Thai spelling reform for data processing

UDOM Warotamasikkhadit
Ramkhamhaeng University

1. Introduction

Those who know Thai know that it is impossible to tell whether a vowel *a*, *o*, *ɔɔ* or null is to be inserted after the first consonant when two consonants stand next to each other. Very often even a native Thai speaker, who knows very well the meaning of the words he reads, makes wrong syllabification cuts in words like ขนมอบกรอบ as ขน-มอ-บก-รอบ *khŏnmɔɔbòkrɔ̂ɔp*, ขน-มอบ-กรอบ *khônmɔ̂ɔpkrɔ̀ɔp*, ขน-มอบ-กร-อบ *khŏnmɔ̂ɔpkɔɔn?òp*, ขน-มอ-บก-รอบ *khŏnmɔɔbòkrɔ̀ɔp*, ข-นม-อบ-กร-อบ *khànŏm?òpkɔɔn?òp*, or ข-นม-อบ-กรอบ *khànŏm?òpkrɔ̀ɔp* where only the last reading is acceptable. It, thus, is very difficult for a foreigner or a machine to syllabify the written words in Thai correctly. This paper aims to present the problems of syllabification, romanization, and data processing of Thai words and to propose the methods of correct syllabification by man or machine.

2. Syllabification problems

2.1 It is difficult to tell whether the covert vowel is an *a*, *ɔɔ*, or *o* as in สรศักดิ์ *sɔ̆ɔrásàk* where the first covert vowel is *ɔɔ* and the second covert vowel is *a*, อมร *?àmɔɔn* where the first covert vowel is *a* and the second covert vowel is *ɔɔ*, ตลบ *tàlòp* where the first covert vowel is *a* and the second covert vowel is *o*.

2.2 A ์, a silence marker, usually silences a character it is placed over, but there are a number of Thai words where a silence marker silences up to three characters. It is almost impossible to tell which characters are to be silent, for example: วิทย์ *wít* where only one character, ย, is silenced, โพธิ์ *phoo* where two characters -- a consonant ธ and a vowel ิ -- are silenced, กาญจน์ *kaan* where two consonants, จ and น, are silenced, ลักษณ์ *lák* where two consonants, ษ and ณ, are silenced, and ลักษมณ์ *lák* where three consonants, ษ, ม and ณ, are silenced.

2.3 It is impossible to tell whether ร or ล is a syllable final or a cluster of an initial, for example: กรมนา can be read as *konmánaa* or *krommánaa*, กลมก can be read as

konlámóp, *klommaphá*, *kàlámóp*, or ***kàlomphá***. Those who know Thai will recognize that only ***krommánaa*** and ***kàlomphá*** are acceptable.

2.4 It is difficult to tell whether a character is a syllable final or an initial of the next syllable. การเวก can be read as *kaanwêek* or ***kaaráwêek***. Those who know Thai will recognize that the second reading is acceptable and *kaanwêek* does not exist in Thai.

2.5 It is hard to tell whether ว is an initial *w*, a second member of a cluster, a vowel *ua*, or a final *w*. Similarly it is hard to tell whether อ is an initial *ʔ* or a vowel *ɔɔ*. ขวนขวาย can be read as *khàwǒnkhàwǎay*, ***khwǒnkhwǎay***, or *khǔankhwǎay*. The correct pronunciation of this word according to the Dictionary of the Royal Institute is ***khwǒnkhwǎay***, but there are a number of people who pronounce it as *khǔankhwǎay*. คราวอด can be read as ***khraawʔòt*** or ***khraawôɔt***, depending on its meaning.

2.6 An initial cluster with ร, ล, or ว presents problems in reading and pronunciation whether ร, ล, or ว is a cluster with the preceding character or is a non-cluster, for example: เปรียญ *pàrian* and เปรียบ *prĭap*.

2.7 A number of Thai words contain a silent ร or ห, such as กำสรด *kamsòt*, เกียรติ *kĭat*, จริง *ciŋ*, ไซร้ *sáy*, ทรวง *suaŋ*, ศิรษะ *sĭisà*, พรหม *phrom*, พราหมณ์ *phraam*.

2.8 A certain number of word finals contain either a silent vowel *Ô*, or *Ø*, such as ชาติ *châat*, ประวัติ *pràwàt*, สมมติ *sŏmmút*, ธาตุ *thâat*, เหตุ *hèet*, but the two vowels are not pronounced.

2.9 Several words in Thai contain ์ in the middle of the word, such as ชอล์ก *chɔ́ɔk*, สาส์น *săan*, โอห์ม *ʔoom*.

3. The existing systems of syllabification

3.1 The use of ์ (yamakkan) over the first member of a consonant cluster and a final. This practice is widely adopted in Buddhist texts and it is still used in the *Book of Prayer* for Thai Buddhists. When a consonant-combination appears in Pali texts, a

yamakkan is placed on top of the first member of the combination[1], for example: ท์วาทะสะโยชะนา dwādasayōjanā.

