ICIC Express Letters
Volume **17**, Number **2**, February **2023**
 ISSN 1881-803X
pp. **131–141**

LINGUISTIC RULES-BASED APPROACH FOR TRANSLATING NYAW LANGUAGE TO THE PHONETIC ALPHABET

Thapani Hengsanankun, Atchara Namburi* and Surasak Tangsakul*

Faculty of Science and Engineering
Kasetsart University
59/5 Moo 1 Chiangkrua, Muang, Sakon Nakhon 47000, Thailand
thapranee.h@ku.th; *Corresponding authors: {adchara; surasak.tang}@ku.th

Received May 2022; accepted July 2022

Abstract. *This paper presents the conversion of the Isarn Thai phonetic alphabet for the Nyaw language using basic rules of linguistics in which the characteristics of input sentences are written in the standard Thai language. The proposed framework consists of 2 main processes: the first process uses the string matching method to search for words from the Nyaw language dictionary, and the second process applies the word segmentation method by comparing the longest word with a Thai dictionary. The first part results are divided into two cases, i.e., the case of the word found in the dictionary and the case of the word not found in the dictionary. For the first case, the words are identified to Nyaw phonetic alphabet. For the second case, the rest words are processed in the second process to re-segmentation using the longest matching algorithm with Thai dictionary. Finally, the word segmentation results will be converted into phonetic alphabets using the basic rules of the Isarn Thai language. To confirm the proposed method, we employed about 568 sentences in the experiments, and we found that the efficiency of the sentence conversion was up to 82.52% greater than the traditional method. As a result, the proposed method satisfies the requirements for building a speech synthesis system.*
Keywords: Nyaw language, Linguistic rule-based, Phonetic alphabet

1. **Introduction.** Phonetic analysis is one of the most challenging problems in the field of natural language processing (NLP). Several literature approaches, such as the Romanian language transcription system, applied the phonetic analysis in the text-to-speech processing component of the Romanian language for grapheme-to-phoneme converting [1]. The phonetic analysis converts lexical orthographic symbols to phonetic representations, along with the possibility of critical information such as stress positions [2,3]. Thai words to phonetic alphabet converting algorithm proposed by [2] applied the linguistic rules and Thai dictionaries, which stored data in a structured form. Then the longest matching word selection method was adopted to separate words and syllables. In addition, a study on transferring Thai written language to Roman characters proposed by [4] solved lexical ambiguity by determining the probability of using the n-gram in choosing the form, scope and syllable pronunciation. [5] presented a study of the development of phonetic alphabet writing by applying a data warehouse of the word segmentation called Tag Corpus Orchid to solve the challenge of reading many phonetic words in Thai.

Central Thai is the most common form of the Thai language. However, there are several Thai dialects, such as Isarn, Lanna Southern and Thai-Nyaw [6]. Although the central Thai language is the official language of Thais, people in localities still use traditional dialects in their daily communication. Communicating with different dialects might not be classified as entirely incomprehensible, but it is often confusing and misunderstood. Thai-Nyaw, also known as Nyaw, is one of the dialects of the Thai language, which is very similar to Laos's language, referred to Isarn Thai dialect. The writing of Nyaw initially has

DOI: 10.24507/icicel.17.02.131

a unique style and recently used Thai characters, but they are still different in phonetic systems. Writing the sentences in both Nyaw and Isan Thai language also used consonants and vowels of the Thai language. However, the Thai phonetic system cannot completely replace pronunciations of the Nyaw language. Therefore, the Isan Thai pronunciation system is used to improve the efficiency of the phonetic conversion system.

This paper proposes the conversion of the Isarn Thai phonetic alphabet for use with the Nyaw language by applying basic linguistic rules. The main contributions of this paper are summarized as follows.

1) We proposed a phonetic translation method that combines the well-known string matching and longest matching algorithms with the linguistic rules-based approach to cope with the known and unknown words of the Nyaw language.

2) The Nyaw sentences are one of the local data that are challenging to collect both online and offline, and a few sentences are not enough to evaluate the performance of the phonetic characters conversion method. Therefore, we made a number of sentences on several situations to evaluate the proposed method properly.

