A Unified Model of Thai Romanization and Word Segmentation

Wirote AROONMANAKUN
Dept. of Linguistics
Chulalongkorn University
Bangkok 10330, Thailand
Wirote.A@chula.ac.th

Wanchai RIVEPIBOON
Dept. of Computer Engineering
Chulalongkorn University
Bangkok 10330, Thailand
Wanchai.R@chula.ac.th

Abstract

Thai romanization is the way to write Thai language using roman alphabets. It could be performed on the basis of orthographic form (transliteration) or pronunciation (transcription) or both. As a result, many systems of romanization are in use. The Royal Institute has established the standard by proposing the principle of romanization on the basis of transcription. To ensure the standard, a fully automatic Thai romanization system should be publicly made available. In this paper, we discuss the problems of Thai Romanization. We argue that automatic Thai romanization is difficult because the ambiguities of pronunciation are caused not only by the ambiguities of syllable segmentation, but also by the ambiguities of word segmentation. A model of automatic romanization then is designed and implemented on this ground. The problem of romanization and word segmentation are handled simultaneously. A syllable-segmented corpus and a corpus of word-pronunciation are used for training the system. The accuracy of the system is 94.44% for unseen names and 99.58% for general texts. When the training corpus includes some proper names, the accuracy of romanizing unseen names was increased from 94.44% to 97%. Our system performs well because it is designed to better suit the problem.

1 Introduction

The attempt to create a system of Romanization for Thai texts began since the 17th century by French missionaries (Griswold, 1960). However, the work was left unnoticed by other foreigners, who tend to romanize Thai words using their own languages notations. Issues of standardizing Thai romanization have been concerned again in the early 20th century (Frankfurt (1906), Petithuguenin (1912),Vajiravudh (1912,1931), Frankfurt et al. (1931)). Some systems of romanization are done on the basis of orthographic form, e.g. the system proposed by the ISO (ISO 11940 : 1998). Some systems are based on the pronunciation, such as the Royal Institute's system, Thiengburanathum's system (in his Thai-English dictionaries). Some are not totally based on orthographic form or pronunciation, e.g. the system proposed by King Rama VI.[1] This could explain why there are many ways to romanize a Thai word. For example, a common name like "เกรียงศักดิ์" can be romanized as "Kriangsak", "Kriengsakdi", "Kriengsak", or "Kreangsak". To lessen this problem, the Royal Institute has established the principle of romanization for Thai.[2] The principle has been endorsed by the United Nation (UN 2002) and slightly modified by the American Library Association and Library of Congress (ALA-LC 1997) for romanizing Thai scripts. Though the standard has been established, it is still not easy for general users to do romanization by hand. People tend to romanize Thai words on their own, rather than adhering to the principle. To help promoting the principle, Thatsanee et al. (1999) proposed an idea to develop an automatic romanization system. We agree with Thatsanee et al. that an automatic Thai romanization system is necessary for ensuring the standard. Such system was first developed in 1975 as a rule-based system (Londe et al. 1975). The accuracy was reported as greater than 95% when testing on Thai words. Unfortunately, the

[1] The system is known as a "graphic system" because it does not romanize words as pronounced in Thai. Words derived from Pali and Sanskrit words will be romanized to reflect the original words.

[2] The first version was proposed in 1939. The latest one is announced in 1999.

system runs on a mainframe and it is not available to the public. In addition, to be useful for the task, the automatic romanization system must be highly accurate, or near perfect. An accuracy of 70-90% on running text reported in many Thai text-to-speech systems is not adequate for this task. Therefore, we aim to develop a romanization system that is highly accurate, at least 99%.

2 Why Automatic Thai Romanization is Difficult?

Systems of Thai romanization that are totally based on the orthographic form like the transliteration system of ISO 1194 : 1998 is easy to be implemented because there is a one-to-one mapping from Thai to roman characters. But the system that is based on the pronunciation is more difficult because mapping from letters to sounds is not a one-to-one mapping. Letter-to-sound or grapheme-to-phoneme conversion systems are those used in text-to-speech applications. The difficulties on this task vary according to the characteristics of the language. The difficulties of transcribing Thai words are already discussed in many research papers, such as Luksaneeyanawin (1989), Meknavin and Kijsirikul. (2000), Khamya et al. (2000), Chotimongkol and Black (2000), Tarsaku et al. (2001), Tesprasit et al. (2003). In sum, Thai is an alphabetical language. There are 44 characters for 21 consonant sounds, 19 characters (including 3 consonant characters) for 24 vowel sounds (18 single vowels and 6 diphthongs), and 4 characters for tone markers (5 tones), and a number of characters for special symbols and numbers in Thai. Most of the following problems of Thai grapheme-to-phoneme conversion have been discussed in previous research. They are rearranged and clarified as follows:[3]

