauk. The various stylistic variants are enabled through the use of the features mechanism Graphite offers.