The remainder of the paper is organized as follows. In the next section, we review the related works on the phonetic alphabet of central Thai, Nyaw, and Isarn Thai language. Section 3 offers the proposed method. In Section 4, we provide details of the experiment and results. Finally, Section 5 provides the conclusion of the paper and suggestions for future work.

2. Related Work.

This section provides the details of Thai, Nyaw, and Isarn Thai languages consisting of consonants, vowels, tones, and their phonetic alphabet. These details are necessary for generating the final phonetic alphabet for the Nyaw language using the proposed method.

2.1. Thai language.

Central Thai is the official language of the Kingdom of Thailand for both spoken and written languages. For the written language, it consists of consonants, vowels, and tonal, which have a particular name and symbol for each. There are 44 consonants in the Thai alphabet, divided into three groups according to sound as low, middle, and high (also known as "Tri-yang") as follows:

High: there are 11 consonants which are ข ฃ ฉ ฐ ถ ผ ฝ ศ ษ ส ห

Middle: there are 9 consonants which are ก จ ฎ ฏ ด ต บ ป อ

Low: there are 24 consonants which are ค ฅ ฆ ง ช ซ ฌ ญ ฑ ฒ ณ ท ธ น พ ฟ ภ ม ย ร ล ว ฬ ฮ

Besides, there are 32 vowels in Thai, i.e., ◌ะ, ◌า, ◌ิ, ◌ี, ◌ึ, ◌ื, ◌ุ, เ◌ะ, เ◌, แ◌ะ, แ◌, เ◌อะ, ไ◌ะ, ใ◌, เ◌าะ, ออ, เ◌ียะ, เ◌ีย, เ◌ียะ, เ◌ีย, ◌ัวะ, ◌ัว, 21 consonant units in which are divided into three groups as nine short sound vowel units, i.e., ◌ะ, ◌ั, ◌ำ, ◌ิ, ◌ึ, ◌ุ, เ◌ะ, แ◌ะ, ไ◌ะ, เ◌าะ, เ◌อะ and nine long sound vowel units, i.e., ◌า, ◌ี, ◌ื, ◌ู, เ◌, แ◌, ใ◌, ◌อ, เ◌อ and three compound vowel units, i.e., เ◌ีย, เ◌ือ, ◌ัว [6]. For the tone marks, there are four forms, i.e., ◌่, ◌้, ◌๊, ◌๋ which consist of five tonal sound units, i.e., ordinary sound, masterpiece sound, master sound, triple sound, and fourth sound (see Table 3).

Thai phonetic alphabet: The phonetic alphabet refers to the letters and symbols used to represent different types of sounds, e.g., symbols for vowels, consonants, tones and other special phonetic marks. Therefore, each language has its own phonetic alphabet that can represent the sound of that language. Those phonetics are represented by the International Phonetic Alphabet (IPA) [2,4,7-9], an alphabetic system of phonetic notation based primarily on the Latin script, as shown in the following table. Table 1 shows the division of consonants into 21 phonemes and their phonetic alphabet. As shown, the number of 21 consonant sounds and 21 vowel phonemes, consists of phonemes at the beginning of the syllable and phonemes at the middle of all syllables, but there are only nine consonant phonemes at the end of the syllable, as shown in Table 2, whereas five phonetic alphabets of the four forms of the tone marks are shown in Table 3.

TABLE 1. Initial consonant unit, and single and compound vowel unit

Initial consonant unit		Single and compound vowel unit	
Consonant unit	Phonetic alphabet	Thai vowel letters	Phonetic alphabet
ก	/k/	_ะ, ั, ำ	/a/
ข ฃ ค ฆ	/kh/	ิ	/i/
ง	/ŋ/	ึ	/ɯ/
จ	/c/	ุ	/u/
ช ฉ ฌ	/ch/	เ_ะ	/e/
ซ ส ศ ษ	/s/	แ_ะ	/ɛ/
ญ ย	/y/	โ_ะ	/o/
ฎ ด	/d/	เ_าะ	/ɔ/
ฏ ต	/t/	เ_อะ	/ə/
ฐ ฑ ฒ ถ ท ธ	/th/	_า	/a:/
น ณ	/n/	ี	/i:/
บ	/b/	ื	/ɯ:/
ป	/p/	ู	/u:/
ผ พ ภ	/ph/	เ	/e:/
ฝ ฟ	/f/	แ	/ɛ:/
ม	/m/	โ	/o:/
ร	/r/	อ	/ɔ:/
ล ฬ	/l/	เ_อ	/ə:/
ว	/w/	เ-ีย	/i:a/
อ	/?/	เ-ือ	/ɯ:a/
ห ฮ	/h/	-ัว	/u:a/

