6K-6

A Proposal on Defining Collation Weights For Myanmar Unicode Texts

Tin Htay Hlaing
tinhtayhlaing@gmail.com
Katsuko T. Nakahira
katsuko@oberon.nagaokaut.ac.jp
Yoshiki Mikami
mikami@kjs.nagaokaut.ac.jp
NAGAOKA UNIVERSITY OF TECHNOLOGY

Introduction

This paper gives a proposal on defining collation elements for Myanmar Unicode texts in accordance with Unicode Collation Algorithm (UCA). Our proposal is constructed based on the sorting order given in Myanmar Spelling Book. Furthermore, we introduced a set of mathematical definitions employing a concept of Complete Order Set (COSET) to describe Myanmar sorting order in a formal way. This formal description gives clear and unambiguous understanding of sorting rules to those who are even non-natives.

1. Myanmar Character Set

According to Unicode version 6.0, Myanmar characters are grouped into consonants, independent vowels, dependent vowels, various signs, dependent consonant signs, great SA, digits, punctuation marks.

For collation, we firstly divide character sets into two categories such as non-ignorable and ignorable. We found that every character group is needed to be considered for collation except punctuation marks.

However, for sorting, we need to re-group characters into five major groups namely consonants C, dependent consonant signs or medials M, ending consonants or finals F (Consonants with ASAT), vowels V (independent and dependent) and diacritics D, to meet its sorting requirements.

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, unsystematic and difficult to understand. Therefore, we introduced a formal description method for collating orders of syllabic scripts by employing a concept of Complete Order Set.

2.1 Complete Ordered Set – COSET

The lexicographic order of a given language is described in the form of Complete Ordered Set (COSET). COSET is defined by a set of letters or syllables and an order relation < defined on them, so that it is written by the pair of two elements (X,<).

2.2 Multilevel Sort

Linguistically meaningful sorting is not done by a simple comparison of code points. Thus, multilevel sort has been introduced to get the results that conform to users` expectations.

We can describe the final order of multilevel sort using *product* of order sets. Product of order is defined as follows.

Let say $x_i \in X$, $y_i \in Y$, $xy \in X \times Y$ and X and Y are COSETs. Then, $x_1y_1 < x_2y_2$ if and only if $x_1=x_2$ and $y_1<y_2$ or $x_1 \leq x_2$. Note that products of orders are not commutative.

Again, the target order of primary comparison for COSET X is (X,<) where we only consider primary differences. Likewise, the target order of COSET Y for secondary comparison is (Y,<). Then, multilevel final combined order can be described using product of these two sets. When × is used as a symbol of product, we can write as (X,<) × (Y,<). If the level of comparison is more than two, the final lexicographic order for each level of comparison can be described by using product of order such as

(X1,<)×(X2,<)×(X3,<) ×………..× (Xn,<)

3 Describing Multilevel Myanmar Collation

3.1 Order defined on Syllable

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. Among them, consonants can also act as nucleus syllable so that other characters can attach to it in different combinations. The generic structure of Myanmar Syllable is 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.

And the order of syllables is further defined by the product of orders defined on consonant, vowel, diacritics, etc. Myanmar syllable order (S,<) is a 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,<)=(C,<)×(M,<)×(F,<)×(V,<)×(D,<) [1]

3.2 Myanmar Characters

Myanmar Language has total 34 consonants for Myanmar by including အ from Myanmar Independent vowels. We will collate them at the primary level.

The vowels that appear alone are independent vowels and those always appear with consonants are known as dependent vowels [2]. We will collate them by giving same collation weights.

There are four basic consonant-conjuncts but different combinations of these characters give seven more conjuncts.

Myanmar Sign Asat (◌်) included in Myanmar Various Signs , is used for devowelizing task. As it always comes with consonants and we call them as finals or ending consonants. We have 33 consonants except အ.

Diacritics alter the vowel sounds of accompanying consonants. There are 2 diacritical marks in Myanmar script. Also, Myanmar Language has its own 10 numerals.

4. Preprocessing of Texts

4.1 Syllabification

As Myanmar is syllable based language and thus syllabification is necessary for collation. But this process needs to return not only syllable boundary but also the type of each component within a syllable. Syllable breaking algorithm with a complete set of syllabification rules are given in [2].

4.2 Reordering

Though some languages such as Lao and Thai store the characters in visual order for backward compatibility, Myanmar characters are stored in this order : <consonant> <medials> <vowel> <ending-consonant> <diacritics>. Therefore, if Unicode encoding is followed, no reordering is required.

4.3 Normalization

One Myanmar independent vowel 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

4.4 Contractions

Some Myanmar dependent vowels, and medials are composed of one or more individual code points. Contraction should be done so that each combined form maps onto a single collation element.

Glyph	Name	Unicode for Contraction
ေ◌ + ◌ာ	Vowel	1031+102C
◌ျ + ◌ွ + ◌ှ	Medial	103B + 103D + 103E
င +◌်	Final	1004 + 103A

5. Myanmar Unicode Collation Weights

Myanmar Collation is different from Unicode Collation Algorithm (UCA). Because, in UCA, only one final sort key is generated for one word. For Myanmar, the number of final sort key is same as number of syllables in a given word.

We define five levels of collation for Myanmar and the respective weight ranges are as follows.

Level	Character Set	Range
Primary	Consonants	02A1..02C2
Secondary	Medials	005A..0064
Tertiary	Finals	0020..0050
Quatery	Vowel	0010..001B
Fifth level	Diacritics	0001..000D

6. Conclusion

This paper proposes syllable-based multilevel collation for Myanmar Language. We also aimed to show how Unicode Collation Algorithm is applied for Myanmar words. We tested our algorithm to sort the words using Spelling Book Order and found that it works. But we are doing some more tests to handle loan syllable, kinzi, great SA and chained syllable so that we can produce more reliable evaluation. Again, this algorithm totally depends on syllabification. This means if syllable breaking function does not work well, it may affect our result.

REFERENCES

[1] Mikami,Y., S.Kodama, W.Ko Ko(2009),A proposal for formal description method of Collating order, Workshop on NLP for Asian Languages, Tokyo,pp.1-8

[2] Zin Maung Maung and Yoshiki Mikami. 2008. A rule-based syllable segmentation of Myanmar text. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

[3] S.Hussain, N.Darrani, A Study on Collation of Languages from developing Asia http://www.panl10n.net (accessed date: Jan,13 2011)