

Dzongkha collation rules

By Pema Geyleg <pgeyleg@dit.gov.bt>



Collation- An Over View

- general term for the process and function of determining the sorting order of strings of characters
 - crucial for the operation of databases, not only in sorting records but also in selecting sets of records with fields within given bounds
 - However, collation is not uniform; it varies according to language and culture
 - for any collation mechanisms to be accepted in the marketplace, algorithms that allow for good performance are crucial
-

Dzongkha Script

- The Dzongkha script also called Bhutanese script is used to write Dzongkha which is the national language of Bhutan
 - written using the letters of the same script system used for writing the Tibetan language (\x0F00 through \x0FCF)
 - Tibetan script is encoded using characters with values from U+0F00 to U+0FFF
 - 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 འ
ཡ ར ལ ཤ ཥ ས ཏ ཐ ད དྷ ན

- ***Vowels***

It consists of four basic vowel signs as show below

ི ཊ ཋ ཌ
i u e o

Table 1: UCS Tibetan Script Characters

	0F0	0F1	0F2	0F3	0F4	0F5	0F6	0F7	0F8	0F9	0FA	0FB	0FC
0													
1													
2													
3													
4													
5													
6													
7													
8													
9													
A													
B													
C													
D													
E													
F													

- Frequently used consonants combinations

They are used to write most of the common words in Dzongkha.

ཀ ཁ ར ལ ཅ ཆ ཇ ཉ ཊ ཋ ཌ ཌྷ

མ ཙ ཛ ཡ ར ལ ཥ ས ཏ ཨ

མ ཙ ཛ ཡ ར ལ ཥ ས ཏ ཨ སྐ སྑ སྒ སྒྷ སྔ སྕ སྖ

ཐ ད དྷ ན པ ཕ བ བྷ མ ཙ ཛ ཡ ར ལ ཥ ས ཏ ཨ

Why do we need a sorting algorithm for Unicode Dzongkha?

- Tibetan script is encoded in Unicode and ISO/IEC 10646 Standards
 - Full support of Dzongkha within a computer environment also requires:
 - Keyboard(s) or other input methods
 - Rendering: readable, printable display of the encoded Dzongkha script data.
 - Collation rules for generating culturally acceptable sorting.
-

Dzongkha- Collation History

- Earlier the Dzongkha sorting rules used single weight sorting model
 - It was generally adequate for sorting native Dzongkha orthographies within a specific application/environment
 - It treated Dzongkha-script sorting in an exclusive, special case fashion; such proprietary sorting methods were not widely implemented
 - present collation rule for Dzongkha script is compliant with the Unicode collation algorithm (UCA) and ISO/IEC 14651(International string ordering and comparison)
-

Features of multi-weight sorting weight sorting methodology

- Well-understood and widely implemented.
 - uses a **collation element table** to achieve culturally acceptable sorting.
 - enables searching at different degrees of precision.
-

Advantages for implementers and users of Unicode Dzongkha

- Collation element table for Dzongkha can “plug into” existing sort logic at the operating system level
 - Robust searching and sorting of Dzongkha data thus becomes automatically available to all compliant applications running within that operating system environment
 - The same collation element table can be used across multiple platforms – resulting in consistent sorting of Dzongkha data within different operating system environments
-

Sorting Unicode Dzongkha using multi weight collation Algorithm

- Words from foreign languages are sorted according to the sort rules of the dictionary's language (and not the sort rules of the origin language)
 - Extending this convention to Dzongkha, all words in a Dzongkha dictionary words including foreign words– are sorted under 30 letters
 - Extending this convention still further, all vowel signs are treated in terms of the 5 standard Dzongkha vowels
 - implicit vowel ཨ
 - 4 explicit vowel signs
-

The multi-weight sorting model for international string ordering

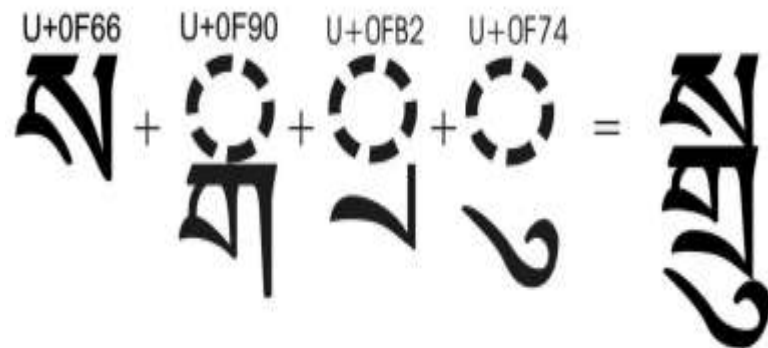
- Weights are generally assigned at three (or more) levels (or more) levels
 - In Latin scripts these levels correspond to:
 - alphabetic ordering = primary level
 - diacritic ordering = secondary level
 - case ordering = tertiary level
 - Additional levels may be used for tie-breaking between strings not distinguished at the first three levels
-

Example of extending multi-weight model to Dzongkha

- ཀ and ཀྱ differ at the primary level.
 - ཀ and ཀྲ differ at the secondary level.
 - ཀ and ཀླ differ at the tertiary level.
 - ཀ and ཀྴ differ at both the secondary level and the tertiary level.
-

