1 The Ordering Principle

The widely accepted standard for Thai lexicographical ordering is defined in the Royal Institute Dictionary 2525 B.E. Edition (1982 A.D.), the official standard Thai dictionary. The ordering principles are:

1. Words are ordered alphabetically, not phonetically. Consonants order is:

ก ข ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท  ธ น บ ป ผ ฝ พ ภ ม ย ร ล รก ฤ ฦ ว ศ ษ ส ห ฬ อ ฮ

(ฤ ฤๅ ฦ ฦๅ are vowels and ligatures, but put in the order according to the sounds they represent.)

2. Vowels are also ordered by written forms, not by sounds. Vowels order is:

ะ จ ั า ำ ิ ี ึ ื ุ ู แ โ ใ ไ

Compound vowels are therefore ordered by their written forms like this:

-ะ -จ -ั -า ำ -ิ -ี ึ ื ุ ู แ โ ใ ไ

(อ ว ย are always ordered as consonants, although they sometimes act as vowels.)

3. Consonants always precede vowels. String comparison is performed from left to right, considering initial consonants before vowels in the same syllable.

4. Tones and diacritics are normally ignored, unless all other parts are equal, in which case the order is:

- - - -

Here is an ordering example:

ก กก กรรม กรรม์ กระแย่ง กราบ กัก ก้าว กำ กิน กี่ กึ๋น กุน เก้ง ก้าว เกี่ยว
2 Algorithmic Aspect

The above principle, with appropriate character code assignment such as TIS-620 and ISO/IEC 10646, almost allows C standard library function `strcmp()` to collate Thai strings without much more complication, except:

1. Leading vowels (เ- แ- ใ- ไ- ๅ- ), which are written before consonants, must be considered after the initial consonant. Therefore, the rearrangement is needed before comparison.

2. Diacritics and tone marks (ๅ- ำ- ง- ้- ๋- ) must be ignored in the first pass, and be considered at later pass if the first pass yields equality

And these are the only two mandatory requirements for Thai string collation algorithms. No syllable structure nor word boundary analysis is required, as Thai lexicons are ordered alphabetically, not phonetically.
2.1 Leading Vocal Rearrangement

To fulfill this requirement, either a preprocessing or collating-element grouping is required. The preprocessor scans the string once and swaps every leading vowel with its succeeding letter. The preprocessed string is then passed to the normal weight calculation process. Another way to manage this is by means of collating-element formation. Every possible pair of leading vowel and consonant is defined as a collating-element, whose weight equals to the weight of the rearranged substring.

Note that the rearrangement of a leading vowel is performed with its immediate succeeding consonant only. No consonant cluster analysis is needed. Instead, doing so would either face ambiguity problem or yield different order from the Royal Institute Dictionary.

For example, if consonant clusters were concerned, “เพลา” could be rearranged as either “พลลา” (two-syllable word พลลา) or “พล零” (one-syllable word, with consonant cluster “พล”).

To illustrate the different ordering caused by consonant cluster analysis, consider this conforming order: (เพล, เพลง, เพศ). If consonant cluster were analyzed, it would be rearranged to (พล, พลเง, พศ), and would yield a different order: (พลง, พล, พศ) (if the consonant cluster “พล” were not grouped as a single collating element); or (พล, พศ, พล) (if the consonant cluster “พล” were treated as a single collating element).

2.2 The Multiple Levels of Character Weights

The second requirement of diacritics and tone marks treatment implies multiple levels of weights. Tone marks and diacritics must be ignored in the first level, and weigh more than consonants and vowels in the second level.

There are ten Thai decimal digits (๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙), each semantically equivalent to Arabic digit 0-9, respectively. Their weights are then equal to their corresponding Arabic digits in the first level, and are different in the second level, to distinguish languages.

When punctuation marks (ๆ ฯ ๏ ๚ ๛) are concerned, another level of weights is required for them. This is correspondent to the fourth level in the Common Template Table. In string ordering, punctuation marks are less significant than any tone marks and diacritics, and must be ignored in all the first three levels.

For example, (ข้างๆ, ข้างกบ, ข้างๆ คู่ๆ, ข้างจัน) is a valid order in the Royal Institute Dictionary. In the first level, the considered weights are ขาง, ข้างกบ, ขางคู, ข้างจัน respectively.