Collation in Dzongkha

Pema Geyleg
Department of Information Technology
pema.geyleg@gmail.com

Abstract

The work on collation rules for Dzongkha was already started by Dzongkha Development Authority and Orient Foundation while trying to incorporate Dzongkha computing support in Microsoft Operating system. Initially, Mr. Sangay Dorji, <ddc@druknet.bt> Director General of Dzongkha Development Authority along with Mr. C. Fynn <cfynn@gmx.net> started the work of deducing the collation rules for Dzongkha. Then it was taken up by Mr. Robert Chilton <acip@well.com>. Final tweaking of the collation rules were carried out at Department of Information Technology.

1. Introduction

Collation, generally termed for the process and function of determining the sorting order of strings of characters. It provides a key function in a computer system; when a list of strings are presented to users, they would like to have it in a sorted order so that finding individual strings would be easy and reliable. Therefore collation i widely used in user interfaces. It is also a 'must-have' for the operation of databases, not only for sorting records but also to select sets of records with fields within given bounds.

But collation is not uniform; it differs from language to language and culture to culture: Germans, French and Swedes sort the same characters in different ways. Variation may also be by specific application: even within same language, dictionaries, phonebooks or book indices may sort differently from each other. East Asian ideographs, an example of non-alphabetic scripts, collation can either be phonetic or based on the character appearance. According to user preference, collation can also be commonly customized, for example: ignoring punctuation or not, putting uppercase before lowercase (or vice versa), and so on. Correct searching Linguistically, needs to be using the same mechanism as "v" and "w" sort as if they were the same base letter in Swedish, a loose search should sort with either of them.

Thus collation implementations must follow an often-complex linguistic convention that people have developed over time for ordering text in their language, and based on user preferences, provide a common customization. Performance is critical while doing all this.

The collation rules for Dzongkha were necessary for GNU C library locale to have Dzongkha collation rules support in Linux Operating System. Another necessity for the collation rule was in Open Office spread sheet and Open Office database.

2. Methods

2.1. Dzongkha script:

The Dzongkha script also called Bhutanese script is used to write Dzongkha. Dzongkha is the national language of Bhutan. The letter used to write is same as the script system used to write the Tibetan laguage (\x0F00 through \x0FCF) assigned in the Universal Character Set defined in the Unicode & ISO 10646 Standards. The Tibetan script is encoded using characters with values from U+0F00 to U+0FFF. The universal Character set defined for Tibetan is shown in the figure 1.

The writing direction for the Dzongkha script is from left to right and the written form consists of multiple stacking of different characters.

Alphabets

It consists of thirty consonants as shown below:

ཀ ཁ ག ང ཅ ཆ ཇ ཉ ཏ ཐ ད ན པ ཕ བ མ ཙ ཚ ཛ ཝ ཞ
ཟ འ ཡ ར ལ ཤ ས ཧ ཨ

2.2. Vowels

It consists of four basic vowel signs as shown below

◌ི ◌ུ ◌ེ ◌ོ

i u e o

3. Frequently used consonants

They are used for writing everyday words in Dzongkha.

རྐ རྒ རྔ རྗ རྙ རྟ རྡ རྣ རྦ རྨ རྩ རྫ།

ལྐ ལྒ ལྔ ལྕ ལྗ ལྟ ལྡ ལྤ ལྦ ལྷ།

སྐ སྒ སྔ སྙ སྟ སྡ སྣ སྤ སྦ སྨ སྩ། ཀྱ ཁྱ གྱ པྱ ཕྱ བྱ མྱ།

ཀྲ ཁྲ གྲ ཏྲ ཐྲ དྲ པྲ ཕྲ བྲ མྲ ཤྲ སྲ ཧྲ།

4. Sorting Unicode Dzongkha using multi weight collation Algorithm

Requirement within a computer environment for full Dzongkha support:

1. Keyboard(s) or other input methods
2. Rendering: readable, printable display of the encoded Dzongkha script data.
3. For generating culturally acceptable sorting, collation rules.

Single weight sorting method was used earlier as rule to sort Dzongkha. Within a specific application/environment, it was generally adequate for sorting native Dzongkha orthographies. Dzongkha-script sorting was treated in an exclusive, special case fashion; such proprietary sorting method was not implemented widely.

All these led to the development of multi weight collation rules for Dzongkha script. The present collation rule for Dzongkha script is compliant with the Unicode collation algorithm (UCA) and ISO/IEC 14651(International string ordering and comparison).

The multi weight sorting methodology which is well-understood and widely implemented uses a collation element table to achieve culturally acceptable sorting and enables searching at different degrees of precision (e.g., case-sensitive searches).

