Sorting Unicode Tibetan using a Multi-Weight Collation Algorithm

Robert R. Chilton

Technical Director,
The Asian Classics Input Project (ACIP)
acip@well.com
www.asianclassics.org

Why do we need a sorting algorithm for Unicode Tibetan?

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

Previous efforts to sort Tibetan data

- Utilized a single-weight sorting model
- Generally adequate for sorting native Tibetan orthographies within a specific application/environment
- Not able to robustly handle foreign transcriptions and other non-standard orthographies

- Designed for Romanized or font-encoded Tibetan, not Unicode Tibetan
- Treat Tibetan-script sorting in an exclusive, special case fashion; such proprietary sorting methods are not likely to be widely implemented

Features of multi-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 (e.g., case-sensitive searches)

*References:

- ISO/IEC 14651 (2001-04) Ed. 1.0
  Information technology -- International string ordering and comparison -- Method for comparing character strings and description of the common template tailorable ordering.
- Unicode Technical Standard #10: Unicode Collation Algorithm (UCA).

Advantages for implementers and users of Unicode Tibetan

- Collation element table for Tibetan can “plug into” existing sort logic at the operating system level
- Robust searching and sorting of Tibetan 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 Tibetan data within different operating system environments

Moving from methodology to algorithm: an overview

- A look at dictionary sorting
- Understanding the multi-weight sorting (“international string ordering”) model and extending this model to Tibetan
- Determining the collation elements needed for sorting Unicode Tibetan

Dictionary sorting of Tibetan

- General agreement (since 1900) on dictionary order of native orthographies
  - Main exception: treatment of *wazur*
- Lack of consensus on details of sorting foreign-origin orthographies
- Universal agreement that all entries must appear under one of the 30 letters (ཀ་ to ཨ་)
  - Example of སྐན་ (Sanskrit: “skandha”): sorts under the collation slot for ས

How foreign words are sorted in dictionaries generally

- 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)
  - a Danish word beginning with **å** appears after letter **Z** in a Danish dictionary
  - the same word is sorted under letter **A** in an English dictionary

- Extending this convention to Tibetan, all words in a Tibetan dictionary – including foreign words – are sorted under 30 letters
- Extending this convention still further, all vowel signs are treated in terms of the 5 standard Tibetan 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
- In Latin scripts these levels correspond to:
  1. alphabetic ordering = primary level
  2. diacritic ordering = secondary level
  3. case ordering = tertiary level

  (Additional levels may be used for tie-breaking between strings not distinguished at the first three levels)

Examples in Latin script

- **role** and **rule** differ at the primary (alphabetic) level
- **role** and **rôle** differ at the secondary (diacritic) level
- **role** and **ROLE** differ at the tertiary (case) level
- **role** and **RÔLE** differ at both the secondary level and the tertiary level

Extending the multi-weight model to Tibetan

- ཏ་ and ཐ་ differ at the primary level
- ཏ་ and ད་ differ at the secondary level
- ཏ་ and ཏ་ differ at the tertiary level
- ཏ་ and ད་ differ at both the secondary level and the tertiary level

Determining collation elements for Unicode Tibetan: an overview

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

Prescripts in Tibetan orthographies

- In English, all 26 letters always have primary weight; thus “**at**” sorts far away from “**vat**”
- In Tibetan, letters written before the radical letter have less than primary weight; thus གན, སྣན, བགན, and བསྣན sort relatively near to each other, under letter ག

- 11 possible prescripts (or “pre-radicals”) might occur before the radical letter:
  - 5 prefix letters: ག ང བ མ འ
  - 3 head letters: ར ལ ས
  - 3 two-letter sequences of བ prefix followed by one of the head letters, i.e.: བར བལ བས

- Grammar rules define which radical letters can take which prescripts
  - For letter ག' there are 7 possible prescripts:

    དག བག རྐ ལྒ སྣ བཀ བསྣ

- No radical letter can take all 11 prescript forms (and some take none at all)

The Unicode encoding model for Tibetan script

- 193 distinct characters defined in Unicode
- The 30 letters (along with conjunct and reversed forms) are encoded twice: in nominal position and in orthographic-subjoined position
  - reflects the fact that the Tibetan script is written from top to bottom as well as from left to right
- Example of བརྒྱན་ encoded as 6 characters:

What is a collation element?

- A ***collation element*** enables clustering of multiple Unicode characters such that they can be treated together as a single item for determining sort weights
- Single characters also function as collation elements
- The weights assigned to the collation elements determine their sort (or collation) order relative to one another

Defining Tibetan pre-scribed radical sequences as collation elements

- For letter ཀ་ we can define each of the 7 prescript + radical clusters (དཀ བཀ ཤ ཧ སྐ བཤ བསྐ) as a collation element (also called a “collation grapheme”)
- We can then assign sort weights to these collation elements such that they sort in a culturally acceptable relative order

Primary-weighted collation elements

- 30 nominal letters: ཀ་ to ཨ་ – which may be either radical letters (མིང་གཞི་) or suffix letters
- 103 multi-letter pre-scribed radical forms
  - In many of these 103 forms the prescript is written at the head line (encoded as 1 or 2 nominal characters) and the radical letter is encoded as a subjoined character
  - In the example of བརྡན་, the radical letter ར is encoded in subjoined position

