A tailor-made approach to Thai word segmentation for topic-specific research

Punjaporn Pojanapunya and Duangjaichanok Pansa
King Mongkut's University of Technology Thonburi

Abstract

Segmenting Thai words for use in corpus-based studies is a complex task. Two major approaches for Thai word segmentation are dictionary-based (DCB) and machine learning-based (MLB). However, it is unclear which method produces the most appropriate segmented text for use in a corpus-based analysis. This paper describes a novel third approach, a two-level segmentation which segments text by using specifically designed criteria. By integrating existing approaches with specific criteria, this method segments Thai text into the shortest syllables or words and then creates longer words from 2-word, 3-word and 4-word clusters by using a reference glossary of terms as the basis for identifying clusters. For this study, all three methods were tested on a corpus of interviews on language teachers' views on assessment. For the first two methods, word units were segmented by ready-made programs, LexTo (DCB) and TLex (MLB). Advantages and drawbacks of these three methods for the purpose of facilitating analysts who prepare Thai texts for corpus linguistics are discussed.

1. Introduction

Corpus-based methods have been extensively used in research in applied linguistics for different purposes. Based on the analysis of a collection of texts, researchers can examine word frequency statistics to explore the particular interesting and remarkable words from the lists (Lu, 2020), study contexts where the words are used, and identify common patterns and outstanding issues in the texts (Bennett, 2010). Moreover, corpus analysis, as a quantitative method for analyzing linguistic data (Hasko, 2013), is popularly implemented in diverse fields of study working with various types of qualitative data. In other words, it can be effectively used in research in which qualitative data is a data source (Egbert et al., 2020).

With regards to systematic qualitative data management, corpus-based analysis has been applied to analyse main concepts or themes of several sorts of qualitative data, e.g., interview transcripts of patients' perspectives (Weetman et al., 2018), veterinarians' responses to the open-ended items in a questionnaire (Huntley et al., 2018), online diary data (Teddiman, 2009), and social media posts and news (Marchi & Taylor, 2009; Touileb & Salway, 2014).

This research is a part of the qualitative data preparation stage for a large-scale study which investigated assessment practices used by Thai teachers (Watson Todd et al., 2020). This qualitative data was from teachers' interviews which were conducted in Thai to gather in-depth details about teacher practices and to gain information reflecting their thoughts more effectively. Most of the corpus analysis tools (e.g., AntConc (Anthony, 2019), WordSmith Tools (Scott, 2020), Wmatrix (Rayson, 2009), and Sketch Engine (2021)) can deal with texts which have clear word boundaries, e.g., the spaces between words existing in Latin-based languages such as English, French, Spanish, and so on. The tools have limitations in handling texts in other languages which have complex systems and no clear boundaries between word items such as Asian languages including Chinese, Korean, Japanese, and Thai.

To prepare Thai texts for corpus analysis tools, therefore, a pre-processing task called segmentation or tokenization (Haruechaiyasak et al., 2008) is a required prerequisite stage. The computer software which separates Thai language into distinct word units is called a tokenizer. Segmentation is also an introductory stage dealing with Thai texts for language processing, such as in Machine Translation (MT) or Information Retrieval (IR) systems (Aroonmanakun, 2007; Haruechaiyasak et al., 2008).

Thai word tokenizers still have room for improvement with regards to accuracy due to the complexity of the Thai language. This study argues that there is no single segmentation program which returns the best results that fit all contexts, especially for research working with a specialized text genre. Further pre-processing tasks after automated tokenization, e.g., text editing and repeated tokenization, is always required to make segmented items which effectively address the potential issues in the specialized corpus. We, therefore, share our experience as researchers in applied linguistics who have explored ways of preparing a Thai interview corpus for a corpus-based analysis. We compare and present word items tokenized by two well-known Thai language tokenizers, LexTo and TLex, and another method referred to as two-level tokenization with a combination of manual segmentation and an n-gram technique. Each of the three methods returned different tokenized word items due to their approaches and algorithms. Strengths and limitations of these methods will be discussed and followed by recommendations for researchers who prepare Thai texts for corpus analysis.