(a) Mapping of characters to sounds depends on the position of that character and its surrounding contexts. For example, the character "ค" is mapped to /kh/ when it is an initial consonant (e.g. "คา"-/**kh**aa0/-{remain}[4]), but it is mapped to /k/ if it is a final consonant (e.g. "นาค"-/naa**k**2/-{Naga}. The character "เ" usually maps to /e/, but if it is followed by a consonant and a vowel form "า", both vowel forms, "เ" and "า", will map to /aw/ (e.g. "เมา"-/m**aw**0/-{drunk}).

(b) It is possible that a vowel sound may not be represented by any character at all. For example, the syllable "กด"-/kot1/-{press} consists of only two characters for initial and final consonants. The vowel sound /o/ in this syllable does not have any corresponding character. In the syllable "สรร"-/san4/, even no vowel character is presented, the syllable is pronounced with a vowel /a/, as /san4/. But in "สร"-/sɔɔn4/, the vowel /ɔɔ/ is added.

(c) In some cases, it is ambiguous whether vowel forms belong to one syllable or two syllables. For example, in "เพลา", this string could be one syllable, "เพลา"-/phlaw0/-{axle}, in which "เ..า" represents the sound /aw/, or two syllables, "เพ-ลา"-/phee0-laa0/-{time}, in which "เ" and "า"represent different vowel sounds, /ee/ and /aa/ respectively.

(d) Since it is possible for two characters to represent one final consonant, it could be ambiguous whether the second character is a part of the final consonant, or it is the initial of the next syllable. For example, in "จักร", "ร" could be a part of the final consonants, "กร", as in "จักร-ยาน"-/cak1+ka1-jaan0/-{bicycle} or it could be the initial consonant of the following syllable, as in "จัก-รี"-/cak1-rii0/.

(e) Some characters map to different phonemes even they are in the same position (initial or final consonant). For example, character "ท", when used as an initial consonant, could map to either /d/ or /th/, as in "บัณ-ฑิต"-/ban0-**d**it1/-{graduate} and "มณ-ฑา"-/mon0-**th**aa0/-{Name of tree}.

(f) In some cases, one character can be both the final and initial consonants of two syllables. For example, the letter "ต" in "อัตรา"-{rate} represents the final consonant of the first syllable as well as the initial consonant of the following syllable, /?a**t**1-**t**raa0/.

---

[3] For each syllable, we will show only its orthographic form and its pronunciation. A gloss is not always provided because a syllable may or may not have a meaning in Thai. (A Thai word may composed of one or more syllables.) A gloss will be shown in { }.

[4] The numbers 0-4 at the end of syllable are used to represent five tones in Thai.

(g) There could be linking sounds between syllables in some words derived from Pali and Sanskrit. For example, in a compound word like "รัฐศาสตร์"-{political science} there is a linking syllable /tha1/ between the two words, "รัฐ"-/rat3/-{state} and "ศาสตร์"-/saat1/-{science}. This word is pronounced as /rat3+**tha1**-saat1/, rather than /rat3-saat1/.

(h) Even letters-to-sound conversion rules can be constructed, some syllables do not follow those rules. For example, the syllable "แม่น"-{precise} should be pronounced by rules with the long vowel as /m**ɛɛ**n2/, but it is actually pronounced with the short vowel as /m**ɛ**n2/. Though the vowel form "ไ" should map to a diphthong /**aj**/, but the syllable "ไห้" can be pronounced either as /h**aaj**2/ (e.g. "ร้อง-ไห้"-/rɔɔŋ3-h**aaj**2/-{cry}) or /h**aj**2/ (e.g. "เสา-ไห้"-/saw4-h**aj**2/-{Name of rice}).[5]

(i) In some cases, a cluster of initial consonants can map to different phonemes. For example, "ปล" can map to a cluster sound /pl/, e.g. "ปลา"-/**pl**aa0/-{fish}, or map to one leading syllable and an initial consonant sound /pa1-l/, e.g. "ปลาส"-/**pa1+l**at2/. The cluster "ทร" in a syllable can map to one phoneme /s/, e.g. "ทราบ"-/**s**aap2/-{know}; or map to a cluster sound /thr/, e.g. "ทรา"-/**thr**aa0/; or map to /tha3-r/, e.g. "ทรา"-/**tha3-r**aa0/, depending on the word it occurs.