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 ཨ, the 4 explicit vowels must be given primary weights; and must be weighted heavier than the nominal letters -- since a radical letter marked with an explicit vowel will sort **after** the same letter not marked by an explicit vowel

* in a stateless implementation

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 Tibetan orthographies:

  ྱ ྲ ླ ྭ

- Remaining 26 are required to treat non-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-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 Tibetan characters have been treated above, leaving 120
- 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

Appendices

- The Unicode (and ISO/IEC 10646) character-encoding chart for Tibetan
  - highlighting characters in example: བཅོས་
- An ordered list of collation elements for Unicode Tibetan

An Ordered List of Collation Elements for Unicode Tibetan

[*Note: a comprehensive Collation Element Table for Tibetan script will include additional collation elements, such as ཱ, ཱ, ཱུ, ྲྀ, ཷ, beyond those listed here.]

A. Primary Weighted Collation Elements

A.1. The 133 radical-initial sequences (also covers the suffix letters):

ཀ ད་ཀ བ་ག ཀ་ ཁ་ ཁ་ བ་ཀ བ་ཁ ང་ མ་ང འ་ང ག དག་ བག་ མག་ འག་ ཅ་ ཆ་
ཚ་ བ་ཅ་ བ་ཚ་ ད ད་ད མ་ད ཇ་ ཛ་ ཛ་ བ་ཇ་ བ་ཛ་ ཉ ག་ཉ བ་ཉ ཏ་ བ་ཏ་ ཐ་ མ་ཐ
འ་ཐ ད ཨ་ད འ་ད ཏ་ ཐ་ བ་ཏ་ ན་ ད་ན ཨ་ན འ་ན ཏ་ ཐ་ བ་ཏ་ བ་ཐ་ ཀ ག་ཀ བ་ཀ ཀ་
ཁ་ ཁ་ བ་ཀ་ བ་ཁ་ བ་ཁ་ པ ཨ་པ འ་པ ད་ ག་ད་ བ་ད་ མ་ད་ འ་ད་ ཀ་ ཁ་ ཁ་ བ་ཀ་ བ་ཁ་
བ་ཁ་ ན་ ག་ན་ མ་ན་ ན་ ཁ་ བ་ན་ བ་ཁ་ བ་ ད་བ་ ཕ་ ཁ་ ཕ་ བ་ འ་བ་ བ་ ད་བ་ འ་བ་ ཏ་ ཐ་
ཁ་ མ་ ད་མ་ མ་ ཁ་ ཅ་ ག་ཅ་ བ་ཅ་ ཇ་ ཁ་ བ་ཇ་ བ་ཁ་ ཆ་ མ་ཆ་ འ་ཆ་ ད་ མ་ད་ འ་ད་
ད་ བ་ད་ ཝ་ ཞ་ ག་ཞ་ བ་ཞ་ ཟ་ ག་ཟ་ བ་ཟ་ འ་ ཡ་ ག་ཡ་ ར་ བ་ར(seen in བ་ན་) ལ་ ཤ་
ག་ཤ་ བ་ཤ་ ས་ ག་ས་ བ་ས་ ཧ་ ཧ་ ཨ་

Key:

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

Blue for (relatively) unambiguous cases of prefixed and/or superscribed radical letters. Note that certain unavoidable ambiguities arise between native orthographies and transcriptions from foreign languages.

Red for 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). Note that certain cases (in Dzongkha) require a 4th codepoint 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., ེ or ོ).

Magenta for an ambiguous case (in Dzongkha) where a 3rd (or possibly 4th) codepoint is required to distinguish the sequence as being a prefixed radical letter (as opposed to a suffix letter ད་ followed by a secondary syllable ་ or ོ).

A.2. The 4 explicit vowels:

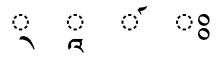
ི ུ ེ ོ

A.3. The 30 post-radicals:

ྐ ྑ ྒ ྔ ྕ ྖ ྗ ྙ ྚ ྛ ྜ ྞ ྟ ྠ ྡ ྡྷ
ྤ ྥ ྦ ྨ ྩ ྪ ྫ ྭ ྮ ྯ ྰ ྱ ྲ ླ

B. Secondary Weighted Collation Elements (have no primary weight)

B.1. The 4 combining marks:



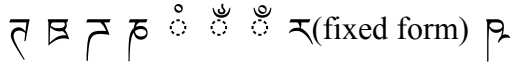
B.2. The 5 signs (used in transliteration):

C. The 120 Remaining Unicode Tibetan Characters

The characters listed above (in items A and B) account for 73 of the 193 Tibetan characters defined in Unicode. This leaves 120 characters, of which 19 can be decomposed into simple elements and thus need not be treated in the collation element table. There is also no need to assign primary secondary or tertiary weights to the 61 characters that function as punctuation marks and other symbols since these generally have no impact on dictionary sort order. [Note that "Syllable OM" at U+0F00 is here treated as an ornamental symbol rather than as having any lexical value due to the fact that there is no canonical or compatibility decompostion specified for this character.]

The digits and half digits account for 20 further characters. 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 listed previously.

9 Nominal letter variants:



3 Vowel variants:

8 Subjoined letter variants: