Collation Weight Design for Myanmar Unicode Texts

Tin Htay Hlaing
Nagaoka University of Technology
Nagaoka, JAPAN.
tinhtayhlaing@gmail.com

Yoshiki Mikami
Nagaoka University of Technology
Nagaoka, JAPAN.
mikami@kjs.nagaokaut.ac.jp

Abstract

This paper proposes collation weights for sorting Myanmar Unicode texts in conformity with *Unicode Collation Algorithm*. Our proposal is constructed based on the sorting order given in Myanmar Spelling Book which is also known as Myanmar Orthography. Myanmar language has complex linguistic features and needs preprocessing of texts before collation. Thus, we examine syllable structure, standard collation orders of Myanmar characters and then some preprocessing steps in detail. Furthermore, we introduced mathematical definitions such as *Complete Order Set (COSET)* and *Product of Order* to describe Myanmar sorting order in a formal way. This formal description gives clear and unambiguous understanding of Myanmar lexicographic order to foreigners as well as natives.

1 Aim of Research

Myanmar Language is the official language of Myanmar spoken by 32 million people as first language and as second language by some ethnic minorities of Myanmar.

However, awareness of Myanmar lexicographic sorting or collation order is lacking among the public. In Myanmar, people rarely look up a word in dictionary using Myanmar lexicographic order because most of the dictionaries published in Myanmar are English-to-Myanmar dictionaries. For example, Engineering dictionary, Medical dictionary and Computer dictionary use English lexicographic order. Likewise, in telephone directory and yellow pages in Web, the Myanmar names are Romanized and sorted according to the conventional Latin alphabet order.

Moreover, Microsoft`s window operating system and office applications for desktop publishing are commonly used in Myanmar. But, until now, Microsoft applications such as Word and Excel simply sorts Myanmar character based on binary code values and it does not meet Myanmar traditional sorting requirements.

Only a few professional linguists compile dictionaries but collation rules used by them are not understood easily by others. Lack of understanding of collation order in turn makes importance of collation unaware.

Consequently, these factors mentioned above act as a barrier to the implementations of Myanmar Language in database managements system and other ICT applications.

Therefore, we design collation weights for Myanmar Unicode texts and employed these collation weights in implementation of lexicographic sorting algorithm for Myanmar texts. Our algorithm is tested to sort the words using Myanmar Spelling Book order or Myanmar Orthography (Myanmar Language Commission, 2006) and proved its proper working. Also, we propose a formal description method for collation order using concepts of ordered set so that Myanmar sorting rules are well understood by both foreigners and natives.

2 Formal Description of Sorting Order

Standard sorting rules are often defined by national level language committee or are given in arrangement of words in the dictionaries. These rules are, however, complex, and difficult to understand well for those who are foreign to that language. Therefore, we introduced a formal description method for collation orders of syllabic scripts by employing a concept of Complete Order Set, COSET.

2.1 Definitions

Definition 1 Complete Order Set

In Set Theory, it is defined that a complete order is a binary relation (here denoted by infix $\leq$) on a set X. The relation is transitive, antisymmetric and total. A set equipped with a complete order is called a complete ordered set. If X is ordered

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

under ≤, then the following statements hold for all *a*, *b* and *c* in *X*:
(Transitivity) if $x \le y$ and $y \le z$ then $x \le z$
(Antisymmetry) if $x \le y$ and $y \le x$ then $x = y$
(Totality) $x \le y$ or $y \le x$
(Moll and et al, 1988)

Definition 2 Product of Order

Let X and Y be COSETs and $x_i \in X$, $y_i \in Y$. Then, the product of order $(X, \le) \times (Y, \le)$ is defined as
$x_1y_1 < x_2y_2 \Leftrightarrow x_1 = x_2$ and $y_1 < y2$ or
$x_1y_1 < x_2y_2 \Leftrightarrow x_1 \le x_2$
As seen from above definition, product of order is not commutative.

Definition 3 Lexicographic Order

Let $(X, \le)$ be a COSET, and X* be the set of all finite-length strings over X, then a lexicographic order on X*, $(X^*, \le_{LEX})$ is defined:
For all $u, v \in X^*$, $u = u_1 \cdots u_n$ and $v = v_1 \cdots v_m$, where $m, n \ge 0$ and $u_j, v_j \in X$, we set $u \le_{LEX} v$ if and only if:
(1) $u_i = v_i$ for $i = 1,..., m$ and $m \le n$; or
(2) there is $j \ge 1$ such that $u_i = v_i$ for $i = 1,...,j-1$ and $u_j < v_j$.

Suppose M be the maximum length of strings in X*, and ø be the null element of the alphabet which always comes before letter, then fixed length representation of a string u is given by $u = u_1u_2..u_n$ ø… ø where ø occurs M-n times.

Then above defined lexicographic order over X^M can be written by using product of order (Mikami and et al, 2010).
$(X^M, \le_{LEX}) = (X,\le) \times \ldots..\times(X,\le)$
$= (X,\le)^M$

2.2 Multilevel Sorting

Sorting that conforms to linguistic requirements of a particular language cannot be done by a simple comparison of code points and a single level of comparison. Multilevel sorting has become necessary to get the results that satisfy with users` expectations.

Let`s think about English lexicographic sorting. The primary level comparison is made based on case-insensitive comparison of alphabet `a` which will be treated as the same order with `A`. Then the secondary level comparison is made based on the case of each letter.

In order to describe above lexicographic order, we introduced two COSETs, (L,<) and (C,<) where (L,<) is a standard alphabet letter order a<b<…..<z and (C,<) is a case order s<c (means small letter comes before capital letter). If a null element is ø is added to both, two CO-SETs finally become:
(L,<) = ({ø,a,..,z}, ø<a<……<z)
(C,<) = ({ø,s,c}, ø<s<c)
Then case-sensitive lexicographic order $(X^M, <_{LEX})$ on length M string of $x \in X$ can be formally described as
$(X^M, <_{LEX}) = (L,<)^M \times (C,<)^M$

3 Myanmar Character Set

In this section, the lexicographic order of Myanmar characters is described using complete order set notation.

Myanmar texts flow from left to right and no space between words or syllables. The categories of Myanmar characters defined in Unicode and in our proposal are shown in Table 1.

Basically, Myanmar script has 33 consonants, 8 vowels, 2 diacritics, 11 medials, ASAT, 10 digits and 2 punctuation marks. Every character group is non-ignorable in sorting process except punctuation marks and some special characters ၌ ၍ ၎ ၏. We describe Myanmar characters using set notation as Myanmar Character Set= {C, V, I, D, M, N, F, ASAT, ၌, ၍ , ၎, ၏ }.

Unicode Category with number of elements	Proposed Category with abbreviation
Consonants (33)	Consonants(C)
Dependent (8) and Independent vowels (8)	Vowels (V , I)
Various Signs (5)	Diacritics(D) and ASAT
Consonant Conjuncts (4)	Medials (M)
Digits (10)	Numerals (N)
Punctuations (2)	Punctuations
	ASAT+ Consonants (F)

"Table 1. Categories of Myanmar Characters."

3.1 Consonants

Myanmar consonant has default vowel sound and itself works as a syllable. The set of consonants is C={က, ခ, ဂ, ဃ, င, စ, ဆ, ဇ, ဈ, ဉ, ည ,ဋ ,ဌ,

ဥ, ပ, ဏ, တ, ထ ,ဒ ,ဓ ,န ,ပ ,ဖ, ဗ, ဘ ,မ ,ယ ,ရ, လ, ဝ, သ, ဟ, ဠ, အ}becomes a COSET if normal consonant order is given. The order defined on C is written as (C,<).

3.2 Vowels

The set of Myanmar dependent vowel characters in Unicode contains 8 elements { ါ, ာ, ိ, ီ, ု, ူ, ေ, ဲ}. But some of them are used only in combined form such as
အ + ိ → အိ or အ + ီ → အီ or အ + ု → အု and some characters which should be contracted before sorting such as အ + ိ + ု → အို. So, the vowel order is defined only on modified vowel set V` composed of 12 elements {အ, အာ, အိ, အီ, အု, အူ, အေ , အဲ, အော, အော်, အံ, အို}which it is written as (V`,<).

The vowel order is also not defined on independent vowel set I={အ, ဣ, ဤ, ဥ, ဦ, ဧ, ဩ, ဪ} because one letter ဦ needs to be normalized to arrange it in a complete order. Thus, vowel order is defined only on modified independent vowel set (I`, <).

(V`,<) and (I`,<) are isomorphic.

3.3 Medials

There are four basic medials characters M={ျ, ြ, ွ, ှ}which combine and produce total 11 medials. The set of those 11 medials have an order (M`, <).

3.4 Finals or Ending Consonants

The set of finals F={က် ,ခ် ,ဂ်, ဃ် ,င် ,စ် ,ဆ် ,ဇ် ,ဈ် ,ည်, ဋ် ,ဌ် ,ဍ် ,ဎ် ,ဏ် ,တ် ,ထ် ,ဒ် ,ဓ် ,န် ,ပ် ,ဖ် ,ဗ် ,ဘ် ,မ် ,ယ် ,ရ် , လ်, ဝ် ,သ် ,ဟ် ,ဠ် }having an order (F,<) is isomorphic to (C,<).

3.5 Diacritics

Diacritics alter the vowel sounds of accompanying consonants and they are used to indicate tone level. There are 2 diacritical marks { ့, း } in Myanmar script and their order is (D,<).

3.6 Digits

Myanmar Language has 10 numerals N={ ၀, ၁, ၂, ၃, ၄, ၅, ၆, ၇, ၈, ၉}with the order (N,<).

3.7 Myanmar Sign ASAT

When there is a consonant at the end of a syllable, it carries a visible mark call ASAT (်) to indicate that the inherent vowel is killed. It usually comes with consonants but sometimes comes with independent vowels.

4 Myanmar Syllable Structure in BNF

Most European languages have orders defined on letters, but those languages which use syllabic scripts, including Myanmar, have an order defined on a set of syllables. In Myanmar Language, syllable components such consonants(C), independent vowel(I) and digits(N) are standalone components while dependent vowels(V), medials(M), diacritics(D) and finals(F) are not. Among them, consonants can also act as nucleus of syllable so that other characters can attach to it in different combinations. The structure of Myanmar Syllable can be illustrated using BNF notation as S:= C{M}{V}[F][D] | I[F] | N where { } means zero or more occurrences and [ ] means zero or one occurrence.

5 Comparison of Multilevel sorting in English and Myanmar

In English, the sorting process does on word level. After completing primary level comparison for a given word, the secondary level comparison is started for this word again. For Myanmar Language, the lexicographic order is defined by the product of five generic components, namely, consonant order (C,<), medial order (M,<), final order (F,<), vowel order (V,<) and diacritics order (D,<). Therefore Myanmar syllable order (S,<) is given by formula

$(S,<)^M = (C,<) \times (M,<) \times (F,<) \times (V,<) \times (D,<)$

Interesting to note here is the fact that the order (F,<) is considered before (V,<) while F comes after V in a coded string.

Thus, Myanmar sorting process does on syllable-wise behavior. For instance, a given word may contain one or more syllables. Sorting is done on first syllable and if there is no difference, the process will move to next syllable.

To sum up, firstly, one sort key is generated for one word in English while it is generated for each syllable in Myanmar. Secondly, multilevel sorting is done on word level in English but it is done within a syllable in Myanmar. Thirdly, in English, it needs whole word information because sorting process goes until it reaches end of word and then goes to next level. In contrast,

Myanmar sorting process goes until it reaches to last comparison level for one syllable and then it moves to next syllable. It means that Myanmar does not need whole word information if there is a difference before the final syllable.

6 Preprocessing of Texts

6.1 Syllabification

Myanmar is syllable based language and thus syllabification is necessary before collation. This process needs to return not only syllable boundary but also the type of each component within a syllable. Maung and Mikami (2008) showed syllable breaking algorithm with a complete set of syllabification rules.

6.2 Reordering

If Unicode encoding is followed, Myanmar characters are not stored in visual order. Myanmar vowels and consonant conjuncts are traditionally being typed in front of the consonants but we store Myanmar syllable components according to this order: <consonant> <medial> <vowel> <ending-consonant> <diacritic>. Therefore, no reordering is required for Myanmar Unicode texts.