3 Model of Automatic Romanization

Since the Royal Institute's standard of romanization is based on pronunciation, it would be better to design the system to transcribe Thai texts first. The output from this system then will be useful not only to the romanization system but also to a Thai text-to-speech system. Besides, transforming the transcription output into roman characters is quite straightforward.

The process of transcription, or grapheme-to-phoneme conversion is a basic research in any languages. Many systems of Thai grapheme-to-phoneme have been proposed. Some are rule-based, such as Londe et al. (1975), Khamya et al. (2000). Some are dictionary and rule-based, such as Luksaneeyanawin (1989). Some are statistical based, such as Chotimongkol and Black (2000), Tarsaku et al. (2001). Some apply a machine learning method, such as Meknavin and Kijsirikul. (2000), Tesprasit et al. (2003). For other languages, statistical methods are commonly used in transcriptions and transliteration tasks, such as Bosch and Daelemans (1993), Knight and Graehl. (1997), Al-Onaizan and Knight (2002).

To design a system of grapheme-to-phoneme for Thai, besides the awareness of difficulties listed above, we have to choose the level of analysis that is right for the problems. In other words, what will be problems and solutions of grapheme-to-phoneme conversion are directly related to the design of the system. For example, for a string "เพลา", if the conversion is treated as a one-step process of mapping from character to sound, there will be a problem of determining whether "เ" should map to /e/ or should it be combined with "า" and mapped to /aw/. But if the conversion is treated as a multi-step process, in which syllable segmentation is the first step, the problem would be to determine whether the string "เพลา" is one syllable or two syllables. Furthermore, when dealing with linking syllables, if the system is designed to work at the character level, these linking syllables have to be generated from a character in a specific context. But if the system is designed to work at the syllable level, these linking syllables can be generated from a specific syllable in a certain context.

In our system[6], we prefer not to view grapheme-to-phoneme conversion in Thai as a one-step mapping from characters to sounds. We think that the process of syllabification is necessary. The input character strings will be converted to a sequence of syllables. During this process, all possible pronunciations of each syllable are generated. Then, the correct pronunciation and word boundaries will

[5] Although the problem of vowel length does not affect the result of romanization since the Royal Institute's romanization system does not differentiate between short and long vowels, we would like to take this problem into consideration since the system developed here is also used for a Thai text-to-speech system.

[6] The system is online and can be downloaded from http://www.arts.chula.ac.th/~ling/tts/

be determined. We share the same view with Meknavin and Kijsirikul. (2000), and Tesprasit et al. (2003) that pronunciation disambiguation should be done simultaneously with word segmentation. We think that it is not always possible to resolve pronunciation ambiguity by doing only syllable segmentation, as Thatsanee et al. (1999) suggested. For example, if "สักวา" is one word, "สักวา"-{a kind of poem}, there would be a linking syllable generated, /sak1+**ka1**-waa0/. But if these are two words, "สัก"-{about} and "วา"-{unit of measurement}, it would be pronounced without a linking syllable as /sak1-waa0/. However, we do think that syllabification is generally useful for pronunciation disambiguation. Unlike Meknavin and Kijsirikul. (2000), and Tesprasit et al. (2003), who jumped directly to the word level, we think that many problems can be solved at the syllable level. And it is at this level that we could deal with problems mentioned in the last section more efficiently.

Problems (a), (b), (c), and (d) could be disregarded if the correct sequence of syllables is determined. When we know the syllable boundary, we would know how to map each character to its corresponding sound (problem (a)); we could add the missing vowel to a syllable that has only consonants[7] (problem (b)); and we will not have problems (c) and (d) at all.

For problems (e), (f), (g), and (h), we think that it should not be solved at the character level, as did in previous research. They are specific characteristics of some syllables. In problem (e), pronunciation ambiguity of "ฑ" could not be predicted at the character level. Actually, there are only a few syllables that "ฑ" should be mapped to /d/. Problems (f) and (g) are found on some loan words from Pali and Sanskrit. They are not a productive process. So, they should be handled as exceptions of some syllables rather than handled by rules. In (h), these are words that their pronunciations do not comply with letter-to-sound rules. So, they should also be treated as exceptions of some syllables. As for problem (i), the ambiguity of pronunciation would not be easily resolved by considering only nearby characters. But it is easier to predict how these clusters should be pronounced by considering surrounding syllables. For example, if "ทรา" occurs after "จัน", as in "จันทรา"-/can0-thraa0/-{moon}, it will be pronounced /thraa0/; but if it occurs in between "กัน" and "กร", as in "กันทรากร"-/kan0-tha3+raa0-kɔɔn0/-{mountain}, it will be pronounced /tha3+raa0/. Therefore, we believe that syllabification is a necessary step for grapheme-to-phoneme conversion.

In our model, unlike Kamya et al. (2000) who use rules for parsing syllables, syllabification is done on the basis of statistics. However, as stated before, we do not use statistical information at the character level as Chotimongkol and Black (2000), and Tarsaku et al. (2001) did, we use a trigram model of syllables to disambiguate syllable segmentations. The most probable syllable sequence is determined from the input characters. All possible pronunciations of each syllable are generated at this step too. After that, the right pronunciation of each syllable is chosen based on the result of word segmentation and the statistical information of pronunciation.

The model of transcription here can be viewed as a probabilistic model as below.

$$
\begin{aligned}
&\arg\max_{w\&phw} P(w_1..w_n \;\&\; phw_1..phw_n \mid c_1..c_m) \\
&= \arg\max_{w\&phw} P(c_1..c_m \mid w_1..w_n \;\&\; phw_1..phw_n) \\
&\quad * P(w_1..w_n \;\&\; phw_1..phw_n) / P(c_1..c_m) \\
&= \arg\max_{w\&phw} P(w_1..w_n \;\&\; phw_1..phw_n) \\
&= \arg\max_{w\&phw} P(w_1..w_n) * P(phw_1..phw_n \mid w_1..w_n) \\
&\approx \arg\max_{w\&phw} P(w_1..w_n) * \prod_{i=1..n} P(phw_i \mid w_i)
\end{aligned}
$$

[7] The missing vowel can be added correctly by considering consonants characters in the syllable.

$P(w_1 \ldots w_n$ & $phw_1 \ldots phw_n \mid c_1 \ldots c_m)$ is the probability that character strings $c_1..c_m$ will be word-segmented as $w_1..w_n$ and pronounced as $phw_1..phw_n$, such that phw_i is the pronunciation of w_i. The second line is equivalent by Bayes' rules. The third line is derived from the fact that $P(c_1..c_m)$ is a constant and $P(c_1..c_m \mid w_1..w_n$ & $phw_1..phw_n)$ is equal to one. Since we assume that the pronunciation of each word is not affected by other words, the probability of pronunciation of word sequence $w_1..w_n$ is estimated as the product of the pronunciation of each word. Thus, as seen in the last line, the model of Thai transcription is viewed as composing of two models: the language model and the pronunciation model.

3.1 Language Model

In this model, word sequences are produced from the sequence of input characters. Since syllabification is very useful for transcription, it will be included as a part of the model. We adopted Aroonmanakun's word segmentation model (2002) for this study. In his model, syllable segmentation is the first process. An input string is segmented into syllables by comparing to syllable patterns and syllable forms that are exceptions. For example, เCRTาะ, เXTอะ, เCRTY are syllable patterns in which X, C, R, Y, T stands for a different group of characters. Pronunciations of syllables matched to syllable patterns are generated by rules. Syllables that do not conformed to the syllable patterns or their pronunciations are different from those generated by rules, such as "ธาตุ"-/thaat2/-{elements}, "อัฐิ"-/?at1/-{ashes}, etc., are listed as exceptions. For example, "ธาตุ"-/thaat2/ is treated as an exception because it is uncommon to have the vowel form "ุ" under the final consonant. There are 220 syllable patterns and 1,935 exceptional syllables used in our system. The results after applying these syllable patterns and exceptions are usually ambiguous. For example, "ประโยคธรรมดา"-{simple sentence} could be syllable-segmented in three ways, namely "ประ-โยค-ธรร-มดา"-/pra1-jook1-than0-ma3+daa0/, "ประ-โยค-ธรรม-ดา"-/pra1-jook1-tham0+ma3-daa0/, "ประ-โยค-ธร-รม-ดา"-/pra1-jook1-thɔɔn0-rom0-daa0/. The most probable syllable segmentation is selected by the use of a trigram model. In this study, a training corpus of 638,277 syllables from newspapers is manually segmented. Witten-Bell discounting is used for smoothing (Chen and Goodman, 1998). Viterbi algorithm is used for determining the best segmentation. The selected sequence of syllables then will be grouped into words. This process is also non-deterministic. There could be many ways to group a given syllable sequence into a word sequence. We adopted Aroonmanakun's maximum collocation approach to select the best word sequence (Aroonmanakun 2002). Collocation strength of a word sequence is the sum of all words' collocation strengths in the sequence (Fw_i). Collocation strength of a word is the sum of collocation strengths between syllables in that word.

$$St = \sum_{i=1}^{n} F_{w_i} \qquad F_{w_i} = \sum_{j=1}^{k-1} C_{s_j, s_{j+1}} \; such\ that\ w_i = s_1 s_2 \ldots s_k \; and \; s_j \; is\ a\ syllable$$

Collocation strength between syllables is the ratio of p(x,y) to q(x,y), where p(x,y) is the probability of finding syllables x and y together, and q(x,y) is the probability of finding any syllable in between x and y (x-ANY-y), or the probability for x and y to be separated by any syllable. The collocation between syllables x-y then is calculated as below:

$$\log \frac{p(x,y)}{q(x,y)} = \log \frac{p(x)p(y \mid x)}{q(x)q(y \mid x)} = \log \frac{p(y \mid x)}{q(y \mid x)} = \log \frac{Count(x,y)/Count(x)}{Count(x,Any,Y)/Count(x)} = \log \frac{Count(x,y)}{Count(x,Any,y)}$$

The output from this model will be the sequence of words, in which each syllable in a word is attached with all possible pronunciations.

3.2 Pronunciation Model

In this model, we assume that pronunciation of a syllable could be determined within the word. Surrounding words do not affect the pronunciation of the target word. In addition, it is assumed that the pronunciation of a syllable is affected only by the preceding and the following syllable forms. The pronunciation then can be estimated as follows:

$$P(phw_i \mid w_i) = P(phs_1..phs_k \mid s_1..s_k)$$
$$\approx \prod_{j=1..k} P(phs_j \mid s_{j-1}s_j s_{j+1})$$

$P(phw_i \mid w_i)$ is the probability that a given word, w_i, will be pronounced phw_i. If the word is composed of syllables $s_1..s_k$, the pronunciation phw_i will be $phs_1..phs_k$, where phs_j is the pronunciation of syllable s_j. $P(phs_j \mid s_{j-1}s_j s_{j+1})$ is the probability that syllable s_j will be pronounced phs_j when the preceding and following syllables are s_{j-1} and s_{j+1}. The probability is estimated from a corpus of word-pronunciation. It is a list of 28,620 words manually aligned between syllable and its transcription. These words are extracted from the Royal Institute dictioncary.

3.3 Romanizing the Transcription

At the final process, the transcription output from the system is adjusted to comply with the Royal Institute's guideline of romanization. Each phonetic sound is replaced with the corresponding roman characters, such as /ŋ/ is changed to "ng", /ɛ/ is changed to "ae", etc. Tone and vowel lengths are deleted. Hyphen is added between two syllables if the final character of the preceding syllable and the initial character of the following syllable may cause reading ambiguity. For example, "samang" is changed to "sam-ang"; "saat" is changed to "sa-at". Space is inserted as word separation, such as "prasop khwam samret" (meet – Nom. Marker – success). The first character of the word is capitalized when romanizing proper names, such as "Dekying Umbun Thongmi" (Girl – First Name – Last Name).

4 Experiments

The system was tested on two data sets: general texts and geographical names. A running text of 18,388 words extracted from a newspaper is used for the first test. The purpose of the first test is to test the accuracy of romanization for general texts. The second test set is the list of 990 Thai geographical names extracted from the Royal Institute's books. Since romanization is generally used for writing Thai geographical names, the purpose of the second test is to check the accuracy of the system for romanizing geographical names. The results are shown in Table 1.

	General texts	Names
Correct	18311 (99.58%)	935 (94.44%)
Incorrect	77 (0.42%)	55 (5.56%)
	18388 (100%)	990 (100%)

Table 1 : Result of romanization

From the results, we can see that the system is highly accurate for general texts, but not for geographical names. For 77 errors found in general texts, 8 errors are caused from abbreviations and 64 errors are caused from English transliterated words, such as "แชมป์"-(champ), "เทรนเนอร์"-(trainer). These words are not supposed to be romanized. They should be processed by a backward transliteration system to its original form in English. If we exclude these 72 instances of errors, the system can romanize Thai texts correctly at the level of 99.97% (18311/18316). Therefore, there are only 5 errors of romanization, caused from the following words: "ดุกรุ่น"-/ku3-krun1/-{glowing}, "ปอดบวม"-/pɔɔt1-buuam0/-{puemonia}, "พิเชษฐ"- /phi3-chet2/-{Name}, "พิพัฒน์พงศ์"-/phi3-phat3-phoŋ0/-{Name}, "อินทรัตน์"-/?in0+

tha3-rat3/-{Name}. Errors of the first two words are caused from errors in syllabification. The word "คุกรุ่น"-{glowing} was incorrectly syllable-segmented as "คุก-รุ่น", rather than "คุ-กรุ่น". The word "ปอดบวม"-{puemonia} was wrongly segmented as "ปอ-ดบวม", rather than "ปอด-บวม".The last three errors are caused from Thai person names.

For 55 errors of geographical names, 4 errors are caused by incorrect syllable segmentation. These four names are segmented as "ขน-อม"-/khon4-?om0/, "บ้าน-เข-ว้า"-/baan2-khee4-waa4/, "ทุ่ง-เส-ลี่-ยม"-/thuŋ2-see4-lii2-jom0/, and "เข-วา-สิน-รินทร์"-/khee4-waa0-sin4-rin0/, while the correct ones should be "ขนอม"-/kha1+nɔɔm4/, "บ้าน-เขว้า"-/baan2-kha1+waw2/, "ทุ่ง-เสลี่ยม"-/thuŋ2-sa1+liiam1/, and "เขวา-สิน-รินทร์"-/khwaw4-sin4-rin0/ respectively. Eleven errors are caused by mispronunciation of seven syllables. For example, "สระ"-/sa1/ was mis-romanized as "sara" rather than "sa"; "แหง"-/hɛɛŋ4/ is mis-romanized as "ngae" rather than "haeng". The rest of errors (40), which are the majority of errors, are caused by incorrect word recognition. Failure to recognize word boundary results in the lack of linking syllables. For example, if the name "ราชเทวี"-/raat2+cha3-thee0-wii0/ is not recognized as one word, it will be incorrectly romanized as "rat thewi". But if it is recognized as one word, it should be romanized correctly with a linking syllable as "rat**cha**thewi". In addition, since many Thai names are composed of loan words from Pali and Sanskrit, their pronunciations are quite different from those of general words, which are used for training in the pronunciation model. Adding a linking syllable is quite common when pronouncing these names. Thus, it is not surprising why the accuracy is dropped to 94.44% when romanizing these geographical names.

To verify whether errors were mainly caused from the unsuitable training corpus, 3,000 person names were added to the training data of the pronunciation model, and these 990 geographical names were retested again. In addition, to handle the missing linking syllable which is caused from wrong word recognition, the system was instructed to treat each input name as one word in this test.

	Names
Correct	961 (97.07%)
Incorrect	29 (2.93%)
	990 (100%)

Table 2 : Result of romanization

From Table 2, we can see that adding person names in the training data could increase the accuracy of romanization. However, some errors still remain. Errors from incorrect syllable segmentation (4 errors) and mispronunciation (11 errors) are not solved because the added names in the training corpus do not have any new information to solve these problems. Many errors caused from the lack of linking syllables are solved at this test because the system is instructed to view each geographical name as one word. For example, "ราชเทวี"-/raat2+cha3-thee0-wii0/ is correctly romanized with a linking syllable as "ratchathewi". However, by treating each name as one word, the system fails to romanize some names because some geographical names may compose of more than one word. For example, "พนัสนิคม"-/pha3+nat3-ni3-khom0/ should be romanized without the linking syllable, "phanat nikhom", because this name is composed of two words "พนัส"-/pha3+nat3/-{forest} and "นิคม"-/ni3-khom0/-{settlement}. But the system mis-romanized it with a linking syllable as "phanat**sa**nikhom".

