B.4 Thai string ordering – a case involving special preparation

Thai 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:

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

ก ข ฃ 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 ฟ ภ ม ย ร ฤ ฤๅ ล ฦ ฦๅ 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.)

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

ะ ั า ำ ิ ี ึ ื ุ ู เ แ โ ใ ไ

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

Consonants always precede vowels. String comparison is performed from left to right, considering initial consonants before vowels in the same syllable.
Tones and diacritics are normally ignored, unless all other parts are equal, in which case the order is:

็ ่ ้ ๊ ๋

Here is an ordering example:

กก	เกล้า	ขนาบ
กรรม	เกลียว	ข้าง
กรรม์	เก้า	ข้างๆ
-กระแย่ง	เกาะ	ข้างกระดาน
กราบ	เกี่ยว	ข้างขึ้น
กะเกณฑ์	เกี๊ยะ	ข้างควาย
กัก	เกือก	ข้างๆ คูๆ
ก้าว	แกง	ข้างเงิน
กำ	แกะ	ข้างออก
กิน	โกน	เขน
กี่	โกร่น	เข็น
กึ่น	ไกล้	เข่น
กุน	ไก่	เข็ด
กูด	ไกล	แข็ง
เก้ง	ข้น	แข่ง

แข้ง
แข้งขวา
แข็งขัน
แข่งขัน
แขน
ครรภ-
ครรภ์
จุมพล
จุ๋พล
ชาย
เฒ่า
เณร
ตลาด
ทูลเกล้า
ทูลเกล้าฯ
ทูลเกล้าทูลกระหม่อม
น้ำ
น้ำ
นี้
บุญหลง
บุญ-หลง
ปา
ป่า
ป้า
ป๊า
ป๋า
ปาน
ผัด
ฯพณฯ
พณิชย์
ย่อง
รอง
ฤทธิ์
ฤษี
ฤาษี
ลลิตา
ฦๅชา
วก
ศาล
หริภุญชัย
หฤทัย
หลง
แห่ง
แห่ง
แหนม
แหนหวง
แหบ
แหม
อาน
ฮา

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:

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

Diacritics and tone marks (◌ั ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู ◌ฺ ◌็ ◌่ ◌้ ◌๊ ◌๋ ◌์ ◌ำ) 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 Vowel Rearrangement

To fullfill 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 that of the rearranged substring.
Note that the rearrangement of a leading vowel is simply performed with its immediate succeeding consonant. No consonant cluster analysis is needed. Indeed, doing so would result in ambiguities or yield a different order than that specified in the Royal Institute Dictionary. For example:

1. Ambiguities. The problem with ambiguity is illustrated by the word "เพลา". It has two potential pronunciations: either as a two-syllable word, "phe-la" (meaning "time"), or as a one-syllable word, "phlao" (meaning "axle" or "abate"). A rearrangement algorithm which follows the distinct pronunciation of the potential cluster 'พล' in this string would result in distinct keys, "พเลา" and "พลเา", and therefore different weights, which are equally legal. Both words need to have the same weight to be sortable, however.

2. Non-conforming Ordering. To illustrate the difference in ordering caused by the treatment of consonant clusters, consider these words, shown in conforming order: "เพล, เพลง, เพศ". The correct rearrangement ignores any clusters and results in the following: "พเล, พเลง, พเศ", which sorts in the order shown.
If, however, pairs of consonants that form legal clusters were grouped as single collation elements (regardless of actual pronunciation where the potential pronunciation is ambiguous), then the results of rearrangement would be "<พล>เ, <พล>เง, พเศ", which would yield the (non-conforming) ordering "เพศ, เพล, เพลง".
Again, if actual clusters were grouped as single collation elements (with some disambiguation effort), then the results of rearrangement would be
"พเล, <พล>เง, พเศ", which would yield the (non-conforming) ordering "เพล, เพศ, เพลง".

2.2 The Multiple Levels of Character Weights

The second requirement of the algorithm, relating to the treatment of diacritics and tone marks, 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 (ฯ ๆ ฿ ◉ ฺ ๛) are concerned, another level of weights is required for them. This corresponds 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 คูๆ, ข้างจัน" is a valid order in the Royal Institute Dictionary. In the first level, the considered weights are ขาง, ขางกบ, ขางคู, ขางจัน respectively.

The third level is not defined for Thai string ordering, but is reserved for tailoring.