A yamakkan or a silent marker, ์, is also used in non-religious texts to signify a final. This system can be found in the manuscript of King Borommakot's version of *Chindamani* where we doubt its genuineness and suspect that the manuscript was created in the early Bangkok period by a distinguished scholar with an intention to honor King Borommakot or to use his name as a pretext, for example: คน์ *khon*, คัน์ *khan*, คาน์ *khaan*, คิน์ *khin*, คีน์ *khiin*, คึน์ *khɨn*, คืน์ *khɨɨn*, คุน์ *khun*, คูน์ *khuun*, เคน์ *kheen*, แคน์ *khææn*, โคน์ *khoon*, คอน์ *khɔɔn*, ควน์ *khuan*, เคียน์ *khian*, เคือน์ *khɨan*, เคอน์ *khəən*, เคิน์ *khəən* คูวน์ *khuan*, คยน์ *khian*.

3.2 The use of ◌̣ a dot below, to replace ์, a yamakkan. It can be easily seen that the yamakkan takes up space above a character and is troublesome in writing. The use of a dot below seems to be neater in a certain aspect. This system is widely adopted in writing the Pali texts with Thai scripts up to present day, for example:

สมเณ พฺราหฺมเณ เจว	ราชวํสิกญาตโย
อมจฺเจ นาคเร จาปิ	สงฺคณฺหาติ ยถารหํ
samaṇe brāhmaṇe ceva	rājavamsikañātayo
amacce nāgare cāpi	sangaṇhāti yathāraham

3.3 Both systems are adopted into the Thai writing system, but they do not completely solve syllabification problems of Thai for romanization and data processing. Problem 2.4 above can be easily solved, but

3.3.1 It is impossible to predict whether บวฺ is to be pronounced *buan* or *bɔɔwɔɔn* (see 2.1 and 2.5).

3.3.2 It is impossible to tell whether a character with a yamakkan on top, or a dot below is a final or the first member of a cluster (see 2.3 above). พอกฺลาย can be read as *phɔɔ klaay* or *phɔ̂ɔklaay* depending on its meaning.

3.3.3 If we employ one of the existing systems, it is impossible to tell whether a character with a yamakkan above or a dot below in the following words is a final or

---

[1]It is believed that the lower part of a yamakkan covers the first member of a cluster and the top part of it covers the second member of a cluster.

the first member of a cluster, such as เกษตรศาสตร์ *kàsèttràsàat*, มาตรฐาน *mâattràthăan*, and นิทรา *nítthraa*.

4. The proposed systems for Thai spelling reform for data processing

4.1 The use of ◌์ to signify a cluster and ◌. to signify a final. If a yamakkan is placed on top of the first member of a cluster (in order to avoid interfering with a tone mark or a vowel written above the line such as ◌่, ◌้, ◌๊, ◌๋) and a dot below is placed under the final, Thai words will be more facilitated for data processing and pronunciation. This system will face difficulty in placing a dot below a consonant when a final already has a vowel ◌ุ, ◌ู written under the line.

4.1.1 The problems in 2.1 will be somewhat solved but not completely. Extra pronunciation rules must be devised such as:

4.1.1.1 If ก, ฆ, จ, ธ, น, ม, ว, ส, ห, or อ stands next to a single ร which is followed by another consonant which is not a final, *ɔɔ* is inserted after it; and if it stands next to a single ร which is followed by a vowel, *a* is inserted after it, for example: กรกฎ *kɔɔrákòt*, กรกฎาคม *kàrákkàdaakhom*, มรณี *khɔɔránii*, มราวาส *kháraawâat*, จรกา *cɔɔrákaa*, จราจร *càraacɔɔn*, ธรณี *thɔɔránii*, ธราธร *tháraathɔɔn*, นรเชษฐ์ *nɔɔrácheet*, นราธิป *náraathíp*, มรดก *mɔɔrádòk*, มรุ *márú*, วรจักร *wɔɔrácàk*, วรุณ *wárun*, สรสิทธิ์ *sɔ̌ɔ rásìt*, สรีระ *sàriirá*, หรดี *hɔ̌ɔrádii*, หริน *hàrin*, อรนุช *ʔɔɔránút*, อริน *ʔàrin*.

4.1.1.2 If บ precedes any consonant, except อ, which is followed by a vowel or a covert vowel, *ɔɔ* is inserted after it, for example: บรม *bɔɔrom*, บริบูรณ์ *bɔɔríbuun*, บดินทร์ *bɔɔdin*, บริบวรณ์ *bɔɔríbuan*[2].

4.1.2 Problem 2.2 and 2.9 will be solved if character deletion rules are provided (see 2.2 and 3.3.3). It must be noted that the character deletion rules must come earlier in rule ordering.

4.1.3 Problems 2.3, 2.4, 2.5, and 2.6 can be completely solved and 2.7 can be partly solved.

---

[2]It must be noted that บวร is pronounced *bɔɔwɔɔn*, but บริบวรณ์ is pronounced *bɔɔríbuan*. Even บวร appears in the latter part of บริบวรณ์, they are pronounced differently.