Unicode encoding model for Dzongkha script

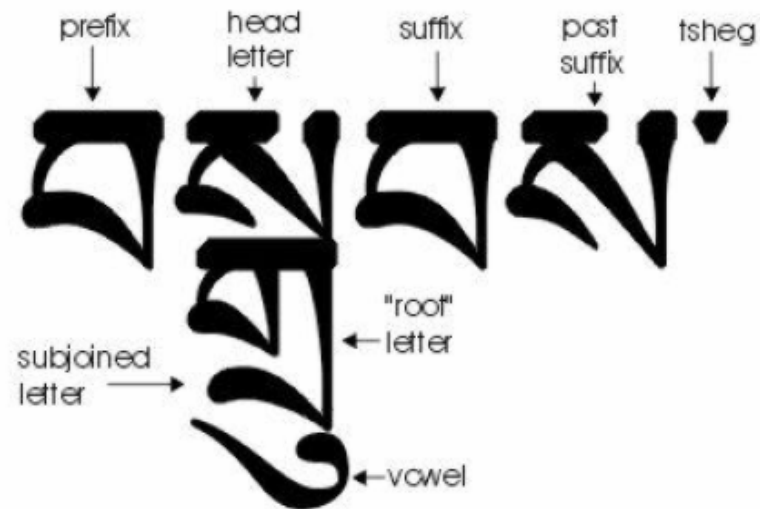
- Dzongkha script has 195 distinct characters defined in Unicode
- The 30 letters normally termed as Dzongkha alphabet are encoded twice in nominal position as well as in orthographic-subjoined position
- This reflects the fact that the Dzongkha script is written from top to bottom as well as from left to right



Syllables & Encoding

- The basic unit of meaning or morpheme in Dzongkha is the *tsheg bar* usually referred to as a “syllable”.
- Each syllable contains a root letter (*ming zhi*) and may additionally have any/or all of the following parts: prefix, head letter, sub-fixed letter, vowel sign, suffix, and post-suffix.
- Syllables are normally delimited by a *tsheg* or another punctuation character.

There are no inter-word spaces in Dzongkha



Determining collation elements for Unicode Dzongkha: an overview

- Prescripts in Tibetan orthographies
 - The Unicode model for encoding Tibetan script
 - What is collation element?
 - There are 167 primary weighted collation elements
 - 9 secondary weighted collation elements
-

Prescripts in Tibetan orthographies

- In English, all 26 letters have primary weight; thus “at” sorts under “a” and “vat” under “v”
- In Dzongkha words, there may be prefix letters written before the radical letter which always have less primary weight than that of the radical; thus

ཀའ་ , ཀླའ་ , བཀའ་ , བཀླའ་

sort relatively near to each other, under letter ཀ

- 11 possible prescripts (pre – radicals) occurring before a radical letter as shown below:
 - 5 prefix letters: ག ད བ མ འ
 - 3 head letters: ར ལ ས
 - 3 two letter sequences of བ prefix followed by one of the head letters: བར བལ བས
-

-
- grammar rules define which radical letters can take which prescripts

ྐ can take 7 possible prescripts as shown below:

དྐ བྐ ཏྐ ལྐ སྐ རྐ བསྐ

- that no radical letter can take all prescript forms while some letters take none at all
-

What is collation element?

- It enables clustering of multiple Unicode characters such that they can be treated as a single item for determining sort weights
 - a single character can also function as collation element
 - weights assigned to the collation elements determine their sort order relative to one another
-

Primary-weighted collation elements

- 30 nominal letters
 - 103 multi-letter prescribed radical forms
 - 4 explicit vowels.
 - 30 post-radical letters (i.e., in orthographic subscribed position)
 - total collation slots at the primary-weight level: $133 + 4 + 30 = 167$
-

Defining the 4 explicit vowels as collation elements

- As collation elements, suffix letters cannot be distinguished from bare radicals
 - Because a nominal letter serving as a radical letter carries the implicit vowel ^ʷ, the 4 explicit vowels must be given primary weights; and must be weighted heavier than the nominal letters- since a radical marked with an explicit vowel will sort after the same letter not marked by an explicit vowel
-

Defining the 30 post-radical letters as collation elements

- Post-radicals = the 30 letters in subjoined position (when not functioning as the radical letter in a pre-scribed radical form)
 - Requires maximum length substring matching
- Only 4 post-radical (subscribed) letters occur in native Dzongkha orthographies:



- Remaining 26 are required to treat non-native native orthographies in a consistent manner
 - Must be given primary weights; and heavier than the 4 explicit vowels
-

Relative order of the 167 primary-weighted collation elements

- First: 30 nominal letters and 103 multi-letter pre letter pre-scribed radical forms (= 133 collation elements)
 - given sort weights such that the 103 pre-scribed radical forms are interleaved as appropriate with the 30 nominal letters
 - Next: 4 explicit vowels
 - Next: 30 post-radical letters (i.e., in orthographic subscribed position)
 - Thus, total collation slots at the primary-weight level: $133 + 4 + 30 = 167$
-

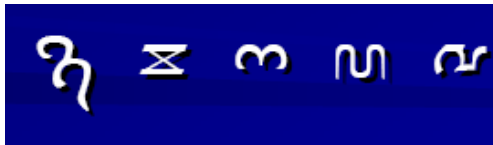
Secondary-weighted collation elements

- These 9 have no primary weight

- 4 combining marks



- 5 signs:



The remaining 120 Unicode Tibetan characters

- $30 + 4 + 30 + 9 = 73$ of the 193 Unicode Dzongkha characters have been treated above, leaving 120 characters
 - 59 of these 120 have a primary weight (in addition to a secondary and/or tertiary weight):
 - 19 can be decomposed into simple elements and thus need not be treated in the collation element table
 - 9 are variants (primary and tertiary weighted) of certain of the 30 nominal letters
 - 3 are variants (primary and tertiary weighted) of certain of the 4 explicit vowels
-

-
- 8 are variants (primary and tertiary weighted) of certain of the 30 subscribed letters
 - 20 are the digits and half-digits
 - The remaining 61 characters are punctuation marks and other symbols which generally have no impact on dictionary sort order and thus have no primary, secondary or tertiary weight
-