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, ɔ 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 ขมนต์บดกกรรช kʰɔnmɔ̂ɔ bɔ̄dkɾɔ̂ɔp, ขมนต์บดกกรรช kʰɔnmɔ̂ɔ bɔ̄dkɾɔ̂ɔp, ขณ-มด-บก-กร kʰαnɔm tɔ̄p kɔ̄njɔp, or ขณ-มด-บก-กร kʰαnɔm tɔ̄p kɔ̄njɔp 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, ɔ, or o as in สรัสรัก sɔrɔrɔs ɔk where the first covert vowel is ɔ and the second covert vowel is a, ɔมข้ามɔɔɔn ɔɔɔn where the first covert vowel is a and the second covert vowel is ɔlosure ɔ่วยɔp where the first covert vowel is a and the second covert vowel is o.

2.2 À, 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: นกทุกสี wít where only one character, ɔ, is silenced, ɔ้รีหู phoo where two characters -- a consonant s and a vowel ɔ -- are silenced, ɔ้รีหู kaan where two consonants, s and ɔ, are silenced, ɔ้รีหู ɔ้าก ɔ้าก where two consonants, ɔ and n, are silenced, and ɔ้รีหู ɔ้าก ɔ้าก where three consonants, ɔ, s and n, are silenced.

2.3 It is impossible to tell whether ɔ or s is a syllable final or a cluster of an initial, for example: ɔ้รีหู can be read as konmɔ́nɔɔ or krommɔ́nɔɔ, ɔ้รีหู can be read as...
konlámóp, klommaphá, kàlámóp, or kàlomphá. Those who know Thai will recognize that only krommáanaa 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. การะ can be read as kaanwèek or kaarâwèek. Those who know Thai will recognize that the second reading is acceptable and kaanwèek does not exist in Thai.

2.5 It is hard to tell whether ฅ is an initial ฅ, a second member of a cluster, a vowel ဗ, or a final ဝ. Similarly it is hard to tell whether ฅ is an initial ฅ or a vowel ဝ. ชาก can be read as khâwônkhâwây, khwónkhwây, or khûankhwây. The correct pronunciation of this word according to the Dictionary of the Royal Institute is khwónkhwây, but there are a number of people who pronounce it as khûankhwây. ควิด can be read as khraawʔòt or khraawʔòt, depending on its meaning.

2.6 An initial cluster with ฅ, ฅ, or ฅ presents problems in reading and pronunciation whether ฅ, ฅ, or ฅ is a cluster with the preceding character or is a non-cluster, for example: เปี้ยน pàrian and เปี้ยน pìap.

2.7 A number of Thai words contain a silent ฅ or ฅ, such as กระดก kamsòt, เกี่ยวด kiat, จริง cinh, ชื่อ sây, ทรง suah, ศิลปะ sîisâ, พระ pam, พราน phram.

2.8 A certain number of word finals contain either a silent vowel ollapse, or ollapse, such as ชาติ châat, ประวัต prâwât, ส้ม sômmuît, ที่ thâat, เธ่ hêet, but the two vowels are not pronounced.

2.9 Several words in Thai contain ฅ in the middle of the word, such as ช่อ chöok, สำน sâan, โฉมʔoom.

3. The existing systems of syllabification

3.1 The use of ฅ (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
yamakan is placed on top of the first member of the combination\(^1\), for example: ท่าที่สะใภทว ด้วดัสภัยจัน.

A yamakan 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 Boromkhot’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 Boromkhot or to use his name as a pretext, for example: คุณ khon, คืน khan, คุณน้า khaan, คุณน้ำ khaan, คุณน้ำ khiin, คุณกัน khan, คุณน้ำ khuun, คุณเก็น keen, คุณเข็น keen, คุณน้า khaon, คุณน้า koon, คุณน้า khan, เคียน klian, เคียน kheen, เคียน kheen, เคียน kheen, คุณน้า khan, คุณน้า khian.

