

Linguistic Sorting and Searching in Mimer SQL

1 Unicode Collation Algorithm (UCA)

<http://www.unicode.org/reports/tr10/>

2 Traditional Indic Collation

Attribute: [Indic]

Function: Traditional Indic collation method

The traditional Indic sort order is as follows:

1. Vowel
2. Vowelless consonant
3. Vowelless consonant + Vowel
4. Vowelless consonant + Vowelless consonant
5. Vowelless consonant + Vowelless consonant + Vowel
6. ... and so on

As the consonant letters in Indic scripts includes an inherent vowel /a/, the following transformations are applied before sorting:

1. Consonant + Virama \Rightarrow Vowelless consonant
2. Consonant + Vowel-sign \Rightarrow Vowelless consonant + Vowel
3. Consonant \Rightarrow Vowelless consonant + A

Transformation example:

क्	ka + virama	\Rightarrow क्	k
क	ka	\Rightarrow क्अ	k + a
कि	ka + i-sign	\Rightarrow क्इ	k + i
कु	ka + u-sign	\Rightarrow क्उ	k + u
के	ka + e-sign	\Rightarrow क्ए	k + e
को	ka + o-sign	\Rightarrow क्ओ	k + o
क्क	ka + virama + ka	\Rightarrow क्कअ	k + k + a

The method for traditional Indic collation effectively works for the following scripts:

- **Devanagari** (Hindi, Marathi, Nepali, and Sanskrit)

<http://developer.mimer.com/charts/hindi.htm>

<http://developer.mimer.com/charts/marathi.htm>

<http://developer.mimer.com/charts/nepali.htm>

<http://developer.mimer.com/charts/sanskrit.htm>

- **Bengali** (Assamese, and Bengali)
<http://developer.mimer.com/charts/assamese.htm>
<http://developer.mimer.com/charts/bengali.htm>
- **Gujarati**
<http://developer.mimer.com/charts/gujarati.htm>
- **Oriya**
<http://developer.mimer.com/charts/oriya.htm>
- **Telugu**
<http://developer.mimer.com/charts/telugu.htm>
- **Kannada**
<http://developer.mimer.com/charts/kannada.htm>
- **Malayalam**
<http://developer.mimer.com/charts/malayalam.htm>

The authoritative Monier-Williams: Sanskrit-English Dictionary is a good reference:

<http://www.ibiblio.org/sripedia/ebooks/mw/>
http://www.ibiblio.org/sripedia/ebooks/mw/0000/mw_0033.html

The traditional Indic collation method also works for **Tamil**, but with different rules as used in the authoritative University of Madras: Tamil Lexicon

<http://dsal.uchicago.edu/dictionaries/tamil-lex/>
<http://developer.mimer.com/charts/tamil.htm>

Punjabi does not need any tailoring, the default order follows the rules in the Punjabi University: Punjabi-English Dictionary ISBN:8173800960.

<http://developer.mimer.com/charts/punjabi.htm>

3 Syllable Collation

3.1 Lao

Attribute: [Lao]

Function: Lao syllabification and collation method

Proper Lao collation requires a syllable sort. Written Lao does not have any delimiters between syllables. A quite complicated regular expression has to be executed in order to determine the syllable boundaries. The syllabification algorithm is described in:

<http://www.panl10n.net/english/final%20reports/pdf%20files/Laos/LAO06.pdf>

Each syllable is sorted primarily on the leading consonants, vowel, and eventually finally consonant, and then secondarily on the tone.

<http://developer.mimer.com/charts/lao.htm>

3.2 Tibetan

Attribute: [Tibetan]

Function: Tibetan syllable collation method

Carefully chosen collation keys will enable advanced search operations.

<http://developer.mimer.com/charts/tibetan.htm>

3.3 Vietnamese

Attribute: [Vietnamese]

Function: Vietnamese syllable collation method

Example of sorted Vietnamese:

bo bít
bo siết
bò lan
bò tót
bỏ cha
bỏ sót
bõ công
bõ ghét
bó chân
bó tay
bọ da
bọ xít

bô bô
bộ xu
bồ chao
bồ quân
bồ bán
bồ vầy
bố cu
bố thí
bộ ba
bộ tịch

bơ bãi
bờ quai
bờ ngõ
bợ đất
bợ đỡ

<http://developer.mimer.com/charts/vietnamese.htm>

4 Chinese Collation

4.1 PinYin

Attribute: [PinYin]

Function: Chinese PinYin collation method

http://developer.mimer.com/charts/chinese_pinyin.htm

4.2 ZhuYin (Bopomofo)

Attribute: [ZhuYin]

Function: Chinese ZhuYin collation method

http://developer.mimer.com/charts/chinese_zhuyin.htm

4.3 WuBiHua (Five Stroke)

Attribute: [WuBiHua]

Function: Chinese WuBiHua collation method

http://developer.mimer.com/charts/chinese_wubihua.htm

5 Japanese Collation

Attribute: [Japanese]

Function: JIS X 4061-1996 collation rules for SOUND/ITERATION MARKS

<http://developer.mimer.com/charts/japanese.htm>

6 Korean Collation

Attribute: [Korean]

Function: Korean collation method

<http://developer.mimer.com/charts/korean.htm>