Collation of Myanmar (Burmese) in Unicode

Sorting Myanmar in Unicode according to “Spelling Book Order”

Introduction

This document presents an algorithm for sorting text in the Myanmar language, which is often still referred to as Burmese. There are two main sorting orders that have been used with Myanmar: “Pali Order” and “Spelling Book Order”. The former was used in older dictionaries, whereas variations on the latter are used in most modern dictionaries.[1]

The algorithm presented here will focus on the “Spelling Book Order” as it is used in modern Myanmar. There are some subtle variations to this order used by different dictionaries, but the order used here will be based on the Myanmar Language Commission Spelling Dictionary.[2]

There are many different non-Unicode encodings for Myanmar text, however, these are based on glyphs rather than linguistic symbols. Using the same encoding it is possible to type the same word in different ways whilst keeping the spelling the same. For example, အမှူး or အမှူး (leader), both are spelled in the same way but using a variation on the glyph for the uu vowel (typed as *trSL;* and $tr^{a};$ respectively in the WinInnwa font). This makes collation hard because lots of combinations have to be worked out.

Collation in Unicode is simpler because in most cases there is no variation in how the word is spelled in terms of code points. The choice of which glyphs to use is made by the font, not the typist. This document will assume that Myanmar[3] is encoded according to the table in Unicode 4, Chapter 10.3[4] and Unicode Technical Note 11.[5]

The notation used here is intended for the purposes of collation only and sometimes may not represent the normal linguistic nomenclature.

Collation Elements

Myanmar is collated based on syllables. A Myanmar syllable encoded in Unicode can be broken into 5 parts for collation:

<consonant><medial><vowel><final><tone>

Only the consonant is always present, one or more of the other parts may be empty in any given syllable. In practice the vowel may be displayed before the consonant e.g. ကေ , but it is encoded as U+1000 (Myanmar letter KA က) U+1031 (Myanmar vowel sign E ေ).

The resulting collation sequence has 5 levels, order of priority: <consonant>, <medial>, <final>, <vowel>, <tone>. Note, that the final and vowel have been switched from their encoded order. Each of these parts of the syllable may be composed of one or more characters as the following tables show.

[1] *Burmese: An Introduction to the Script*, John Okell, 1994, SOAS, Appendix 6: Alphabetic Order.

[2] Myanmar Spelling Dictionary, Myanmar Language Commission, Second Edtion, 2003.

[3] http://www.unicode.org/charts/PDF/U1000.pdf

[4] http://www.unicode.org/versions/Unicode4.0.0/ch10.pdf

[5] http://www.unicode.org/notes/tn11/

Consonant

Collation Order – read left to right and then down. The data is presented in the traditional layout of the alphabet.

Glyph		Code	Glyph		Code	Glyph		Code	Glyph		Code	Glyph		Code	Glyph		Code
C1	က	U+1000	C2	ခ	U+1001	C3	ဂ	U+1002	C4	ဃ	U+1003	C5	င	U+1004			
C6	စ	U+1005	C7	ဆ	U+1006	C8	ဇ	U+1007	C9	ဈ	U+1008	C9	ဉ	U+1009	C10	ည	U+100A
C11	ဋ	U+100B	C12	ဌ	U+100C	C13	ဍ	U+100D	C14	ဎ	U+100E	C15	ဏ	U+100F			
C16	တ	U+1010	C17	ထ	U+1011	C18	ဒ	U+1012	C19	ဓ	U+1013	C20	န	U+1014			
C21	ပ	U+1015	C22	ဖ	U+1016	C23	ဗ	U+1017	C24	ဘ	U+1018	C25	မ	U+1019			
C26	ယ	U+101A	C27	ရ	U+101B	C28	လ	U+101C	C29	ဝ	U+101D	C30	သ	U+101E			
			C31	ဟ	U+101F	C32	ဠ	U+1020	C33	အ	U+1021						

Note 1: The relative order is the same as the code points themselves.

Note 2: C33 is actually the A vowel, but it behaves in many ways like a consonant in regards to the other parts of the syllable.

Note 3: It may be appropriate to append the "Various Signs" U+104C ... U+104F at the end of this class – see comments in *Other Myanmar Characters* below.

Medials

The case where there is no medial is also included so that the relative sequence is clear.

Order	Glyph(s)	Unicode Sequence
M0	–	-
M1	္ယ	U+1039 U+101A
M2	္ရ	U+1039 U+101B
M3	္ဝ	U+1039 U+101D
M4	္ဟ	U+1039 U+101F
M5	္ယ္ဝ	U+1039 U+101A U+1039 U+101D
M6	္ယ္ဟ	U+1039 U+101A U+1039 U+101F
M7	္ရ္ဝ	U+1039 U+101B U+1039 U+101D
M8	္ရ္ဟ	U+1039 U+101B U+1039 U+101F
M9	္ဝ္ဟ	U+1039 U+101D U+1039 U+101F
M10	္ရ္ဝ္ဟ	U+1039 U+101B U+1039 U+101D U+1039 U+101F

Note 1: the combined medials are treated as one unit for collation not as a sequence of component medials.

Note 2: when the consonant is U+1004 a Zero Width Joiner (U+200D) is inserted before the medial to disambiguate it from the case of Kinzi (U+1004 U+1039 င်္), which is actually a final of U+1004. Since U+200D is normally ignored for collation it is not included in the table.

Vowels

Order	*Glyph*	*Unicode Sequence*	*Order*	*Glyph(s)*	*Unicode Sequence*
V0	-	-	V6	ေ◌	U+1031
V1	◌ါ/◌ာ	U+102C	V7	◌ဲ	U+1032
V2	◌ိ	U+102D	V8	ေ◌ါ / ေ◌ာ	U+1031 U+102C
V3	◌ီ	U+102E	V9	ေ◌ါ်/ ေ◌ာ်	U+1031 U+102C U+1039 U+200C
V4	◌ု	U+102F	V10	◌ံ	U+1036 – see note below
V5	◌ူ	U+1030	V11	◌ို	U+102F U+102D

Note 1: V9 is actually the low tone form of V8, but it is included here, because it does not use the normal tone marks. There is never a final after V3, V5, V9 or V10.

Note 2: V10 is not really a vowel, however, when there is no other vowel it is treated as one for collation. When it occurs in the sequence U+102D U+1036 or U+102F U+1036, it is instead the U+1036 is collated as if it was a final U+1019 U+1039, which is what it is linguistically.[6]

Finals

Finals may are marked with a Myanmar sign virama U+1039 character in Unicode. Normally this is a visible virama, which is represented as U+1039 U+200C. However, the U+200C Zero Width Non Joiner can usually be ignored for collation.

If the U+200C is not present, then the consonant of the following syllable will be displayed underneath the final. In a few rare cases, a ligature of the final and the following consonant is used instead, but these are rendering issues and are not relevant for collation.

Glyph		*Code*	*Glyph*		*Code*	*Glyph*		*Code*	*Glyph*		*Code*	*Glyph*		*Code*	*Glyph*		*Code*
F0	–	-	F1	က်	U+1000 U+1039	F2	ခ်	U+1001 U+1039	F3	ဂ်	U+1002 U+1039	F4	ဃ်	U+1003 U+1039	F5	င်	U+1004 U+1039
F6	စ်	U+1005 U+1039	F7	ဆ်	U+1006 U+1039	F8	ဇ်	U+1007 U+1039	F9	ဈ်	U+1008 U+1039	F9	ဉ်	U+1009 U+1039	F10	ည်	U+100A U+1039
F11	ဋ်	U+100B U+1039	F12	ဌ်	U+100F U+1039	F13	ဍ်	U+100D U+1039	F14	ဎ်	U+100E U+1039	F15	ဏ်	U+100F U+1039			
F16	တ်	U+1010 U+1039	F17	ထ်	U+1011 U+1039	F18	ဒ်	U+1012 U+1039	F19	ဓ်	U+1013	F20	န်	U+1014 U+1039			
F21	ပ်	U+1015 U+1039	F22	ဖ်	U+1016 U+1039	F23	ဗ်	U+1017 U+1039	F24	ဘ်	U+1018	F25	မ်	U+1019 U+1039			
F26	ယ်	U+101A				F28	လ်	U+101F U+1039				F30	သ်	U+101E U+1039			
						F32	ဠ်	U+1020									