3.2 The use of ฯ, a dot below, to replace 八大以来, a yamakan. It can be easily seen that the yamakan 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:

<table>
<thead>
<tr>
<th>คำที่คืบ</th>
<th>คำที่คืบ</th>
</tr>
</thead>
<tbody>
<tr>
<td>สมมาเน พระแมว เจ้า</td>
<td>ราชสิทธิบุตรไก</td>
</tr>
<tr>
<td>อภิจิษฐานศรี จารี</td>
<td>สุวรรณธาตุ ยุพาราน</td>
</tr>
<tr>
<td>samaâne brâhmanâ ceva</td>
<td>râjavamsikanâtatayo</td>
</tr>
<tr>
<td>amacece nâgare câpi</td>
<td>sanganâhâti yathârahram</td>
</tr>
</tbody>
</table>

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 ป枸杞 is to be pronounced buan or ขวํวํคoon (see 2.1 and 2.5).

3.3.2 It is impossible to tell whether a character with a yamakan on top, or a dot below is a final or the first member of a cluster (see 2.3 above). พระคลาย can be read as phoo klaay or phôoklaay 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 yamakan above or a dot below in the following words is a final or

\(^1\)It is believed that the lower part of a yamakan 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 kāsēttrāsāt, มาตรฐาน māattrāthān, and นิทราว nīthraa.

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 ʰ, ʷ, ʰ, ʰ) 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 ḥ, ḫ 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 น, ภ, ผ, ฝ, ฤ, ษ, ห, or ฮ stands next to a single ṭ which is followed by another consonant which is not a final, ṭ was inserted after it; and if it stands next to a single ṭ which is followed by a vowel, a was inserted after it, for example: ต้น kārākōt, กระดำ kārakkādaakhom, ข้าม khārānīi, พระวัด khāraawāat, จาระ cōrrākāa, ราชจักร cāraacīo, ธรรม theoīrānīi, ธรรม thāraathōon, นางเช้า nooīrāchēet, นางปรี narāthīp, مصر mūrū, (JFrame) พอร์tāk, วัน wārun, สะทิม sōc rāsīt, ฟ้า sāriirā, ฟ้า hōorādīi, ฮาริ narīn, ฮาริ pōo rānūt, ฮาริ pārīn.

4.1.1.2 If ḫ precedes any consonant, except ḡ, which is followed by a vowel or a covert vowel, ṭ was inserted after it, for example: บน bōrom, บริบูรณ์ bōoribīo, บริบูรณ์ bōoribōo, บริบูรณ์ bōoribuan.

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 บว is pronounced bōorīo, but บริบูรณ์ is pronounced bōoribuan. Even บว appears in the latter part of บริบูรณ์, they are pronounced differently.
4.1.3.1 An intruded vowel a rule must applied to  krámmana 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 การบก is pronounced kaarâwêek because r does not have a dot below and it is not a final of the preceding syllable.

4.1.3.3 ขั้นชาย is pronounced khwôngkhwâay because r functions as the second member of a cluster entity, not a vowel ua, and u with a dot below functions as a final. Thus a covert vowel in the first syllable is o.

4.1.3.4 เบรียน pârian and เบรียน prâap are no longer a problem because r in เบรียน pârian is not the second member of a cluster while r in เบรียน prâap is.

4.1.3.5 เกียด, พรหมณ์ and พรมณ are to be transcribed as kïat, phraam, and phrom respectively by devising a rule deleting r or u 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 ə or ə, at word final is not pronounced, and must be deleted, such as เภี้ยด kïat, ชาต châat, ปราวât, สามม์ sömмút, ที่เท่ thâat, เน็ยหèet.

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: สัตตราวาจา sàattracaan, ม้าตระาจ mâattrâthâan, สุทธิสาน sûtthisâan. 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 หนุน, ไอ 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 ə, such as ที่เท่ thâat, มาต mâat, เม่น meen, เน็ยหèet. 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