TABLE 2. Final consonant unit

Sound unit	Consonant unit	Consonant letters	Phonetic alphabet
1	/ก/	ก ข ค ฆ	-/k/
2	/ง/	ง	-/ŋ/
3	/ย/	ย ไอ ใอ	-/y/
4	/ด/	ด จ ช ซ ศ ษ ส ฎ ฏ ฑ ท ด ฐ ฒ ถ ท ธ	-/t/
5	/น/	น ญ ณ ล ร ฬ	-/n/
6	/ป/	บ ป พ ฟ ภ ผ ฝ	-/p/
7	/ม/	ม อำ	-/m/
8	/ว/	ว เอา	-/w/
9	/อ/	*No final consonant for short vowel.*	-/?/

2.2. **Nyaw language.** Nyaw ethnic group are originally lived in large numbers in Hongsa, and Khammuan, a territory in Laos, and then they are immigrated to settle in Thailand [10]. In the areas of Northeast and the East of Thailand it was found that the Nyaw ethnic groups moved to settle in many provinces, e.g., Nong Khai, Udon Thani, Maha Sarakham, Sakon Nakhon, Nakhon Phanom, Mukdahan, Khon Kaen, Nakhon Sawan, Saraburi, Prachin Buri and Sa Kaeo.

TABLE 3. Tones of Thai language

Tonal sound units		
Tones	Tone marks	Phonetic alphabet
Ordinary sound	-	-
Masterpiece sound	◌่	/ˋ/
Master sound	◌้	/ˆ/
Triple sound	◌๊	/ˊ/
Fourth sound	◌๋	/ˇ/

In addition, it also found that the Nyaw ethnic groups were living in Cambodia, such as Banteay Meechai and Udon Meechai. Regarding their written language used in the past, they used the Dharma character script or Thai Noi characters like Thai Isarn people. However, they are currently using Thai characters, and their language was declared on the National Intangible Cultural Heritage List in 2014.

The distinctive features of the Nyaw language are similar to the Isarn Thai language and Lao language in the dialect of Luang Prabang, which is considered one of the fascinations of the Nyaw language. In this regard, the Nyaw language consists of 19 consonants, 18 single vowels, three compound vowels, four tones, six diphthongs [11], e.g., a vowel "เออ" of a word that corresponds to the vowel "เ-"of standard Thai language, for example, "หัวเฮ้อ" = "หัวใจ" (heart), "เฮอ" = "ให้" (give), "ผู้เญ่อ" = "ผู้ใหญ่" (adult), "เตอ" = "ใต้" (under), "เกอ" ="ใกล้" (near), "เม้อ" ="ใบไม้" (leaf). Another distinctive feature is the question word, e.g., "เตอ" or "เบอะเตอ" = "อะไร" (what?), "นั่นเตอ" or "นั่นเบอะเตอ" = "นั่นอะไร" (what is that?), "เลอเล่อ" or "จั้งเลอ" = "อย่างไร" (how?), e.g., "เขาเว่อเลอเล่อ" = "เขาพูดอย่างไร", "เจาชิเว่าจั้งเลอ" = "คุณจะพูดอย่างไร" (how do you say?), "เมื่อเลอ" = "เมื่อใด/เมื่อไร" (when?), "กะเลอ" = "ที่ไหน/ไหน", e.g., "เจาชิไปกะเลอ" = "คุณจะไปไหน" (where are you going?). Moreover, the word in interrogative sentences about people, for instance, "พอ" or "เพอเลอ" = "ใคร" (who?), can be used at the beginning and the end of the sentence, e.g., "เพอมาหาข้อย" = "ใครมาหาฉัน" (who came to me?), "เพอเลอเอิ้นข้อย" = "ใครเรียกฉัน" (who called me?), "นั่นคือเพอ" = "นั่นคือใคร" (who is that?).