[6] This is in contrast to John Okell, who collates U+1036 as equivalent to U+1019 U+1039. e.g. *Burmese/Myanmar Dictionary of Grammatical Forms*, John Okell & Anna Allott, 2001, Curzon Press.

Tones

Order	*Glyph*	*Unicode*
T0	-	-
T1	◌့	U+1037
T2	◌း	U+1038
T3	◌့း	U+1037 U+1038

Note 1: T3 is not normally found in a dictionary, it marks a genitive in some situations. It might however be found in book indexes etc.

Independent Vowels

The Independent vowels are collated as if they were written with အ U+1021 (Myanmar letter A) and the corresponding vowel. In some cases they may be followed by း U+1038 (Myanmar sign visarga), which collates in the same way as normal.

Order	*Glyph*	*Unicode Sequence*	*Equivalent Representation*	*Equivalent Sequence for Collation*	*Equivalent Collation Elements*
IV1	ဣ	U+1023	အိ	U+1021 U+102D	C33 V2
IV2	ဤ	U+1024	အီ	U+1021 U+102E	C33 V3
IV3	ဥ	U+1025	အု	U+1021 U+102F	C33 V4
IV4	ဦ	U+1026 or (U+1025 U+102E)	အူ	U+1021 U+1030	C33 V5
IV5	ဧ	U+1027	အေ	U+1021 U+1031	C33 V6
IV6	ဩ	U+1029	အော	U+1021 U+1031 U+102C	C33 V8
IV7	ဪ	U+102A	အော်	U+1021 U+1031 U+102C U+1039 U+200C	C33 V9

Note 1: although these are equivalent for the purposes of collation, usually only one representation is correct in a given word.

Note 2: the independent vowels may take finals, tones and even V4 when combined with a final.

Contractions

There are a few words which are written with a repeated consonant omitted, but which should be collated as if the consonant was still present. The number of these is small, though this list is probably not complete.

Word	*Meaning*	*Unicode Representation*	*Collation Equivalent*	*Collation Elements*
ယောက်ျား	man	U+101A U+1031 U+102C U+1000 U+1039 U+200C U+1039 U+101A U+200C U+1038	U+101A U+1031 U+102C U+1000 U+1039 U+200C **U+1000** U+1039 U+101A U+200C U+1038	C26 V8 F1 C1 M1 V1 T2
ကျွန်ုပ်	1st person singular	U+1000 U+1039 U+101A U+1039 U+101D U+1014 U+1039 U+200C U+102F U+1015 U+1039 U+200C	U+1000 U+1039 U+101A U+1039 U+101D U+1014 U+1039 U+200C **U+1014** U+102F U+1015 U+1039 U+200C	C1 M5 F20 C20 V4 F21

Short Forms

These are variations on the normal spelling, that are still found in current use. If they need to be collated, then they should be collated with their normal spelling not the variant. However, implementing this level of collation support for Myanmar can probably be regarded as optional. They can be regarded as "short forms" because their left to right width is less.

Word	*Short Form*	*Normal Spelling*	*Short form Unicode Sequence*	*Normal Unicode Sequence*	*Collation Elements*
daughter	သ္မီး	သမီး	U+101E **U+1039** U+1019 U+102E U+1038	U+101E U+1019 U+102E U+1038	C30 C25 V3 T2
cooked rice	ထ္မင်း	ထမင်း	U+1011 **U+1039** U+1019 U+1004 U+1039 U+200C U+1038	U+1011 U+1019 U+1004 U+1039 U+200C U+1038	C17 C25 F5 T2
tea	လ္ဘက်	လက်ဖက် or လက်ဘက်	U+101C U+1039 U+1018 U+1000 U+1039 U+200C	U+101C **U+1000** U+1039 **U+200C** *U+1018* U+1000 U+1039 U+200C	C28 F1 C25 F1

Note 1: In the first 2 examples the U+1039 should be ignored and is purely a trick to get the correct rendering.

