

A STRUCTURAL GRAMMAR FOR UNSUPERVISED LANNA WORD SEGMENTATION FRAMEWORK

Natsima Suradet^{1*} and Wilawan Yathongkhum²

Received: January 30, 2019; Revised: May 13, 2019; Accepted: May 14, 2019

Abstract

This paper presents a new unsupervised approach to Lanna word segmentation. In a previous work, we proposed the Lanna structural grammar for computational linguistics processing. Accuracy evaluation shows that the grammar is feasible, valid, and effective. Additionally, we developed the unsupervised word segmentation algorithm that applied the Lanna structural grammar to predict words from character sequences and a search for the optimal segmentation. The proposed algorithm was evaluated on Lanna phrases from a variety of document types. The experimental results show that the segmentation accuracy is effective and practical. It is concluded that the Lanna structural grammar is appropriate and capable for unsupervised Lanna word segmentation. Consequently, our approach is suitable for computational linguistics processing and has the potential to improve the coverage of morphological analyzers for the Lanna language without explicit word boundary markers.

Keywords: Lanna structural grammar, Lanna word segmentation, unsupervised word segmentation, Lanna word segmentation framework

Introduction

The Lanna language is a tonal language in which the different phonemic tones may lead to a different intended meaning of a word. This language is still spoken by local people in Northern Thailand. The Lanna script is a descendant of the Old Mon script and can predominantly be found in old religious

manuscripts, texts, and temple scriptures (Maneechedta, 1994; Pansook, 2008; Prongthura, 1982). Figure 1 shows an example of Lanna script.

The Lanna writing system does not have a standard form and can be represented using various writing styles, since the Lanna learning

¹ Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Lanna, Tak, 63000, Thailand. E-mail: natsima@hotmail.com

² Ban Mae Thoei School, Paholayothin Road, Mae Tuen, Li, Lamphun, 51110, Thailand.

* Corresponding author

system uses a teacher- centered instruction format (Maneechedta, 1994). Lanna text is written continuously without a delimiter between words causing several difficulties for describing the language characterization.

Word segmentation is an important task in natural language processing for languages without word delimiters. Most existing approaches are supervised. However, unsupervised approaches are more adaptive and can cooperate with supervised ones to overcome the drawbacks of both approaches. The unsupervised word segmentation methods are represented by various research works. Goldwater *et al.* (2006) presented an alternative framework for word segmentation that relies on Bayesian statistics through a method based on Dirichlet processes and inference by using Gibbs sampling. Mochihashi *et al.* (2009) proposed a combination of dynamic programming and an accurate spelling model from a Bayesian perspective. Zhikov *et al.* (2010) presented an algorithm for unsupervised word segmentation with branching entropy and minimum description length that

utilizes the local predictability of adjacent character sequences. Wang *et al.* (2011) proposed an iterative process ESA that consist of 3 phases: evaluation, selection, and adjustment. The research suggests that it is effective on Chinese corpora. Spiegler *et al.* (2009) described a Promodes algorithm for unsupervised word decomposition which was based on a probabilistic generative model and applied statistics of a small training set instead of building a morphological dictionary and a rule base.

Many previous methods focus on the observation that transitions between units (characters, phonemes, or syllables) within words are generally more predictable than transitions across word boundaries, while statistics have been proposed for measuring these differences. Several methods are based solely on local statistics and a few works exist on unsupervised word segmentation from linguistic knowledge. However, there is some research focusing on unsupervised approaches through a knowledge of the structure of a language. Somsap and Seresangtakul (2013)

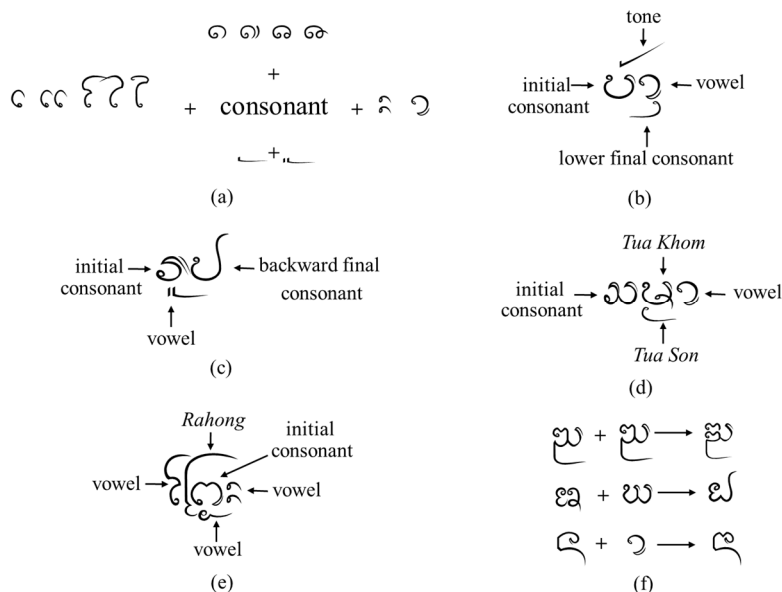


Figure 1. The structures of Lanna words: (a) no final consonant type, (b) lower final consonant, (c) backward final consonant, (d) KhomSon consonant, (e) Rahong cluster, and (f) special consonant

where the final consonant appears below the initial consonant or it can appear below the vowels as shown in Figure 1(b).

3. Backward final consonant (BFC) is a word where the final consonant directly follows the initial consonant when the vowel or vowel combination takes up a space below the preceding consonant as shown in Figure 1(c).

5. *Rahong* cluster is a word where the *Rahong* (ꠘꠞꠤ) symbol appears directly in front of the initial consonant as shown in Figure 1(e).

The Lanna writing system is complex, and it has the particularity to be used to write different forms with different rules and specific characters. Unfortunately, nowadays, we still lack a Lanna corpus and dictionary for computational language processing. Therefore, we have designed the Lanna structural grammar based on the Lanna writing system that is described above. In our previous work, we proposed the Lanna structural grammar for computational linguistics processing. We applied this grammar for a word segmentation system without using a dictionary. The experimental results demonstrated the effectiveness of the grammar (Suradet and Yathongkhum, 2014). However, the word segmentation system cannot fit in the case of words that contain a special character. Some Lanna words indicate that the special characters have their own specific writing system that are excluded in this grammar. Thus, we have designed additional grammar for Lanna special characters. The proposed adaptive Lanna structural grammar can be stated as follows:

Lanna Structural Grammar

1. No final consonant type which consists of only 3 components: consonant, vowel, and tone mark. The vowels can occur before, after, above, and below the consonants or combinations of these places as shown in Figure 1(a).

2. Lower final consonant (LFC) is a word

Word	(No_final_consonant LFC BFC KhomSon Rahong_cluster Special_consonant)*
No_final_consonant	→ ([consonant] [consonant_cluster]) (ᵀ ᶜ ᵉ ᵒ ᶠ ᶡ ᶢ) [tone]? ([consonant] [consonant_cluster]) [tone]? (ᵁ ᵃ ᵇ) ᵈ (ᵋ ᵌ ᵍ) ([consonant] [consonant_cluster]) [tone]? :? ᵊ (ʔ ʔ̣) ([consonant] [consonant_cluster]) [tone]? (ᵐ ᵎ) ᶑ ([consonant] [consonant_cluster])_ḥ ([consonant] [consonant_cluster]) (˘ ˙) (ᶛ [tone]? ˚) c ([consonant] [consonant_cluster]) (ᶞ ᶟ ˘? [tone]?_j ᶝ ˘? [tone]? ɹ) c ([consonant] [consonant_cluster]) (ᵉ ᵓ [tone]?_ḿ; :?)
LFC	→ [consonant] (ᵀ ᶜ ᵉ ᵒ ᶠ ᶡ ᶢ) [tone]? [lower_final_consonant] ([consonant] [consonant_cluster]) [tone]? ɹ [lower_final_ consonant] c (ᵋ ᵌ) [consonant] [tone]? [lower_final_consonant] c [consonant] ((lower_final_consonant) ; ᵉ [tone]? [lower_final_ consonant])) ᶑ [consonant] · ? [tone]? [lower_final_consonant]
BFC	→ ([consonant] [consonant_cluster]) (ᵱ ᵲ ᵳ) [tone]? [backward_ final_consonant] ([consonant] [consonant_cluster]) [tone]? (ᵴ ᵵ ᵶ) [backward_ final_consonant] c ([consonant] [consonant_cluster]) (ᵷ ᵸ) [tone]? ˗ [backward_ final_consonant] [consonant_cluster] (ᵺ ᵻ ᵼ ᵽ ᵾ ῀) [tone]? [backward_ final_consonant] c (ᵋ ᵌ) [consonant_cluster] [tone]? [backward_final_consonant] c [consonant_cluster] ([BFC] ; ᵉ [tone]? [backward_final_ consonant])) ᶑ [consonant_cluster] · ? [tone]? [backward_final_consonant]
KhomSon	→ c [consonant] c [KhomSon_consonant] ɹ c [consonant] (ᶤ ᶦ) [KhomSon_consonant] c [consonant] [KhomSon_consonant] (ᵅ ᵆ ᵇ)? [consonant] (ᵋ ᵌ ᵍ ᶫ) [KhomSon_consonant] [consonant] [KhomSon_consonant]? (ᵱ ᵲ) c [KhomSon_consonant] ɹ? [consonant] [KhomSon_consonant]{1,2} (ᵅ ᵆ ᵇ ᵏ)? [consonant]{1,2} (ᵱ ᵲ ᵳ ᵴ ᵵ ᵶ)[KhomSon_consonant][consonant]? (ᵅ ᵆ ᵇ ᵏ z ᵱ) [consonant] [KhomSon_consonant] _ [KhomSon_consonant] [KhomSon_consonant]
Rahong_cluster	→ (c ᵋ ᵌ ᵍ ᵍᵏ) ? ƭ [consonant] [tone]? [lower_final_consonant]? ƭ [consonant] [tone]? ɹ [lower_final_consonant]? ƭ [consonant] (ᵱ ᵲ ᵳ) [tone]? [backward_final_consonant]? ƭ [consonant] (ᵷ ᵸ) [tone]? [lower_final_consonant]? ƭ [consonant] (· ᵴ)? [tone]? [lower_final_consonant] ƭ [consonant] · ɹ [tone]? ƭ [consonant] · [tone]? ˚. c ƭ [consonant] ᵕ [tone]? [lower_final_consonant]? ƭ [consonant] ɹ [KhomSon_consonant] ˘ [consonant] (ᵉ ᵱ ᵲ ᵳ) (ᵰ ᵱᵰ) (ᵉ ᵏ) (j)? [consonant] (ᵉ ᵱ ᵳ) ᵰᵰ [consonant] (ᵉ ᵱ ᵳ) ᵰᵰᵰᵰ [backward_final_consonant]? [consonant] (ᵅ ᵆ) ᵰᵰ
Special_consonant	→ ᵯ ᵱ ᵲ ᵳ ᵴ ᵵ ᵶ ᵷ ᵸ ᵹ ᶀ ᶁ ᶂ ᶃ ᶄ ᶅ ᶆ ᶇ ᶈ ᶉ ᶊ ᶋ ᶌ ᶍ ᶎ ᶏ ᶐ ᶑ ᶒ ᶓ ᶔ ᶕ ᶖ ᶗ ᶘ ᶙ ᶚ ᶛ ᶜ ᶝ ᶞ ᶟ ᶠ ᶡ ᶢ ᶣ ᶤ ᶥ ᶦ ᶧ ᶨ ᶩ ᶪ ᶫ ᶬ ᶭ ᶮ ᶯ ᶰ ᶱ ᶲ ᶳ ᶴ ᶵ ᶶ ᶷ ᶸ ᶹ ᶺ ᶻ ᶼ ᶽ ᶾ ᶿ ᷀ ᷁ ᷂ ᷃ ᷄ ᷅ ᷆ ᷇ ᷈ ᷉ ᷊ ᷋ ᷌ ᷍ ᷎ ᷏ ᷐ ᷑ ᷒ ᷓ ᷔ ᷕ ᷖ ᷗ ᷘ ᷙ ᷚ ᷛ ᷜ ᷝ ᷞ ᷟ ᷠ ᷡ ᷢ ᷣ ᷤ ᷥ ᷦ ᷧ ᷨ ᷩ ᷪ ᷫ ᷬ ᷭ ᷮ ᷯ ᷰ ᷱ ᷲ ᷳ ᷴ ᷵ ᷶ ᷷ ᷸ ᷹ ᷺ ᷻ ᷼ ᷽ ᷾ ᷿ ᷀ ᷁ ᷂ ᷃ ᷄ ᷅ ᷆ ᷇ ᷈ ᷉ ᷊ ᷋ ᷌ ᷍ ᷎ ᷏ ᷐ ᷑ ᷒ ᷓ ᷔ ᷕ ᷖ ᷗ ᷘ ᷙ ᷚ ᷛ ᷜ ᷝ ᷞ ᷟ ᷠ ᷡ ᷢ ᷣ ᷤ ᷥ ᷦ ᷧ ᷨ ᷩ ᷪ ᷫ ᷬ ᷭ ᷮ ᷯ ᷰ ᷱ ᷲ ᷳ ᷴ ᷵ ᷶ ᷷ ᷸ ᷹ ᷺ ᷻ ᷼ ᷽ ᷾ ᷿ [consonant] ˘ (ᶛ ᶜ) [tone]? [lower_final_consonant]? ᵈ [tone]? (ᵋ ᵌ ᵍ [lower_final_consonant]?)