2.3. **Isarn Thai language.** Isarn Thai is a popular language spoken in various provinces in the Northeast of Thailand. The sound form of the language can be divided into consonant sounds, vowels sounds, and tones with tonal and vowel sounds which are different in some locals or some provinces. For the Isarn Thai language, there is no diphthong in which it is considered a distinct difference from standard Thai [6,12]. The differences in consonants, vowels and tones between standard Thai and Isarn Thai are shown in Tables 4, 5 and 6, respectively. An example of the differences in the sound of standard Thai and Isarn Thai are shown in Table 7.

TABLE 4. Consonant sounds of Thai and Isarn Thai language phonetic alphabet

Consonant sounds	
Thai Language Phonetic alphabet	Isarn Thai Language Phonetic alphabet
y	ñ
ch	s or c
r	l
l	m
kh	h

TABLE 5. Vowel sounds of Thai and Isarn Thai language phonetic alphabet

Vowel sounds	
Thai language phonetic alphabet	Isarn Thai language phonetic alphabet
u	o
ɯ	ə
ɯa	ia
ɯa	əː

TABLE 6. Tones of Isarn Thai language phonetic alphabet

Tones	Isarn Thai language phonetic alphabet
Masterpiece sound	/ˋ/
Master sound	/ˆ/
Triple sound	/ˊ/
Fourth sound	/ˇ/
Mids rise	/~/
Ordinary sound	-

TABLE 7. Example of the differences in the sound of standard Thai and Isarn Thai language

Thai language phonetic	Isarn Thai language phoneme
chá:ŋ	sâ:ŋ
chì:k	cì:k
ya:w	ña:w
raw	haw

Both Nyaw and Isan Thai languages employed Thai consonants and vowels in their phrases, but their phonetic systems are still different. However, the writing of Nyaw initially has a unique style, and the Thai phonetic system cannot completely replace all pronunciations of the Nyaw language. In this paper, the Isan Thai pronunciation system is introduced to improve the efficiency of the proposed phonetic conversion system.

2.4. **Dictionary.** The proposed method applied the dictionary of the National Electronics and Computer Technology Center [13] and also used the 500 words of Nyaw dictionary [14] which relied on the pronunciation system of the Isarn Thai to determine the Nyaw phonetic alphabet. The structure and example of the Nyaw dictionary are shown in Table 8.

3. **Proposed Method.** The Nyaw language still uses the Thai alphabet for writing, and most of their sentences are also short sentences or phrases. However, they are different in the pronunciation system in which a spoken of the Nyaw language in the locality is similar to the Isarn Thai language. Therefore, the type of sentences or phrases used in this research is a sentence that is only used in daily life.

TABLE 8. The structure and example of the Nyaw dictionary

Nyaw vocabulary	Phoneme
กก	kok
ขี้แก๊ม	khi: kɛ:m
เข่า	khàw
ไข่	khày
คึ๊ด	khɯ´t
ม่วน	mũan

The processing steps of the proposed method can be divided into two main processes as follows.

First step, we implement a string matching algorithm for determining the known or unknown words from the Nyaw dictionary. The known words are then directly translated to Nyaw phonetic alphabet from the dictionary.

Second step, only the unknown words got from the first step are segmented by the longest matching algorithm with a Thai dictionary.

Then, the Nyaw phonetic alphabets are achieved by the proposed Isarn Thai rules-based language translation. The main process of the proposed method is shown in Figure 1 and details of all processes are provided as follows.

