Linguistic Sorting and Searching in Mimer SQL

This is a summary of some language specific sort rules and collating details regarding linguistic sorting and searching in Mimer SQL. For further information, see the chapter called *Collations and Linguistic Sorting* in the Mimer SQL Reference Manual, http://download.mimer.com/pub/developer/docs/latest/MIMSQLEN.PDF.

1 Unicode Collation Algorithm (UCA)

The Default Unicode Collation Element Table (DUCET) is stated in the specification for the Unicode Collation Algorithm (UCA), http://www.unicode.org/reports/tr10/. This table provides a mapping from characters to collation elements.

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 => Vowelless consonant
2. Consonant + Vowel-sign => Vowelless consonant + Vowel
3. Consonant => Vowelless consonant + A

Transformation example:

क्	ka + virama	=>	क्	k
क	ka	=>	क्अ	k + a
कि	ka + i-sign	=>	क्इ	k + i
कु	ka + u-sign	=>	क्उ	k + u
के	ka + e-sign	=>	क्ए	k + e
को	ka + o-sign	=>	क्ओ	k + o
क्क्	ka + virama + ka	=>	क्क्अ	k + k + a

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

- **Devanagari** (Hindi, Marathi, Nepali, and Sanskrit)
  - http://download.mimer.com/pub/developer/charts/hindi.htm
  - http://download.mimer.com/pub/developer/charts/marathi.htm

    - http://download.mimer.com/pub/developer/charts/nepali.htm
    - http://download.mimer.com/pub/developer/charts/sanskrit.htm
- **Bengali** (Assamese, and Bengali)
    - http://download.mimer.com/pub/developer/charts/assamese.htm
    - http://download.mimer.com/pub/developer/charts/bengali.htm
- **Gujarati**
    - http://download.mimer.com/pub/developer/charts/gujarati.htm
- **Oriya**
    - http://download.mimer.com/pub/developer/charts/oriya.htm
- **Telugu**
    - http://download.mimer.com/pub/developer/charts/telugu.htm
- **Kannada**
    - http://download.mimer.com/pub/developer/charts/kannada.htm
- **Malayalam**
    - http://download.mimer.com/pub/developer/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://download.mimer.com/pub/developer/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://download.mimer.com/pub/developer/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://download.mimer.com/pub/developer/charts/LaoSyllableAlgorithm.pdf

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

http://download.mimer.com/pub/developer/charts/lao.htm

3.2 Tibetan

Attribute: [Tibetan]
Function: Tibetan syllable collation method

Carefully chosen collation keys will enable advanced search operations.

http://download.mimer.com/pub/developer/charts/tibetan.htm

3.3 Vietnamese

Attribute: [Vietnamese]
Function: Vietnamese syllable collation method

Example of sorted Vietnamese:

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

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

bơ bải
bờ quai
bỡ ngỡ
bợ đít
bợ đỡ

http://download.mimer.com/pub/developer/charts/vietnamese.htm

4 Chinese Collation

4.1 PinYin

Attribute: [PinYin]
Function: Chinese PinYin collation method

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

4.2 ZhuYin (Bopomofo)

Attribute: [ZhuYin]
Function: Chinese ZhuYin collation method

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

4.3 WuBiHua (Five Stroke)

Attribute: [WuBiHua]
Function: Chinese WuBiHua collation method

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

5 Japanese Collation

Attribute: [Japanese]
Function: JIS X 4061-1996 collation rules for SOUND/ITERATION MARKS

http://download.mimer.com/pub/developer/charts/japanese.htm

6 Korean Collation

Attribute: [Korean]
Function: Korean collation method

http://download.mimer.com/pub/developer/charts/korean.htm