Its main advantages are as follows:

1. At the operating system level Dzongkha collation element table can "plug into" existing sort logic.
2. Therefore robust searching and sorting of Dzongkha data becomes automatically available to all complaint applications running within that operating system environment.
3. Across multiple platforms the same collation element table can be used – resulting in consistent Dzongkha data sorting within different operating system environment.

Under one of the 30 letters generally called Dzongkha alphabet, it was agreed that all entries must appear. The sorting of words from foreign language was according to the sort rules of the dictionary's language and not the sort rules of the origin language. For example, a Danish word beginning with å appears after letter Z in a Danish dictionary but the same word is sorted under letter A in an English dictionary. Because of this convention all words in a Dzongkha dictionary – including foreign words are sorted under 30 letters. Further extending this convention, all vowel signs are treated in terms of the 5 standard Dzongkha vowels

- implicit vowel ཨ་
- 4 explicit vowel signs

For international string ordering, the multi-weight sorting model assigns weights at three (or more) levels. In Latin scripts these levels correspond to:

- primary level = alphabetic ordering
- secondary level = diacritic ordering
- tertiary level = case ordering

For tie-breaking between strings additional levels may be used not distinguished at the first three levels.

An example in Dzongkha of extending multi-weight model is shown below.

- at the primary level ཏ and ཐ differ.
- at the primary level ཏ and ཏྟ differ.
- at the primary level ཏ and ཊ differ.
- at the primary level ཏ and ཊྟ differ.

The brief note on Unicode encoding model for Dzongkha script before I go to the specific Dzongkha collation rule is as follows.

Dzongkha script in unicode is defined with 195 distinct characters. In the 30 letters known as Dzongkha alphabet encoding is done twice, in nominal position as well as in orthographic- subjoined position.

The fact it reflects is that the Dzongkha script is written from top to bottom as well as from left to right. An example of Dzongkha encoding is shown below.

What is collation element? For determining sort weights, it enables clustering of multiple unicode characters such that they can be treated as a single item. But a single character can also work as collation element. Their sort order relative to one another can be determined by the weights assigned to the collation elements.

There are 167 primary weighted collation elements along with 9 secondary weighted collation elements. All 26 letters have primary weight in English; therefore "at" sorts under "a" and "vat" under "v". Letter written before the radical letter in Dzongkha have always had less primary weight than that of the radical; thus

ཀན , སྐན , བཀན , བསྐན relatively sorts near each other, under letter ཀ.

All in all, letter 11 possible prescripts (pre – radicals) are there occuring before a radical as shown below:

- ག ད བ མ འ : 5 prefix letters
- ར ལ ས : 3 head letters
- བར བལ བས : 3 two letter sequences of བ prefix followed by one of the head letters

Dzongkha grammar defines a rule for specifying which radical letters can take which prescripts. For an example letter ཀ can take 7 possible prescripts as: shown below.

- དཀ བཀ རྐ ལྐ སྐ བརྐ བསྐ

Note has to be made that no radical letter can take all prescript forms while some letters take none at all.

Prescribed radical values are defined as collation element in Dzongkha collation rules and are assigned sort weights such that sorting is done in culturally accepted relative order.

Roughly we have 167 primary weighted Dzongkha letters in the collation rule both as a single letter as well as a collating element. The relative order for these 167 letters is shown below:

- 133 collation elements = 30 nominal letters and 103 multi-letter prescribed radical forms
- 4 explicit vowels
- in orthographic subscribed position = 30 post-radical letters
- 133 + 4 + 30 , total collation slots at the primary-weight level
- 167 total

The 9 letters without primary weight are shown below:

- 4 combining marks: ◌ྀ ◌ཱ ◌ི ◌ཿ
- 5 signs: ྈ ྉ ྆ ྊ ྇

These 9 letters combine to give secondary weighted collation element.

Among the remaining 122 Unicode Dzongkha characters

- from 122 of these, 59 have a primary weight,
- since 19 can be decomposed into simple elements, it need not be treated in the collation element table,
- of the 30 nominal letters, 9 are variants (primary and tertiary weighted) of certain,
- of the 4 explicit vowels, 3 are variants (primary and tertiary weighted) of certain,
- of the 30 subscribed letters, 8 are variants (primary and tertiary weighted) of certain,
- 20 are the digits and half-digits and

The 63 remaining characters are punctuation marks and other symbols having no impact on dictionary sort order and therefore have no primary, secondary or tertiary weight.