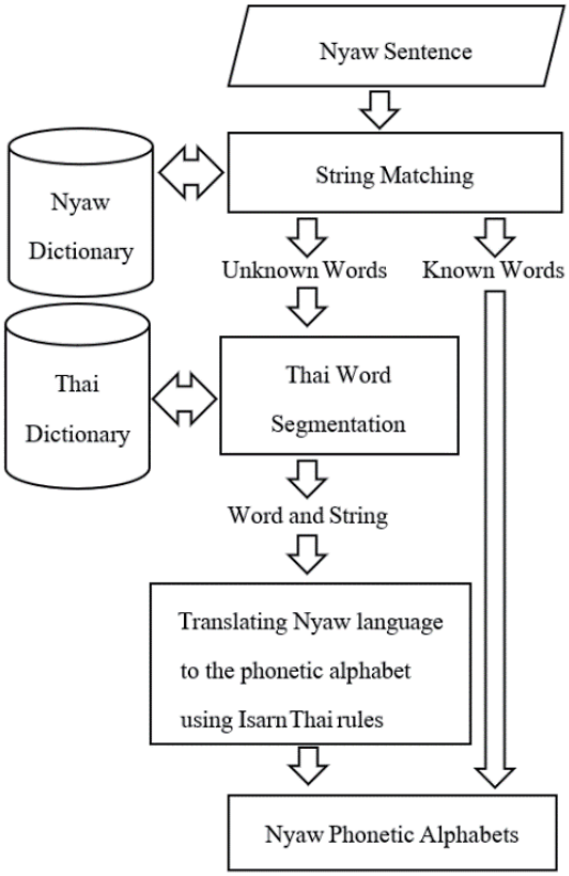

FIGURE 1. The main process of the proposed method

3.1. **Word segmentation using the Nyaw dictionary.** In the first step, the string matching algorithm [15] is applied to determining the word boundary in any sentence or text according to the word in the Nyaw dictionary [14]. This algorithm finds the start position and end position of the word in which the process can be performed according to Algorithm 1. For example, T = "แก้วไปไสมา" (Kaew, where have you been?), the result from this process is W = {w7 (ไส)} and we then also got a phoneme ไส = SĂY from the dictionary.

Algorithm 1 String matching algorithm.

Let:
T is an input sentence (S).
P is a pattern of character found in the dictionary.
W is a set of output word.
i is the first position of P.
j is the first position of T.
k is the character found in the dictionary.
P_k is the position of the character found in the dictionary.
Initialisation:
$i = 1, j = 1, k = 1$
Loop Process:
while $j \leq T$ and $k \leq P$ **do**
  if $T_j == P_k$ **then**
    $j = j + 1$ and $k = k + 1$
  else
    $i = i + 1$, $j = i$, and $k = 1$
  end if
  if $k > P$ **then**
    return W
  end if
end while

3.2. **Word segmentation for the unknown word.** For the remaining sentences from the previous process that consist of the character or string that do not appear in the Nyaw dictionary, it will be segmented using the longest matching algorithm [16,17] according to Algorithm 2. The process also uses the dictionary to decide an appropriate word boundary in which it considers the sentence to finding the word in the left to the right direction. This process is performed until finding the end of the sentence, e.g., spacing or new paragraph. Then, the part of the sentence will be decided as a word if it is founded in the dictionary. Otherwise, the length of the sentence will be decreasing. The process will be repeated until the sentence can be paired with some word in the dictionary. We add the marker after the position that the word is founded as the backtracking point to repeat the longest matching process for finding the next words. For instance, T = "แก้วไป--มา", the result from this process is W = {w1 (แก้ว), w5 (ไป), w9 (มา)}.

3.3. **Conversion rules for Nyaw phonetics.** Finally, the words achieved by all segmentation processes will be converted to the Nyaw phonetics according to the rules of the pronunciation system of the Isarn Thai language [6]. To build these conversion rules, we considered the basics of the Nyaw language, which consists of 20 consonants, 21 vowels, nine short vowels, nine long vowels, three compound vowels, and tonal sounds of six phonemes. It can be converted into Nyaw language by categorizing three types of syllable structures by defining symbols to build the conversion rules as

c = Consonant, T = Tone, v = Verb and Short vowel, and vv = Long vowel

Algorithm 2 Longest matching algorithm.

0: Let:
T is an input sentence (S).
D is a set of word in the Thai dictionary.
W is a set of output word.
N is the length of the sentence.
i is the position of w_i.
j is the current starting position marker.
LP is the last of position marker.
TN is the length of remaining sentence.
0: *Loop Process*:
1: **while** $TN > 0$ **do**
2: **if** $T[j \text{ to } LP] \in D$ **then**
3: $W[i] = T[j \text{ to } LP]$
4: $i = i + 1,\ j = LP + 1$
5: $TN = length(T),\ LP = N$
6: **else if** $LP == j$ **then**
7: $W[i] = T[i]$
8: $i = i + 1,\ j = LP + 1$
9: $TN = length(T),\ LP = N$
10: **else**
11: $LP = TN - 1$
12: **end if**
13: **end while**
13: *Termination*:
14: **return** W

There are three types of phonetic conversion rules for Mono-syllabic, Di-syllabic and Tri-syllabic as follows.