Note 2: The third example is different in that the final က် has been dropped from the first consonant, so the U+1000 and U+200C have been removed. The second consonant is normally now spelled as U+1016 ဖ (PHA) not U+1018 ဘ (BHA), but it should probably still be collated as U+1018.

Other Myanmar Characters

The Myanmar symbols are not normally collated, however, it is probably legitimate to append them to consonant class, though they would never take any of the other syllable components. If they are collated they should be collated below the consonants in the order: ၌ locative U+104C; ၍ completed U+104D; ၎င်း aforementioned U+104E; ၏ genitive U+104F.

Myanmar sign little section (U+104A) and Myanmar sign section (U+104B) can normally be ignored or treated at a lower level similar to collation of punctuation in other languages.

Myanmar digits can be treated at a primary level as equal to the digits of other languages and and at a secondary level on a script basis as per the Unicode standard.

Implementation

Glibc

An implementation has been written for Glibc. This combines the collation elements for <consonant><medial> into one unit to avoid ambiguities of the sequence <consonant> + U+1039 which occurs both with a medial and as a final. <vowel><final> are also combined to allow the <final> to take precedence over the vowel. This gives a large number of collation elements, but gives correct results. Performance is probably sub-optimal because of the large number of collation elements used.

ICU

An implementation has also been written for ICU. It uses a large number of collation elements, combining <consonant><medial> and <vowel><final> to ensure the correct sequence. There is probably a lot of scope to optimise it, but it might require changes to the ICU source code.

Examples

The table below shows a selection of words to illustrate the 5 different orders of collation.

Word	*Collation Element in 1st Syllable*					*Unicode Sequence*
	Consonant	*Medial*	*Final*	*Vowel*	*Tone*	
ကခုန္‌	C1	M0	F0	V0	T0	**U+1000** U+1001 U+102F U+1014 U+1039 U+200C
ကာ	C1	M0	F0	V1	T0	**U+1000 U+102C**
ကား	C1	M0	F0	V1	T2	**U+1000 U+102C U+1038**
ကိရိယာ	C1	M0	F0	V2	T0	**U+1000 U+102D** U+101B U+102D U+101A U+102C
ကုဗပေ	C1	M0	F0	V4	T0	**U+1000 U+102F** U+1017 U+1015 U+1031
ကေဒာ	C1	M0	F0	V6	T0	**U+1000 U+1031** U+1012 U+102C
ကဲလ္ဝန္‌	C1	M0	F0	V7	T0	**U+1000 U+1032** U+101C U+1039 U+101D U+1014 U+1039 U+200C
ကဲ့	C1	M0	F0	V7	T1	**U+1000 U+1032 U+1037**
ကောလီက္ရေ	C1	M0	F0	V8	T0	**U+1000 U+1031 U+102C** U+101C U+102E U+1000 U+1039 U+101B U+1031
ကော့လန္‌	C1	M0	F0	V8	T1	**U+1000 U+1031 U+102C U+1037** U+101C U+1014 U+1039 U+200C
ကော္‌လံ	C1	M0	F0	V9	T0	**U+1000 U+1031 U+102C U+1039 U+200C** U+101C U+1036
ကံ	C1	M0	F0	V10	T0	**U+1000 U+1036**
ကုိ	C1	M0	F0	V11	T0	**U+1000 U+102F U+102D**
ကက္ကရာ	C1	M0	F1	V0	T0	**U+1000 U+1000 U+1039** U+1000 U+101B U+102C
ကက္‌ကင္‌းဓာတ္‌	C1	M0	F1	V0	T0	**U+1000 U+1000 U+1039 U+200C** U+1000 U+1004 U+1039 U+200C U+1038 U+1013 U+102C U+1010 U+1039 U+200C
ကုက္ကလံ	C1	M0	F1	V4	T0	**U+1000 U+102F U+1000 U+1039** U+1000 U+101C U+1036
ကောက္‌ခံ	C1	M0	F1	V8	T0	**U+1000 U+1031 U+102C U+1000 U+1039 U+200C** U+1001 U+1036

