

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*, *๑๑* 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-บ-กรอบ *khǒnm๑๑bòkr๑p*, ขน-มอบ-กรอบ *khǒnm๑๑pkr๑p*, ขน-มอบ-กร-อบ *khǒnm๑๑pkr๑๑p*, ขน-มอ-บ-กรอบ *khǒnm๑๑bòkr๑p*, ข-นม-อบ-กร-อบ *khànm๑๑pkr๑๑p*, or ข-นม-อบ-กรอบ *khànm๑๑pkr๑p* 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*, *๑๑*, or *o* as in สรรค์ *sǎwrasàk* where the first covert vowel is *๑๑* and the second covert vowel is *a*, อมร *ʔàmm๑๑n* where the first covert vowel is *a* and the second covert vowel is *๑๑*, 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, โฟล์ *phoo* where two characters -- a consonant ฟ and a vowel อล์ -- are silenced, กาญจน์ *kaan* where two consonants, ก and ญ, are silenced, ลักขณ *lák* where two consonants, ล and ข, are silenced, and ลักขมณ *lák* where three consonants, ล, ข, ม and ณ, are silenced.

2.3 It is impossible to tell whether *ร* or *ด* is a syllable final or a cluster of an initial, for example: กรมฯ can be read as *konmánaa* or *krommánaa*, กรมฯ 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 *khàwǒnkhàwǎy*, *khwǒnkhwǎy*, or *khǔankhwǎy*. The correct pronunciation of this word according to the Dictionary of the Royal Institute is *khwǒnkhwǎy*, but there are a number of people who pronounce it as *khǔankhwǎy*. คราวอด can be read as *khraawʔòt* or *khraawɔɔt*, depending on its meaning.

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

2.7 A number of Thai words contain a silent ฏ or ฐ, such as กำสด *kamsòt*, เกียรติ *kiat*, จริง *cin*, ไซ้ *sáy*, ทรวง *suaŋ*, ศีรษะ *sǐisà*, พรหม *phrom*, พรหมณ์ *phraam*.

2.8 A certain number of word finals contain either a silent vowel *Ō*, or *Ø*, such as ขาติ *cháat*, ประวัติ *pràwàt*, สมนติ *sǒmmút*, 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 *chǎók*, สำน *sǎan*, โฉม *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 combination¹, for example: ท้าทะสะโยทะนา dwādasayōjanā.

A yamakkan or a silent marker, ๅ, 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*, คั่น *khan*, คาน์ *khaan*, คิน์ *khin*, คิั่น *khiin*, คิั่น *khin*, คิั่น *khiin*, คุั่น *khun*, คุั่น *khuun*, เค็น *kheen*, แคน์ *khææn*, โคน์ *khoon*, คอน์ *khoon*, ควน์ *khuan*, เคียน์ *khian*, เคือ่น *khiian*, เคอ่น *khææn*, เค็น *khææn* คุวน์ *khuan*, คียน์ *khian*.

3.2 The use of ๅ a dot below, to replace ๅ, 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:

สมณ พุราหมณ เจว	ราชวสิกญาดโย
อมจเจ นาคเร จาปี	สงคณหาติ ยถารห
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 บว is to be pronounced *buan* or *bwaawwa* (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). พกฺลย can be read as *phwa klaay* or *phōw klaay* 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 *mâattràthǎan*, and นิทรา *nítthraa*.

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

4.1 The use of ˆ 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, ข, จ, ฉ, ช, ฅ, ฆ, or ฮ stands next to a single ะ which is followed by another consonant which is not a final, ะ is inserted after it; and if it stands next to a single ะ which is followed by a vowel, ะ is inserted after it, for example: กรกฎาคม *kòorákòt*, กรกฎาคมคม *kàrákkàdaakhom*, ขรณี *khòoránii*, ขรวัส *kháraawâat*, จรกา *còorákaa*, จราจร *càraacoon*, ธรณี *thòoránii*, ธารธร *tháraathoon*, นรเชษฐ์ *nooráchêet*, นราธิป *náraathíp*, มรดก *mòorádòk*, มรุ *máru*, วรจักร *wòorácàk*, วรุณ *wárun*, สรัสสี *sàw rásìt*, สรีระ *sàriirá*, หรดี *hòorádii*, หริน *hàrin*, อรุณ *òoránút*, อริน *òàrin*.