Mono-syllabic consists of three phoneme conversion rules as follows:

Rule 1. cvvT Rule 2. cvcT Rule 3. cvvcT

Di-syllabic consists of ten phoneme conversion rules as follows:

Rule 1. cvTcvvT	Rule 2. cvvTcvvT	Rule 3. cvTcvvcT
Rule 4. cvvTcvvcT	Rule 5. cvTcvcT	Rule 6. cvcTcvcT
Rule 7. cvvcTcvcT	Rule 8. cvvcTcvvcT	Rule 9. cvcTcvvT
Rule 10. cvvcTcvvT		

Tri-syllabic consists of eight phoneme conversion rules as follows:

Rule 1. cvTcvcTcvvT	Rule 2. cvTcvvcTcvvT	Rule 3. cvcTcvTcvcT
Rule 4. cvcTcvTcvvcT	Rule 5. cvcTcvTcvT	Rule 6. cvcTcvTcvvT
Rule 7. cvvTcvTcvvT	Rule 8. cvTcvvcTcvcT	

However, to determine whether any word is Mono-syllabic, Di-syllabic, or Tri-syllabic can be performed by considering the word structure; the consonants and vowels are compared to the structure of phonetic conversion rules to discover the rule that matches the word structure.

Table 10 illustrated examples of applying the rules for phoneme conversion from the Nyaw language "แก้วไป--มา" (Kaew, where have you been?) to their phoneme.

Finally, we then combine the results obtained from Section 3.1, Section 3.2, and Section 3.3 to build the final results as T = {แก้ว (kɜˆ:w), ไป (pay), ไส (săy), มา (ma:)}.

4. **Experiment and Results.** In the experiment, we found that the efficiency of the phonetic conversion method is also dependent on the efficiency of the word segmentation

TABLE 9. Example of applying three types of phonetic conversion rules

Rule	Syllable structure	Nyaw language word	Phoneme
Mono-syllabic			
Rule 1. c [_า, ี,ื,...] T, [เ_, แ_, ไ_,...] c T	cvvT	ไป, แม่	pay, mɛ̆:
Rule 2. c [_ะ,ิ,ึ,...] c T	cvcT	มน	mon
Rule 3. c [_า, ี,ื,...] c T, [เ_, แ_, ไ_,...] c c T	cvvcT	ก้วย, เบิ่ง	kûuay, bə̆:ŋ
Di-syllabic			
Rule 1. c [_ะ,ิ,ึ,...] T c [_า, ี,ื,...] T	cvTcvvT	กะทา	kà tha:
Rule 2. c [_า, ี,ื,...] T [เ_, แ_, ไ_,...] c T	cvvTcvvT	โสเล่	sŏ: lé:
Rule 3. c [_ะ,ิ,ึ,...] T c[_า, ี,ื,...] c T	cvTcvvcT	กะปอม	ka pɔ:m
Rule 4. c [_า, ี,ื,...] T c [_า, ี,ื,...] c T	cvvTcvvcT	ตาเวน	ta: we:n
Rule 5. c [_ะ,ิ,ึ,...] T c [_ะ,ิ,ึ,...] c T	cvTcvcT	ขะลำ	ka lam
Rule 6. c [_ะ,ิ,ึ,...] c T c [_ะ,ิ,ึ,...] c T	cvcTcvcT	สิงไค	sĭŋ khay
Rule 7. c [_า, ี,ื,...] c T c [_ะ,ิ,ึ,...] c T	cvvcTcvcT	ปานดง	pa: doŋ
Rule 8. c [_า, ี,ื,...] c T c [_า, ี,ื,...] c T	cvvcTcvvcT	จอนฟอน	cɔ:n fɔ:n
Rule 9. c [_ะ,ิ,ึ,...] c T [เ_, แ_, ไ_,...] c T	cvcTcvvT	คันแท	khan thɛ:
Rule 10. c [_า, ี,ื,...] c T [เ_, แ_, ไ_,...] c T	cvvcTcvvT	ปานแห	pa:n hɛ̆:
Tri-syllabic			
Rule 1. c [_ะ,ิ,ึ,...] T c [_ะ,ิ,ึ,...] c T c [_า, ี,ื,...] T	cvTcvcTcvvT	กระดันงา	ka dan ŋa:
Rule 2. c [_ะ,ิ,ึ,...] T [เ_, แ_, ไ_,...] c c T c [_า, ี,ื,...] T	cvTcvvcTcvvT	กะโบงดา	ka bo:ŋ ta:
Rule 3. c [_ะ,ิ,ึ,...] c T c [_ะ,ิ,ึ,...] T c [_ะ,ิ,ึ,...] c T	cvcTcvTcvcT	สักลัน	sák ka lan
Rule 4. c [_ะ,ิ,ึ,...] c T c [_ะ,ิ,ึ,...] T [เ_, แ_, ไ_,...] c c T	cvcTcvTcvvcT	สักเทียม	sàk ká thiam
Rule 5. c [_ะ,ิ,ึ,...] c T c [_ะ,ิ,ึ,...] T c [_ะ,ิ,ึ,...] T	cvcTcvTcvT	สำมปี	săm ma pĭ?
Rule 6. c [_ะ,ิ,ึ,...] c T c [_ะ,ิ,ึ,...] T c [_า, ี,ื,...] T	cvcTcvTcvvT	สังกลี	săŋ ka li:
Rule 7. c [_า, ี,ื,...] T c [_ะ,ิ,ึ,...] T c [_า, ี,ื,...] T	cvvTcvTcvvT	อาธนา	hâ: tha na:
Rule 8. c [_ะ,ิ,ึ,...] T c [_า, ี,ื,...] c T [เ_ะ,...] c c T	cvTcvvcTcvcT	ตะบองเพ็ด	ka bɔ:ŋ phét

TABLE 10. Example of applying the conversion rules

Position	Nyaw language	Rule	Phoneme
1	แก้ว	cvvcT	kɜ̂:w
5	ไป	cvvT	pay
9	มา	cvvT	ma:

algorithm. Therefore, the performance evaluation of the proposed method consists of two parts, i.e., the results from known words and the results from unknown words. In this regard, we used precision, recall, and F-measure [18] according to (1) as follows:

$$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{1}$$

where precision is the number of correct conversions of the word from the proposed method divided by the total conversion results from the proposed method, and recall is the number of correct conversions of the word from the proposed method divided by the results from the expert.

This research intended to consider only the word that appears in the dictionary with not more than three syllables. Due to the fact that the Isan Thai language has a limitation in the syllable structure, and it is used only by the local people, it causes most of the words to arise and appear to be cognate words. For the result of all segmented words, the efficiency of word segmentation will be determined in two levels (syllable and word level) before applying the phonetic conversion rule. The results are shown in Table 11.

TABLE 11. Word segmentation result on syllable and word level

Level	Precision (%)	Recall (%)	F-measure (%)
Syllable	78.42	82.86	80.53
Word	84.26	80.51	82.34

In the experiments, we used 68 sentences of Nyaw language that were used in daily life from the book of Nyaw language and food [19], and 500 sentences made by the researcher with all 4,098 words. To build the 500 additional sentences, we collected the data from the expert interviews, the words from the Nyaw dictionary [14], modified the sentences from the book [19], and the sentences collected from daily life used in the Nyaw dialect. We compared phonetic translation efficiency with the longest matching, the basis and the popularity of the Thai word segmentation algorithm [2,15,16,20] to evaluate the proposed method, as shown in Table 12. We found that the phonetic translation efficiency of the longest matching algorithm only achieved 72.78%, 76.54%, and 74.61% for precision, recall, and F-measure, respectively. Whereas the efficiency of the proposed phonetic translation method achieved precision 80.20%, recall 84.97%, and F-measure 82.52%.

TABLE 12. Comparison of phonetic translation efficiency obtained by the longest matching algorithm and the proposed method

Methods	Precision (%)	Recall (%)	F-measure (%)
Longest matching	72.78	76.54	74.61
Propose method	**80.20**	**84.97**	**82.52**

5. **Conclusions and Future Work.** The Nyaw dialect has a limitation of syllable structure because it is localized to only local people different from the standard Thai language, cognate terms. However, the Nyaw language is similar to Isarn Thai but different in the spoken accent, in which some sounds cannot be substituted. Therefore, this paper proposed converting Nyaw words into phonetic characters using linguistic rules that applied the conversion rules of the Isarn Thai language pronunciation system. The results showed that the proposed method could gain good effectiveness and be useful. However, the missing character in the words or some differences in phoneme of the Nyaw language can be further developed in the future.