Word	*Collation Element in 1st Syllable*					*Unicode Sequence*
	Consonant	*Medial*	*Final*	*Vowel*	*Tone*	
ကုိက္‌	C1	M0	F1	V11	T0	**U+1000 U+102F U+102D U+1000 U+1039 U+200C**
ကင္‌မရာ	C1	M0	F5	V0	T0	**U+1000 U+1004 U+1039 U+200C** U+1019 U+101B U+102C
ကင္‌းစီး	C1	M0	F5	V0	T2	**U+1000 U+1004 U+1039 U+200C U+1038** U+1005 U+102E U+1038
ကုမ္ပဏီ	C1	M0	F25	V4	T0	**U+1000 U+102F U+1019 U+1039** U+1015 U+100F U+102E
ကုံင	C1	M0	F25	V4	T0	**U+1000 U+102F U+1036** U+1004
ကုံး	C1	M0	F25	V4	T2	**U+1000 U+102F U+1036 U+1038**
ကယ္‌ခ္ယ္ဝတ္‌	C1	M0	F26	V0	T0	**U+1000 U+101A U+1039 U+200C** U+1001 U+1039 U+101A U+1039 U+101D U+1010 U+1039 U+200C
က္ယ	C1	M1	F0	V0	T0	**U+1000 U+1039 U+101A**
က္ယာ	C1	M1	F0	V1	T0	**U+1000 U+1039 U+101A U+102C**
က္ရ	C1	M2	F0	V0	T0	**U+1000 U+1039 U+101B**
က္ရောင္‌့	C1	M2	F4	V8	T1	**U+1000 U+1039 U+101B U+1031 U+102C U+1004 U+1039 U+200C U+1037**
က္ဝာခ္ရား	C1	M3	F0	V1	T2	**U+1000 U+1039 U+101D U+102C** U+1001 U+1039 U+101B U+102C U+1038
က္ယ္ဝေး	C1	M5	F0	V6	T2	**U+1000 U+1039 U+101A U+1039 U+101D U+1031 U+1038**
က္ရ္ဝားဝါ	C1	M7	F0	V1	T2	**U+1000 U+1039 U+101B U+1039 U+101D U+102C U+1038** U+101D U+102C
ခမ္ယာ	C2	M0	F0	V0	T0	**U+1001** U+1019 U+1039 U+101A U+102C
အိတ္‌ကပ္‌	C33	M0	F16	V2	T0	**U+1021 U+102D U+1010 U+1039 U+200C** U+1000 U+1015 U+1039 U+200C
ဣတ္ထိလိင္‌	C33	M0	F16	V2	T0	**U+1023 U+1010 U+1039** U+1011 U+102D U+101C U+102D U+1004 U+1039 U+200C
အုတ္‌	C33	M0	F16	V4	T0	**U+1021 U+102F U+1010 U+1039 U+200C**

Note that although several of the words are multi-syllable, the first syllable (code points in bold) is sufficient to determine the sort order in all cases in this example. (The one exception is for the 2 examples with C1 M0 F1 V0 T0, where the second syllable controls sorting).

Conclusions

An algorithm for Myanmar Collation has been presented in terms of 5 levels of collation elements within a syllable. In order of precedence these are: <consonant>, <medial>, <final>, <vowel>, <tone>, where <vowel> is encoded before <final> according to Unicode. The independent vowels should be collated as equivalent to the same vowel sound written with Myanmar letter A (U+1021). The collation is complicated because the same code sequence may be found in several places within the syllable, so a partial context analysis may be required to disambiguate. In addition, a complete implementation should take account of contractions and short forms.

August 22, 2005
Revision 295
Keith Stribley