Lanna Word Segmentation Algorithms

The purpose of the unsupervised Lanna word segmentation algorithm is to discover correct word boundaries in a Lanna sentence. The Lanna structural grammar was applied for matching a character sequence in the forward detection and backward detection algorithms. The matching process, with forward or backward matching will be executed to obtain all the possible segmentation candidates. Then, a search for the optimal segmentation will determine word boundaries by the word segmentation algorithm.

We introduce here the notation that will be used throughout the algorithms in this paper. Let $S[c_1, c_2, \dots, c_n]$ be an array of characters in the Lanna sentence where n is the number of all characters. Let T be a set of word types as described in section 2. Let P_1 and P_2 be an $n \times n$ array of possible word types in S that are detected by the forward detection algorithm and backward detection algorithm, respectively. Let W denote the set of all words that are segmented by the proposed unsupervised Lanna word segmentation algorithm.

Forward Detection (S)

```

1  for  $i \leftarrow 1$  to  $n$ 
2    do for  $j \leftarrow i$  to  $n$ 
3      do detect word boundary by
        Lanna structural grammar
4        if  $S[c_i, \dots, c_j] \in T$ 
5          then  $P_1[i, j] \leftarrow 1$ 
6        else  $P_1[i, j] \leftarrow 0$ 
7  Word segmentation ( $P_1, S$ )
```

Backward Detection (S)

```

1  for  $i \leftarrow n$  down to 1
2    do for  $j \leftarrow i$  down to 1
3      do detect word boundary by
        Lanna structural grammar
4      if  $S[c_j, \dots, c_i] \in T$ 
5        then  $P_2[j, i] \leftarrow 1$ 
6      else  $P_2[j, i] \leftarrow 0$ 
7  Word segmentation ( $P_2, S$ )
```