6.3 Normalization

Normalization is required if a letter or ligature is encoded in composed form as well as decomposed form. One Myanmar character has multiple representations and thus normalization is required.

Decomposed Form	Unicode for Decomposed forms	Equivalent Composed form	Unicode for Composed Form
ဥ + ီ	1025 102E	ဦ	1026

"Table 2. Normalization of a Vowel."

6.4 Contractions

For some Myanmar dependent vowels, and medials, two or more characters clump together to form linguistic unit which has its own identity for collation. This group is treated similarly as a single character. These units may not be directly encoded in Unicode but are required to be created from their constituent units which are encoded. This process is called *contraction* (Hussain and Darrani, 2004).

Glyph	Unicode for Contraction	Description
ေ + ာ	1031+102C	Vowel sign E + AA
ေ + ာ +်	1031+102C+103A	Vowel sign E+AA+ASAT
ိ + ူ	102D + 102F	Vowel sign I + UU

"Table 3. Vowel Contractions."

Glyph	Unicode for Contraction	Description
ျ + ွ	103B + 103D	Consonant Sign Medial YA + WA
ြ + ွ	103C + 103D	Consonant Sign Medial RA + WA
ျ + ှ	103B + 103E	Consonant Sign Medial YA + HA
ြ + ှ	103C + 103E	Consonant Sign Medial RA + HA
ွ + ှ	103D + 103E	Consonant Sign Medial WA + HA
ျ + ွ + ှ	103B + 103D + 103E	Consonant Sign Medial YA+WA + HA
ြ + ွ+ ှ	103C + 103D + 103E	Consonant Sign Medial YA+WA + HA

"Table 4. Consonant Conjuncts Contractions."

Similarly, we need contractions for ending consonants or finals which is a combination of consonants and Myanmar Sign ASAT. Some of them are shown in table below.

Glyph	Unicode for Contraction	Description
ဘ + ်	1000+103A	Letter KA + ASAT
ခ + ်	1001+103A	Letter KHA+ ASAT
ဂ + ်	1002 + 103A	Letter GA + ASAT

"Table 5. Ending Consonant Contractions."

7 Myanmar Collation

Myanmar Language presents some challenging scenarios for collation. Unicode Collation Algorithm (Davis and Whistler, 2010) is to be modified for Myanmar Collation. Because, in UCA, only one final sort key is generated for one word. But, one Myanmar word is divided into a sequence of syllables and sort key is generated at the syllable level. We use five levels of collation with consonants getting the primary weights and conjunct consonants having the secondary weights. At the tertiary and quaternary levels, ending consonants and vowels will be sorted respectively. Finally, the quinternary level is used to sort diacritical marks. We ignore digits intentionally as there is no word with digits in dictionary. The collation levels and their respective weight ranges are depicted in table below.

Level	Components	Range
Primary	Consonants	02A1..02C2
Secondary	Consonant Conjuncts	005A..0064
Tertiary	Ending Consonants	0020..0050
Quaternary	Vowel	0010..001B
Quinternary	Diacritics	0001..000D

"Table 6. Collation Levels and Weights Range for Myanmar."

7.1 Collation Element Table for Myanmar

Some of the Unicode collation elements for Myanmar Language are given in Appendix A.

8 Conclusion

This paper proposes syllable-based multilevel collation for Myanmar Language. We also aimed to show how Unicode Collation Algorithm is employed for Myanmar words. We tested our algorithm to sort the words using Myanmar Orthography or Spelling Book Order and found that it works. But we have to do some more tests to handle loan syllable, kinzi, great SA and chained syllable so that we can produce more reliable evaluation. Myanmar language has some traditional writing styles such as consonant stacking eg. ဗုဒ္ဓ (Buddha), မန္တလေး (Mandalay, second capital of Myanmar), consonant repetition eg. တက္ကသိုလ် (University), kinzi eg. အင်္ဂတေ (Cement), loan words eg. ဘတ်(စ်) (bus). Although we write using above traditional styles, we read them in a simple way, for example, တက္ကသိုလ် will be read as တက်+ က + သိုလ် by inserting invisible Virama sign automatically. Therefore, syllabification process has to provide syllable boundary according to the way we read. If syllable breaking function does not work well, it may affect our result.