2. Approaches to Thai word segmentation and previous studies on Thai word segmentation

The principle behind Thai word segmentation is word creation. Thai words can be created on the basis of different rules with three main ways of creating words: combining morphemes into a word, reduplicating words, and compounding words (Aroonmanakun, 2018). These word creation methods increase the complexity of Thai language and consequently make word segmentation complicated. For example, a compound word is a new meaningful word created from a combination of two words containing specific meanings. For instance, ชี้ [chée] (point, verb) วัด [wát′] (verb: measure; noun: temple) can be combined to coin a new word ชี้วัด [chée-wát′] (verb: indicate). ชี้วัด [chée-wát′] can be considered one word (see Sentence 1) or two words, chée and wát′ (see Sentence 2) depending on particular contexts as shown below.

Sentence 1: ความสำเร็จจะถูกประเมินตามตัว**ชี้วัด**ที่กำหนด
[kwam-săm′-rét′-jà′-tòok-bhrà′-mern-dham-dhua-**chée-wát′**-têe-gam′-nòt′]
Success will be evaluated according to specified **indicators**.

Sentence 2: พระ**ชี้วัด**ที่เพิ่งซ่อมแซมเสร็จ
[prá′-**chée-wát′**-têe-pêrng-sâwm-sæm-sà′-rèt′]
A monk **pointed** to the **temple** which has just been renovated.

In both sentences, chée and wát′ are written consecutively without spaces, so they could be recognized as one or two words according to their contexts and meanings. This context-dependency clearly makes automated word segmentation burdensome and complicated.

Several automated techniques and algorithms have been developed since Thai word segmentation was first researched in the 1980s. Some main approaches of segmentation include rule-based techniques (RB), dictionary-based techniques (DB), and machine learning-based techniques (MLB) (Tapsai et al., 2021). Recently, a hybrid approach which employs a range of various algorithms for segmentation has been increasingly used among researchers in the field.

The first method, a rule-based (RB) approach, acknowledges some rules of Thai language such as character clusters and syllable structures (see Chamyapornpong, 1983; Thairatananond, 1981), syntax, grammar (see Mahatthanachai et al., 2015), and collocations (see Aroonmanakun, 2002) as the basis for segmentation. The second method, a dictionary-based (DB) approach, relies on dictionary entries to segment words into smaller units (e.g., in Poowarawan, 1986). A third approach, machine learning-based (MLB), utilizes a group of datasets to perform word recognition (see Bheganan et al., 2009; Haruechaiyasak & Kongyoung, 2009; Kruengkrai et al., 2006). In more recent papers, the hybrid method which employed, for example, RB and MLB approaches (in Paripremkul & Sornil, 2021), DB, MLB and RB (Hengsanankun & Namburi, 2020), and DB and corpus-based approaches (Tanantong et al., 2020) to word segmentation has been investigated. These approaches and their corresponding technical issues of segmenting words are beyond the scope of this study. Those who are interested in the technical aspects can consult the references provided.

As applied linguistics researchers, we have searched for the methods which would describe how applied linguists have dealt with Thai corpora, especially in a pre-processing stage for corpus analysis. Surprisingly, no research collecting data in Thai language employing corpus analysis was found. Most studies relevant to Thai word segmentation tools have been conducted by software developers, focusing particularly and deeply on technical issues, e.g., algorithms and programing. Design, development, and evaluation of algorithms and programs were taken into account with regards to improving the accuracy of segmentation in each program – many of them aim to solve the complications in terms of word boundary ambiguity and unknown word problems (e.g., Hengsanankun & Namburi, 2020; Jucksriporn & Sornil, 2011; Nararatwong et al., 2018; Thangthai et al., 2013). Most studies associated with pre-processing texts have been conducted as a part of developing or testing programs for text analysis, for example, segmentation as to pre-process texts for the study of keyword extraction (Ousirimaneechai & Sinthupinyo, 2018), the development of summarizing programs (Apisuwankun & Mongkolnavin, 2013; Thumrongluck & Mongkolnavin, 2011), machine translation (Unlee & Seresangtakul, 2016), classification systems (Chanta & Porrawatpreyakorn, 2013; Khowrurk & Kongkachandra, 2020; Kurdkit et al., 2015), and information extraction (Chantaraj & Rungrattanaubol, 2020).