Word Segmentation (P, S)

```

1   $i \leftarrow 1$ 
2   $m \leftarrow 1$ 
3  while  $i \leq n$ 
4    do for  $j \leftarrow n$  down to 1
5      do if  $P[i, j] = 1$ 
6        then  $k \leftarrow j$ 
7       $W[m] \leftarrow S[c_i \dots c_k]$ 
8       $m \leftarrow m + 1$ 
9     $i \leftarrow k + 1$ 
```

Unsupervised Lanna Word Segmentation Framework

The purpose of the unsupervised Lanna word segmentation framework, as shown in Figure 2, is to discover correct word boundaries in a Lanna sentence. The input of our framework is the sequence of characters. We develop the forward detection and backward detection algorithms applied with the Lanna structural grammar to predict words from the character sequence. The aim is to compare the word segmentation correctness between the 2 algorithms. The matching process, with forward or backward matching will be executed to obtain all the possible segmentation candidates. Then, the word segmentation algorithm is applied to select the most appropriate segmentation choice from the candidates.

Experiments

We validate the proposed framework by randomly selecting Lanna phrases which contain 8,773 Lanna words from a variety of document types. The datasets are separated into 6 groups as below:

Group 1 is a set of Lanna phrases that contains 3,631 Lanna words which were randomly selected from Lanna Khaosaw, the Setthi Hua Wiang (Punjaburi, 2014). Khaosaw is a form of literature that uses aesthetic and rhythmic language.

Group 2 is a set of Lanna phrases that contains 912 Lanna words which were randomly selected from idioms in Lanna legends (Lackmuang, 2009).

Group 3 is a set of Lanna phrases that contains 144 Lanna words which were randomly selected from Lanna inscriptions for worshipping in Chedi Phra Mahathat Si Wiang Chai, Lamphun, Thailand. The characteristics of words in this group have grammatical features and terminology similar to the Pali language which uses subscript consonant symbols, double consonants such as the *Khomson* type that tend to compress and shorten a vowel sound.

Group 4 is a set of Lanna phrases that contains 655 Lanna words which were randomly selected from the lessons of basic Lanna Dhamma script (Inchai and Wongthong, 2018). The words in this group include simple and compound words.

Group 5 is a set of Lanna phrases that contains 310 Lanna words which were randomly selected from news columns of Lanna Kham Muang (The Center for the Promotion of Arts and Culture, Chiang Mai University, 2018).

Group 6 is a set of Lanna phrases that contains 3,121 Lanna words which were

randomly selected from old Lanna machine-printed document images (Chaijaruwanich, 2009). The dataset in this group is from old Lanna literature.

In this research, the segmentation accuracy is evaluated by means of precision, recall, and F-measure:

$$Precision = \frac{W_{correct}}{W_{segment}} \quad (1)$$

$$Recall = \frac{W_{correct}}{W_{all}} \quad (2)$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

where $W_{correct}$ is the number of correctly segmented words, $W_{segment}$ is the total number of segmented words by the algorithm, and W_{all} is the number of words identified by a