5 Discussion

The results show that the accuracy can be increased if appropriate training data is added. Since the current training data are mostly general texts, more training data on proper names should be added to the system. In addition, from the experiments, our system cannot romanize homographs correctly. For example, "สระ" can be romanized as either "sa" or "sara" depending on its meaning. Since we assume that pronunciation of syllables could be determined within the word, the system will not know how to disambiguate the pronunciations in these cases. Luckily homographs with different sounds are rare in

Thai. (There are only 8 words.) Most of them are hardly used in general texts. Only two words, "สระ"-{pond ; vowel} and "แหน"-{Name of plant ; be jealous}, need to be solved in the future development.

The results also show that the recognition of word boundary is very important for romanization. Incorrect word segmentation can cause an error in romanization. Thus, performance of romanization system relies heavily on the word segmentation algorithm. Unfortunately, none of Thai word segmentation algorithms yields perfect results. In addition, to romanize any texts according to the guideline of the Royal Institute, the system must be able to identify not only new words, but also proper names and their types. If the system is romanizing a person name, it should analyze that name as one word rather than composition of words. But if it is romanizing a geographical name, it should break down that name into words and insert a space for word separation. For example, if "แก่งหางแมว"-/kɛɛŋ1-haaŋ4-mɛɛw0/-{rapids-tail-cat} is a person name, it should be romanized as "Kaenghangmaeo". But if it is a geographical name, it should be romanized as "Kaeng Hang Maeo". Therefore, identification of new words and proper names should also be implemented in the next development. But at present, to make the system suitable as a public tool for romanization, the system must allow users to specify whether the input text should be treated as a single word or running texts, and allow users to add new words in a user dictionary. This would lessen the problems.

The results also indicate that the statistical model works very well for the transcription task. Our system which is based on a simple n-gram model trained with a moderate-size corpus can perform quite well. In our view, the accuracy depends on the system design rather than the statistical method. Our system is designed to work at the syllable level, which is more suitable to the Thai transcription problem than systems that work mainly at the character level. Therefore, any NLP systems should be carefully designed to suit the analysis of the linguistic problems. Statistical models are not by itself a magic box which can solve any linguistic problems. Linguistic analysis of the problems should be done first to understand the nature the problems. Moreover, we can see that error analysis is necessary for improving the system's performance. By locating where the problems are, we can prepare the training data that are suitable for the task. We can also see what could be improved and what would be the limitation of the current design.

Acknowledgements

We would like to thank the Thailand Research Fund and the Commission on Higher Education for funding this research. A Thai dictionary used in this system is mainly extracted from the Royal Institute dictionary, which is made available by the Thailand's Information Research and Development Division, National Electronics and Computer Technology Center.

References

ALA-LC 1997. *Romanization tables : transliteration schemes for non-roman scripts 1997*, approved by the Library of Congress and the American Libraty Association, Randall K. Barry (editor). Washington, D.C.: Library of Congress.

Al-Onaizan, Y. and K. Knight. 2002. Machine Transliteration of Names in Arabic Text. In *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages*, pages 34-46, Philadelphia, USA.

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. In *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*, pages 68-75, Pathumthani: Sirindhorn International Institute of Technology.

Bosch, A. van den and W. Daelemans. 1993. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 45-53, Utrecht, Netherland.

Chen, Stanley F., and Joshua Goodman. 1998. *An Empirical Study of Smoothing techniques for Language Modeling*. TR-10-98. Harvard University.

Chotimongkol, Ananlada and Alan W Black. Statistically trained orthographic to sound Models for Thai, In *Proceedings of ICSLP 2000*, Beijing, China October 2000.

Frankfurt, O. 1906. Some Suggestions for Romanizing Siamese. In *Journal of the Siam Society*, Vol.3, Part 2, 52-61.

Frankfurt, O., P. Petithuguenin, and J. Crosby. 1931. Proposed System for the Transliteration of Siamese Words into Roman Characters. In *Journal of the Siam Society*, Vol.10, Part 4, 1-22.

Griswold, A. B. 1960. Afterthoughts on the Romanization of Siamese. In *Journal of the Siam Society*, Vol.48, Part 1, 29-66.