Acknowledgment. This study was a good success with the support to Faculty of Science and Engineering which funded this research, Kasetsart University Chalermphrakiat Sakon Nakhon Province Campus which provided research resources and working.

REFERENCES

[1] M. A. Ordean, A. Saupe, M. Ordean, M. Duma and G. C. Silaghi, Enhanced rule-based phonetic transcription for the Romanian language, *2009 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pp.401-406, 2009.

[2] B. Sriman and P. Seresangtakul, Thai text to phonetic transcription using dictionary and linguistic rule, *KKU Research Journal (Graduate Studies) Khon Kaen University*, vol.6, no.2, pp.58-67, 2006.

[3] X. Huang, A. Acero, H.-W. Hon and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 2001.
[4] T. Charoenporn, A. Chotimongkol and V. Sornlertlamvanich, Automatic Romanization for Thai, *Proc. of the 2nd International Workshop on East-Asian Language Resources and Evaluation (ORIENTAL COCOSDA'99)*, Taipei, Taiwan, pp.137-140, 1999.
[5] A. Khamya, L. Narupiyakul and B. Sirinaovakul, Speech computing in Thai polysyllabic words, *KMUTT Research and Development Journal*, vol.23, no.1, pp.57-70, 2000.
[6] W. Kingkham, *ThaiTin*, Publisher Kasetsart University, Bangkok, 2013.
[7] P. Seresangtakul and T. Takara, Synthesis of polysyllabic sequences of Thai tones using a generative model of fundamental frequency contours, *IEEJ Transactions on Electronics, Information and Systems*, vol.125, no.7, pp.1101-1108, 2005.
[8] S. Aunkaew, M. Karnjanadecha and C. Wutiwiwatchai, Constructing a phonetic transcribed text corpus for southern Thai dialect speech recognition, *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp.69-73, 2015.
[9] N. Phaiboon and P. Seresangtakul, Isarn Dharma alphabets lexicon for natural language processing, *2017 9th International Conference on Knowledge and Smart Technology (KST)*, pp.211-215, 2017.
[10] K. Singvongsa and P. Seresangtakul, Lao-Thai machine translation using statistical model, *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp.1-5, 2016.
[11] Intangible Cultural Heritage, *Ethnic Languages (Northeastern Region): Nyaw Language*, http://ich.culture.go.th/, 2021.
[12] P. Janyoi and P. Seresangtakul, An Isarn dialect HMM-based text-to-speech system, *2017 2nd International Conference on Information Technology (INCIT)*, pp.1-6, doi: 10.1109/INCIT.2017.8257873, 2017.
[13] National Electronics and Computer Technology Center, *Lexitron Dictionary*, http://lexitron.nectec.or.th, 2019.
[14] T. Hengsanankun and S. Sansri, The conversion of Nyaw-Thai vocabulary to Thai phonetic alphabet by using rule-based approach, *2019 International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology*, pp.363-369, 2019.
[15] Y. Poovorawan and V. Imarom, Dictionary-based thai syllable segmentation (in Thai), *The 9th Electrical Engineering Conference*, 1986.
[16] V. Sornlertlamvanich, Word segmentation for Thai in machine translation system, *Machine Translation*, pp.556-561, 1993.
[17] S. Somsap and P. Seresangtakul, Isarn Dharma word segmentation using a statistical approach with named entity recognition, *ACM Transactions on Asian and Low-Resource Language Information Processing (TAL-LIP)*, vol.19, no.2, pp.1-16, 2020.
[18] J. Davis and M. Goadrich, The relationship between precision-recall and ROC curves, *Proc. of the 23rd International Conference on Machine Learning*, pp.233-240, 2006.
[19] Nakhon Phanom Library, *Local Language and Food of Thai Nyaw*, http://www.nakhonphanombook.ml.ac.th/, 2021.
[20] S. Somsap and P. Seresangtakul, Isarn Dharma word segmentation, *Proc. of the International Conference on Control, Automation and Information Sciences (ICCAIS'13)*, pp.53-57, 2013.