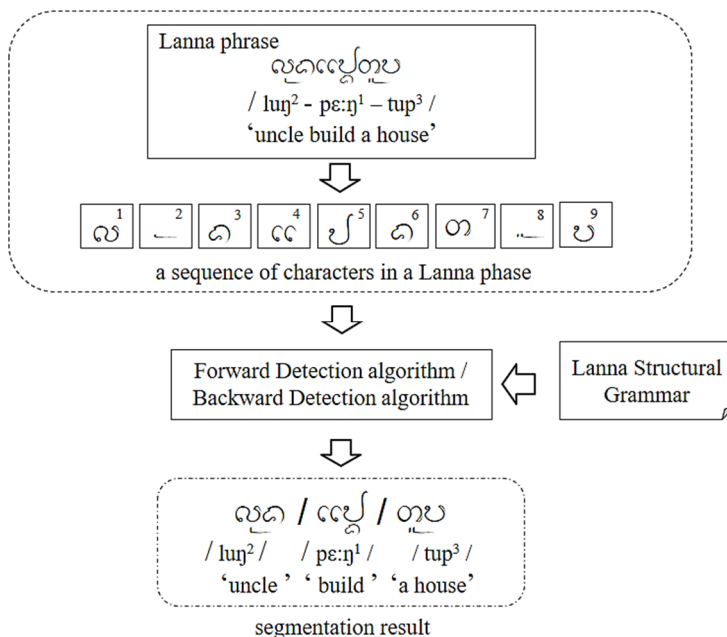


Figure 2. The unsupervised Lanna word segmentation framework

native writer. The evaluation and comparison of different segmentation algorithms are mentioned for the average of the precision and recall for all Lanna phrase groups.

Results and Discussion

The experimental results in Table 1 show the performance of the segmentation accuracy between the forward detection and backward detection algorithms. The precision result of the dataset of group 3 for the forward detection algorithm has a lower segmentation accuracy than the others. It can be seen clearly that the

characteristics of the dataset of group 3 have an impact on the proposed Lanna word segmentation framework, particularly the overlapping ambiguity of the segmentation results from the forward detection algorithm as shown in Figure 3. The segmentation result of the forward detection algorithm is obviously wrong, while the backward detection algorithm can get the correct result. The phonetic symbols that are shown in Figures 2 and 3 followed the phonology of the Lanna language which was found in a Northern Thai dialect book (Wimonkasem, 2006).

From the performance of the segmentation, precision, recall, and F- measure, it was

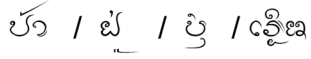
Lanna phrase	
	
/pa:⁵ - yu:³ - bon¹ - hu:a²/	
'Aunt is on the boat'	
Forward Detection algorithm	
	
/pa:⁵ / /yu:p⁴/ /hu:a²/	
'Aunt' 'touch' 'the boat'	
Backward Detection algorithm	
	
/pa:⁵ / /yu:³/ /bon¹/ /hu:a²/	
'Aunt' 'is' 'on' 'the boat'	

Figure 3. The segmentation results from the forward detection and backward detection algorithms

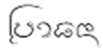
Table 1. The results of the segmentation accuracies

Lanna phrase	Forward detection algorithm			Backward detection algorithm		
	Recall	Precision	F-measure	Recall	Precision	F-measure
Group 1	0.98	0.93	0.96	1.00	0.99	1.00
Group 2	0.97	0.95	0.96	0.99	0.98	0.98
Group 3	0.92	0.59	0.72	0.99	0.92	0.95
Group 4	0.94	0.87	0.90	0.95	0.84	0.89
Group 5	0.97	0.87	0.92	1.00	0.89	0.94
Group 6	0.97	0.94	0.96	1.00	1.00	1.00
All	0.97	0.93	0.95	0.98	0.98	0.98

demonstrated that the backward detection algorithm has a higher accuracy than the forward detection algorithm. It can effectively eliminate most of the overlapping ambiguity which may arise from the forward detection algorithm.

Conclusions

In this research, we proposed a new framework for Lanna word segmentation by developing the unsupervised word segmentation algorithms that were applied to the Lanna structural grammar to search for the optimal segmentation to determine word boundaries. The best overall performance was obtained with the backward detection algorithm, with a precision of 98% and recall of 98% which is high enough to justify its practical applicability and potential. Surprisingly, the performances of the segmentation algorithms do not depend on a variety of document types and are not degraded with the increasing number of words in a phrase. However, after detailed investigation, we found out that the Lanna structural grammar is still not suitable for a group of consonants which have no

intervening vowel such as  (pha:1th a2na:2, which means “want”). Nevertheless, it is a short ‘a’ sound in compound words that is rarely found in Lanna documents but is commonly used in Lanna inscriptions.