ISO/FDIS 11940. 1998. *Information and documentation --- transliteration of Thai*, (ISO 11940 : 1998)

Khamya, A., L. Narupiyakul, and B. Sirinaovakul, 2000. SATTS : Syllable Analysis for Text-To-Speech System, In *The 4th Symposium on Natural Language Processing (SNLP 2000)*, pages 336-340, May 10-12, Chiang Mai Plaza Hotel, Chiang Mai.

Knight, Kevin and Jonathan Graehl. 1997. Machine Transliteration. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 128-135, Madrid, Spain.

Londe, David L., U. Warotamasikkhadit, and N. Kanchanawan. 1971. *TRACTS : Thai-Roman Computerized Transliteration System.* Research Report sponsored by the Advanced Research Projects Agency for the Thai-US Military Research and Development Center.

Luksaneeyanawin, Sudaporn. 1989. A Thai Text to Speech System. In *Proceeding of the Conference of the Regional Workshops on Computer Processing of Asian Languages*, pages 305-315, Asian Institute of Technology.

Meknavin, Surapant and Boonserm Kijsirikul. 2000. Thai Grapheme-to-Phoneme Conversion. In Burnham, Denis, et.al., editors, *Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing of Language and Speech*. NECTEC: Bangkok.

Petithuguenin, P. 1912. Method for Romanizing Siamese. In *Journal of the Siam Society*, Vol.9, Part 3, 1-12.

Royal Institute. 1999. *Principles of Romanization for Thai Script by Transcription Method*.

Tarsaku, Pongthai, Virach Sornlertlamvanich, and Rachod Thongpresirt. 2001. Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser. In *Proceedings of Eurospeech 2001*, Aalborg, Denmark, Sept 2001.

Tesprasit, Virongrong, Paisarn Charoenpornsawat, and Virach Sornlertlamvanich. 2003. A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis. In *Human Language Technology Conference 2003* (HLT-NAACL 2003), Edmonton, Canada, May 2003

Thiengburanathum, Wit, 1977. *Thai-English Dictionary.* Ruamsan Co.,Ltd., Bangkok: 2539.

United Nation. 2002. *Report on the current status of the United Nation romanization systems for geographical names*.

Vajiravudh, King. 1912. The Romanization of Siamese Words. In *Journal of the Siam Society*, Vol.9, Part 4, 1-10.

Vajiravudh, King. 1931. Notes on the Proposed System for the Transliteration of Siamese Words into Roman Characters. In *Journal of the Siam Society*, Vol.10, Part 4, 24-33.

Table of the Royal Institute's Thai Romanization

Consonant form	Romanized character		Vowel form	Romanized character
	Initial con.	Final con.		
ก	k	k	อะ, ั (reduced form of อะ) , รร (with final consonant), อา	a
ข ฃ ค ฅ ฆ	kh	k	รร (without final consonant)	an
ง	ng	ng	อำ	am
จ ฉ ช ฌ	ch	t	อิ, อี	i
ซ ทร (pronounced ซ) ศ ษ ส	s	t	อึ, อื	ue
ญ	y	n	อุ, อู	u
ฎ ฑ (pronounced ด) ด	d	t	เอะ, เ็- (reduced form of เอะ), เอ	e
ฏ ต	t	t	แอะ, แอ	ae
ฐ ฑ ฒ ถ ท ธ	th	t	โอะ, -(reduced form of โอะ), โอ , เอาะ, ออ	o
ณ น	n	n	เออะ, เ็- (reduced form of เออะ) , เออ	oe
บ	b	p	เอียะ, เอีย	ia
ป	p	p	เอือะ, เอือ	uea
ผ พ ภ	ph	p	อัวะ, อัว, -ว- (reduced form of อัว)	ua
ฝ ฟ	f	p	ใอ, ไอ, อัย, ไอย, อาย	ai
ม	m	m	เอา, อาว	ao
ย	y	–	อุย	ui
ร	r	n	โอย, ออย	oi
ล ฬ	l	n	เอย	oei
ว	w	–	เอือย	ueai
ห ฮ	h	–	อวย	uai
			อิว	io
			เอ็ว, เอว	eo
			แอ็ว, แอว	aeo
			เอียว	iao
			ฤ (pronounced รึ), ฤๅ	rue
			ฤ (pronounced ริ)	ri
			ฤ (pronounced เรอ)	roe
			ฦ, ฦๅ	lue