However, few substantial resources and guidelines are available on how to complete word segmentation as a pre-processing task to prepare a corpus for analysis. In this paper, hence, we share our experience on word segmentation by presenting three possible ways of segmenting words. The outputs or segmented word units provided by different programs and methodologies will be shown and compared. The fact that different programs employed distinctive approaches and algorithms to segment words or units has been taken into account. These approaches should be a key criterion for choosing the program for analysis, but the accessibility has been prioritized because of two reasons. First, over half of the segmentation procedures implemented in the previous studies we reviewed are specially designed procedures or programs which are not accessible to general users or researchers from different fields (e.g., in Bheganan et al., 2009; Hengsanankun & Namburi, 2020; Mahatthanachai et al., 2015; Nararratwong et al., 2018; Tanantong et al., 2020; Vichianchai, 2014). In addition, some of the web-based tools are not practical for those who have no background knowledge in Natural Language Processing (NLP) or programming. With respect to those reasons, LexTo and TLex, two of the most commonly used programs, were chosen (Apisuwankun & Mongkolnavin, 2013) to tokenize our corpus. This choice was mainly based on their 1) accessibility, 2) user friendliness, and 3) reliability suggested by previous studies.

As mentioned, most papers associated with Thai word segmentation have focused on technical aspects, but rarely provide practical guidelines for users or researchers who plan to have a corpus-based analysis of Thai data as a part of a study. We believe that our study, one of a few papers, can address practical issues and provide recommendations for the audience with limited background knowledge and skills in NLP and programming to help them prepare a Thai corpus.

Based on the experience of the applied linguists working on segmenting Thai words in a corpus of interviews with Thai teachers on their assessment practices, three possible segmentation methods were chosen. The first two methods are automated while the third is a self-designed method called two-level segmentation. The third method, which we argue is most usable for specific-topic research, is a combination of a manual segmentation and the n-gram function available in AntConc. The outputs or the segmented units provided by these three methods were compared. Some practical suggestions and constructive guidelines for researchers who prepare Thai corpus for corpus-based analysis are also provided based on the findings.

3. The study

This study is a part of the qualitative data preparation stage of a large-scale project funded by the Thailand Science Research and Innovation (TSRI) (Watson Todd et al., 2020).

3.1 Data

The data for this project is a corpus of interviews with 29 English teachers in primary and secondary school on their views and use of assessment practices in their classes. Each interview lasts for 30 – 45 minutes. This data set is utilized in this study to represent a Thai corpus of topic-specific texts.

3.2 Software

Three distinctive software programs have been implemented as the main tools in this study. The first two tools are web-based programs for word segmentation, i.e., the Thai Lexeme Tokenizer (LexTo) developed by National Electronics and Computer Technology Center, Thailand (NECTEC) in 2007 and the Thai Lexeme Analyser (TLex) developed by Haruechaiyasak and Kongyoung (2009). These two programs share a similar purpose which is to segment chunks of Thai language texts into single units separated by spaces. This process was conducted to prepare a compatible and readable corpus for corpus processing tools, e.g., AntConc (Anthony, 2019). Major differences between LexTo and TLex are the approaches and algorithms each of the programs employs to segment words. The first relies on a DB approach while the latter uses an MLB approach. Only the application of these two programs to segmenting a topic-specific corpus will be mentioned in this study. Technical issues and concerns can be found in studies by Haruechaiyasak & Kongyoung (2009), Haruechaiyasak et al., (2008), and Tapsai et al., (2021). The last tool is AntConc (Anthony, 2019), whose several functions are available for processing and analyzing a corpus, such as word list, keyword list, concordance, clusters and collocates.

3.2.1 Thai Lexeme Tokenizer (LexTo)

LexTo, which is an automated program, employs a longest matching algorithm based on a dictionary-based approach (DB) – namely using the Lexitron dictionary (Apisuwankun & Mongkolnavin, 2013). The software examines an input text, reading from left to right, and selects the longest unit result matching with a dictionary entry (Poowarawan, 1986).

It has been claimed by technical papers that Lexto is accurate and reliable (Thumrongluck & Mongkolnavin, 2011), flexible in recognizing words which are not in the dictionary (Apisuwankun & Mongkolnavin, 2013), and is compatible with other instruments (Chanta & Parrawatpreyakorn, 2013).