Extensive experiments on the various datasets confirm that the Lanna structural grammar can be applied for the word segmentation system without using a corpus or dictionary. Consequently, our approach is suitable for computational linguistics processing and has the potential to improve the coverage of morphological analyzers for the Lanna language.

References

Chaijaruwanich, J. (2009). Wannaphim Lanna: Literature published in Lanna Dharma 60 books. 1st ed. Chiang Mai University Press, Chiang Mai, Thailand, 6,402p.

- Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. Proceedings of the 21st International Conference on Computational Linguistics; July 17-18, 2006; Sydney, Australia, p. 673–680.
- Inchai, S. and Wongthong, A. (2018). Introduction to Lanna Dhamma alphabet. Chiangrai: The Art and Culture Office Chiangrai, Chiang Rai Rajabhat University. Available from: <http://ctc.rru.ac.th/topic/708>. Accessed date: May 3, 2018.
- Lackmuang, P. (2009). Conceptual analysis of idioms in Lanna legends, [M.Ed. thesis]. Department of Education Program in Teaching Thai Language, Graduate School, Chiang Mai University, Chiang Mai, Thailand, 470p.
- Maneechedta, W. (1994). Enculturation process on Lanna Dhama alphabets, [M.Ed. thesis]. Department of Nonformal Education, Graduate School, Chiang Mai University, Chiang Mai, Thailand, 87p.
- Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian unsupervised word segmentation with nested Pitman- Yor language modeling. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing; August 2-7, 2009; Singapore, p. 100-108.
- Pansook, A. (2008). Status and role changes of the Lanna language, [M.A. thesis]. Department of Lanna Language and Literature, Graduate School, Chiang Mai University, Chiang Mai, Thailand, 188p.
- Prongthura, N. (1982). Dhamma scripts of Northern Thailand, [M.A. thesis]. Department of Oriental Language, Graduate School, Silpakorn University, Bangkok, Thailand, 313p.
- Punjaburi, N. (2014). An analytical study of a Lanna Khaosaw, the Setthi Hua Wang, [M.A. thesis]. Department of Oriental Language, Graduate School, Silpakorn University, Bangkok, Thailand, 424p.
- Somsap, S. and Seresangtakul, P. (2013). Isarn Dharma word segmentation. Proceedings of the 2013 International Conference on Control, Automation and Information Sciences; November 25-28, 2013; Nha Trang, Vietnam, p. 53-57.
- Siriratphiriya, K. (2005). Lanna language: the northern Thai characters and how to read them through the structure of words. 1st ed. Sukhothai Thammathirat University Press, Nonthaburi, Thailand, 234p.
- Spiegler, S., Golenia, B., and Flach, P. A. (2009). Unsupervised word decomposition with the Promodes algorithm. Proceedings of the 10th Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments; September 30 - October 02, 2009; Corfu, Greece, p. 625-632.
- Suradet, N. and Yathongkhum, W. (2014). Lanna morphology grammar approach for computational linguistics processing. Proceedings of the 18th International Computer Science and Engineering Conference; July 30 - August 1, 2014; Khon Kaen, Thailand, p. 80–85.

- The Center for the Promotion of Arts and Culture, Chiang Mai University. (2018). Lanna Kam Muang. Bangkok: Matichon Public Co., Ltd. Available from: www.matichonweekly.com/magazine-column/lanna. Accessed date: Apr 27, 2018.
- Wang, H., Zhu, J., Tang, S., and Fan, X. (2011). A new unsupervised approach to word segmentation. *Comput. Linguist.*, 37(3):421-454.
- Wimonkasem, K. (2006). Northern Thai Dialect. 2nd ed. Silpakorn University Press, Bangkok, Thailand, 366p.
- Zhikov, V., Takamura, H., and Okumuran, M. (2010). An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*; October 9-11, 2010; Cambridge, MA, USA, p. 832-842.