

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 is 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 language (\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 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:

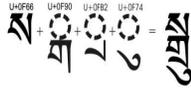
ཀ ཁ ག ན ལ ཌྷ ཎ ཏ ཐ ད དྷ ན པ བ མ ཙ ཚ ཛ ཛྷ ཞ ཟ འ

ཟ འ ཡ ར ལ ཤ ས ཧ ཉ

2.2. Vowels

It consists of four basic vowel signs as shown below

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 a 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, ལན, འཀན, འལན relatively sorts near each other, under letter ཀ.

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

- ཀ ལ འ ལ འ : 5 prefix letters
- འ ལ འ : 3 head letters
- འ འ ལ འ : 3 two letter sequences of འ 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.

- འཀ འལ ཀ ལ ལ འལ འལཀ

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: འ འ འ འ
- 5 signs: འ འ འ འ འ

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)**

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.