4.1.1.2 If ๗ precedes any consonant, except ฮ, which is followed by a vowel or a covert vowel, ๗ is inserted after it, for example: บรม *bòorum*, บริบูรณ์ *bòoríbuun*, บดินทร์ *bòodin*, บริวารณ์ *bòoríbuan*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.

²It must be noted that บวร is pronounced *bòowpòon*, but บริวารณ์ is pronounced *bòoríbuan*. Even บวร appears in the latter part of บริวารณ์, they are pronounced differently.

4.1.3.1 An intruded vowel *a* rule must applied to กรัมนา *krommána*. The final must be repeated an *a* is inserted before a nasal provided that the preceding consonant is not the same as the one followed.

4.1.3.2 การเวก is pronounced *kaaráwéek* because *ɾ* does not have a dot below and it is not a final of the preceding syllable.

4.1.3.3 ข้วนวิชัย is pronounced *khwǒnkhwǎy* because *ɔ* functions as the second member of a cluster *ɔɔ*, not a vowel *ua*, and *ɯ* with a dot below functions as a final. Thus a covert vowel in the first syllable is *o*.

4.1.3.4 เปรี๊ยะ *pàrian* and เปรี๊ยะ *pri̇ap* are no longer a problem because *ɾ* in เปรี๊ยะ *pàrian* is not the second member of a cluster while *ɾ* in เปรี๊ยะ *pri̇ap* is.

4.1.3.5 เกียรติ, ฟ้าหม่อม and ฟ้าหม่อม are to be transcribed as *kĭat*, *phraam*, and *phrom* respectively by devising a rule deleting *ɾ* or *ɯ* in between an overt or a covert vowel and a final. Other words in 2.7 must be treated as exceptions.

4.1.3.6 A vowel *̇* or *̣* at word final is not pronounced, and must be deleted, such as เกียรติ *kĭat*, ขาติ *cháat*, ประวัติ *pràwàt*, สัมมติ *sǒmmút*, ธาตุ *tháat*, เหตุ *hèet*.

4.2 The use of *̣* a dot below under the first member of a cluster to signify a cluster and a space in front of a final. King Ramkhamhaeng who devised the Thai scripts had adopted a space in front of a final as part of the Thai writing system, for example: ศา สตราจา ทย *sâattraacaan*, มา ตรรฐา น *mâattràthǎan*, สุทธิสา ร *sùtthísǎan*. This system is not much different from 4.1 in its efficiency in syllabification, romanization and data processing of Thai words. Its advantages over 4.1 are that it uses only one special marker, signifying a cluster and words with a consonant with *̣* like สาส์ น, โฉน์ ม in the middle of the word will be no longer a problem of syllabification, language data processing and romanization (see 2.7). It eliminates an interference of a dot below and a vowel *̣*, such as ธา ตู *tháat*, มา ตู *máat*, เม รุ *meen*, เหตุ *hèet*. It is incredible that King Ramkhamhaeng's ingenuity concerning the use of a space in front of a final is very unique. His system reduces one marker less than the other system, and no one will miss a word final for its position after a space is obvious.

5. Conclusion

It is inevitable that exceptions exist in all languages and Thai is no different. The major problems of syllabification, romanization, and data processing of Thai will be solved, but not completely. There will be very few exceptions left to be treated as exceptions. The second proposed system is preferred to the first because only is needed in the system to signify a cluster the use of a space in front of a final devised by King Ramkhamhaeng will solve the problem of a word final containing a vowel above or below as in 1.8. It is realized that the second system proposed here may be hard to accept by the public because the use of a space in front of a final was discontinued almost 700 years ago, and people all over the world are accustomed to a space being used as a word boundary.

REFERENCES

- International Standard Organization. 1998. *ISO-11940-1 Information and documentation ---- Transliteration of Thai, Part 1*.
- International Standard Organization. forthcoming. *ISO-11940-2 Information and documentation---Simplified Transcription of Thai, Part 2*.
- Royal Institute, The. 1996. *Dictionary of the Royal Institute of B.E. 2525*. 6th Printing. Bangkok: The Royal Institute.
- Udom Warotamasikkkhadit and David Londe. 1961. "Computerized alphabetization of Thai." *Memoriam Phya Anuman Rajadhon*. Bangkok: The Siam Society.