References

Mark Davis and Ken Whistler. 2010. *Unicode Collation Algorithm, Version 6.0.0.*

Myanmar Language Commission. 2006. *Myanmar Orthography,* Third Edition, University Press, Yangon, Myanmar.

Robert M. Moll, Michael A. Arbib, and A.J. Kfoury. 1988. *An Introduction to Formal Language Theory*, Springer-Verlag, New York, USA.

Sarmad Hussain and Nadir Darrani. 2008. *A Study on Collation of Languages from Developing Asia*, National University of Computer and Emerging Sciences, Lahore, Parkistan.

Yoshiki Mikami, Shigeaki Kodama, and Wunna Ko Ko. 2009. *A proposal for formal description method of Collating order*, Workshop on NLP for Asian Languages, Tokyo.1-8.

Zin Maung Maung and Yoshiki Mikami. 2008. *A Rule-based Syllable Segmentation of Myanmar Texts.* Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages, Hyderabad, India, 51-58.

Appendix A. <Myanmar Collation Element Table>

Glyph	Unicode	Collation Elements	Unicode Name
<- Consonants ->			
က	1000	02A1 005A 0020 0010 0001	MYANMAR LETTER KA
ခ	1001	02A2 005A 0020 0010 0001	MYANMAR LETTER KHA
ဂ	1002	02A3 005A 0020 0010 0001	MYANMAR LETTER GA
.	.	…….	……..
အ	1021	02C2 005A 0020 0010 0001	MYANMAR LETTER A
<- Consonant Conjunct Signs or Medials ->			
ျ	103B	0000 005B 0020 0010 0001	MYANMAR CONSONANT SIGN MEDIAL YA
ြ	103C	0000 005C 0020 0010 0001	MYANMAR CONSONANT SIGN MEDIAL RA
ွ	103D	0000 005D 0020 0010 0001	MYANMAR CONSONANT SIGN MEDIAL WA
.	…	…..	.....
ျွှ	103B103D103E	0000 0064 0020 0010 0001	MYANMAR CONSONANT SIGN MEDIAL YA+WA+HA
ြွှ	103C103D103E	0000 0065 0020 0010 0001	MYANMAR CONSONANT SIGN MEDIAL RA+WA+HA
<-Independent Vowels ->			
ဣ	1023	0000 0000 0000 0012 0001	MYANMAR LETTER I
ဤ	1024	0000 0000 0000 0013 0001	MYANMAR LETTER II
ဥ	1025	0000 0000 0000 0014 0001	MYANMAR LETTER U
ဦ	1025102E	0000 0000 0000 0015 0001	MYANMAR LETTER U + MYANMAR LETTER II
ဦ	1026	0000 0000 0000 0015 0001	MYANMAR LETTER UU
.	….	.....	…..
ဪ	102A	0000 0000 0000 0019 0001	MYANMAR LETTER AU
<- Dependent Vowels ->			
ါ	102B	0000 0000 0000 0011 0001	MYANMAR VOWEL SIGN TALL AA
ာ	102C	0000 0000 0000 0011 0001	MYANMAR VOWEL SIGN AA
ိ	102D	0000 0000 0000 0012 0001	MYANMAR VOWEL SIGN I
.	…..	…..	…..
ော	1031102C	0000 0000 0000 0018 0001	MYANMAR VOWEL SIGN E + AA
ော်	1031102C103A	0000 0000 0000 0019 0001	MYANMAR VOWEL SING E+AA+ASAT
ံ	1036	0000 0000 0000 001A 0001	MYANMAR SIGN ANUSVARA
ို	102D102F	0000 0000 0000 001B 0001	MYANMAR VOWEL SING I + U
<- Diacritics ->			
့	1037	0000 0000 0000 0000 000C	MYANMAR SIGN DOT BELOW
း	1038	0000 0000 0000 0000 000D	MYANMAR SIGN VISARGA
<- Ending Consonants or Finals ->			
ကဲ့	1000103A	0000 0000 0021 0010 0001	MYANMAR LETTER KA + MYANMAR SIGN ASAT
.	…..	…..	…..
အဲ့	1021103A	0000 0000 0042 0010 0001	MYANMAR LETTER A + MYANMAR SIGN ASAT