5. Results

Tibetan unicode ordered list of collation elements

Collation elements of primary weight

133 radical-initial sequences (also covers the suffix letters):

ཀ དཀ བཀ རྐ ལྐ སྐ བརྐ བསྐ ཁ མཁ འཁ ག དག བག མག འག རྒ ལྒ སྒ བརྒ བསྒ ང དང མང རྔ ལྔ སྔ བརྔ བསྔ ཅ གཅ བཅ ལྕ བལྕ ཆ མཆ འཆ ཇ མཇ འཇ རྗ ལྗ བརྗ ཉ གཉ མཉ རྙ སྙ བརྙ བསྙ ཏ གཏ བཏ རྟ ལྟ སྟ བརྟ བལྟ བསྟ ཐ མཐ འཐ ད གད བད མད འད རྡ ལྡ སྡ བརྡ བལྡ བསྡ ན གན མན རྣ སྣ བརྣ བསྣ པ དཔ ལྤ སྤ ཕ འཕ བ དབ འབ རྦ ལྦ སྦ མ དམ རྨ སྨ ཙ གཙ བཙ རྩ སྩ བརྩ བསྩ ཚ མཚ འཚ ཛ མཛ འཛ རྫ བརྫ ཝ ཞ གཞ བཞ ཟ གཟ བཟ འ ཡ གཡ ར བར (seen in བརླ) ལ ཤ གཤ བཤ ས གས བས ཧ ལྷ ཨ

Key:

30 nominal letters are in black. Note that in native orthographies, any of these 30 can serve as a bare radical, 10 of these can also appear in suffix position.

Relatively unambiguous cases of prefixed and/or superscribed radical letters are in Blue. This is to be noted that between native orthographies and transcriptions from foreign languages certain unavoidable ambiguities arises.

Ambiguous cases where a 3rd codepoint is required to distinguish the sequence as being a prefixed radical letter (as opposed to a root letter followed by a suffix) are in red. In order to distinguish a case of a prefixed radical letter from a case of a suffix letter followed by a secondary syllable that involves a vowel (i.e., ནི or སོ) requires a 4th code point in certain cases in Dzongkha.

Ambiguous case (in Dzongkha) where a 3rd (or possibly 4th) code point is required to distinguish the sequence as being a prefixed radical letter (as opposed to a suffix letter ད followed by a secondary syllable པ or པོ) are in magenta.

A.2. the 4 explicit vowels:

ི ེ ོ

ུ

A.3. the 30 post-radicals:

◌◌ྐ ◌◌ྑ ◌◌ྒ ◌◌ྔ ◌◌ྕ ◌◌ྖ ◌◌ྗ ◌◌ྙ ◌◌ྟ ◌◌ྠ ◌◌ྡ

◌◌ྣ ◌◌ྤ ◌◌ྥ ◌◌ྦ ◌◌ྨ ◌◌ྩ ◌◌ྪ ◌◌ྫ ◌◌ྭ ◌◌ྮ ◌◌ྯ

◌◌ཱ ◌◌ྱ ◌◌ྲ ◌◌ླ ◌◌ྴ ◌◌ྶ ◌◌ྷ ◌◌ྸ

Collation Elements of Secondary Weight (have no primary weight)

B.1. the 4 combining marks:

◌྄ ◌◌ཱ ◌༹ ◌ཿ

B.2. the 5 signs (used in transliteration):

ྈ ྉ ྊ ྋ ྌ

C. The 120 Remaining Unicode Tibetan Characters

73 of the 195 Dzongkha characters defined in unicode are listed above. This leaves 122 characters, from which 19 can be decomposed into simple elements and therefore need not be treated in the collation element table. Since these generally have no impact on dictionary sort order, there is no need to assign primary secondary or tertiary weights to the 61 characters that function as punctuation marks and other symbols. [Note that due to the fact that there is no canonical or compatibility decomposition specified for this character, "Syllable OM" at U+0F00 is here treated as an ornamental symbol rather than as having any lexical value.]

20 further characters are in account for the digits and half digits. As listed previously, the remaining 20 characters are variations (i.e., having both primary and tertiary weights) of certain of the 30 nominal letters, 4 vowels, and 30 subjoined post-initial letters.

9 Nominal letter variants:

ཊ ཋ ཌ ཎ ◌ཾ ◌ྂ ◌ྃ ཪ (fixed form) ཥ

3 Vowel variants:

◌ྀ ◌ཻ ◌ཽ

8 Subjoined letter variants:

◌◌ྚ ◌◌ྛ ◌◌ྜ ◌◌ྞ ◌◌ྺ ◌◌ྻ ◌◌ྼ ◌◌ྵ

6. Conclusion

The collation rules for Dzongkha have been successfully implemented in GNU C Locale and have also been submitted to Common repository in Unicode.

Open office also supports Dzongkha collation rules. The Dzongkha Collation rules can be downloaded from http://dzongkha.sourceforge.net

7. Reference:

[1] R. Chilton, Sorting Unicode Tibetan using a Multi Weight collation algorithm. In Proceedings of *Tenth Seminar of the International Association for Tibetan Studies*, Oxford University, September 2003, pg 6-12.