To our knowledge, however, there has not been any research which utilized or reviewed the application of LexTo in preparing texts for corpus-based analysis of qualitative data, such as interviews or open-ended questions from a questionnaire.

It was suggested in Kurdkit et al. (2015) that LexTo allows users to add unknown, specific, and target words to the dictionary database so that those words can be recognized by LexTo. This function would help make great contributions to research which analyzes a specialized or a specific-topic corpus. However, we found that the web-based LexTo which is accessible by public users does not provide this particular function. Other studies have also indicated this limitation of LexTo which only identifies words contained in the reference dictionary (Ousirimaneechai & Sinthupinyo, 2018). The program only works well with well-prepared documents whereas unformatted text such as those with disorganized spaces or indents, Thai numbers or different spelling formats (Chantaraj & Rungrattanaubol, 2020), misspelled words (Chanta & Parrawatpreyakorn, 2013), or medical jargon (Kurdkit et al., 2015) could cause obstacles to the analysis.

3.2.2 Thai Lexeme Analyser (TLex)

Due to the limitations of LexTo, TLex was developed using a machine learning-based approach (ML) with a conditional random field algorithm and an implementation of BEST2009, a 5 million-word corpus, as a sample corpus for the ML training (Haruechaiyasak & Kongyoung, 2009; Ousirimaneechai & Sinthupinyo, 2018). TLex is the most accurate segmentation program as claimed by Hirankan et al. (2013). It was trained on a large corpus in order to facilitate it to read texts, especially those with difficult words which are not identified by LexTo. Yet, some limitations in segmentation was found, e.g., segmenting words in some categories which do not exist in BEST2009 such as social media lists (Ousirimaneechai & Sinthupinyo, 2018).

3.2.3 AntConc

AntConc (Anthony, 2019) is one of the most regularly used tools for processing and analyzing a corpus. Several functions provided in the toolkit include concordances, concordance plots, file views, clusters/n-grams, collocates, word lists, and keyword lists, which are all useful for researchers who are interested in a certain aspect of a corpus. (More information about the use of AntConc can be found in the user handbook which is available on the software's website at https://www.laurenceanthony.net/software/antconc/releases/AntConc343/help.pdf.) In this study, the two main functions of the word list and n-gram were used. The word list function was employed to generate a list of words contained in a corpus together with the frequency of each word. Moreover, a list of word clusters or multiword units which commonly occurred in the corpus were generated by the n-gram function. Users can specify the number of words in each cluster depending on their own objectives; 2-, 3- and 4-word clusters are generally implemented.

3.3 Two-level segmentation

The process of two-level segmentation includes segmenting Thai text into the smallest units and then creating longer units (clusters of 2-, 3- and 4-words). In this study, a glossary of terms was also used as the basis for identifying these word clusters. To apply the two-level segmentation method, several steps were conducted.

Level 1: Manual segmentation and data editing

In this study, the main goal for word segmentation is to segment the interview data into units appropriate to the purpose of the study. All Thai interview transcriptions were segmented into smallest units manually and were then edited for consistency of word use in the interview corpus.

The text was segmented into the smallest units following a set of criteria, but not the expected units at this stage (including terms and longer units of words treated as single lexical units) to reduce inconsistencies which might arise from multiple people doing the manual segmentation. This segmenting process cannot be done via an automatic program since the programs available to us also combine some single units into one word automatically. The combination of some single units, especially those associated with the important concepts in this study will affect the frequency data of those concepts and is likely to affect the analysis as a whole. Therefore, the smallest units will be combined into more meaningful units using the n-gram technique afterwards.

Segmentation:

- All words were segmented into the smallest units even though the segmentation returns single words with new meanings. For example, a compound word แบบสอบถาม (questionnaire) was segmented as three single words: แบบ (type) สอบ (test) ถาม (ask).
- Words with prefixes which change a part of speech of that root word were tokenized as two words, such as การวัด (measurement) as การ (prefix) วัด (measure)

Editing:

- All abbreviations were replaced with their full words. For example, 'ป.ตรี' replaced with 'ปริญญาตรี' which means undergraduate level of university.
- Proper nouns that are commonly addressed in English abbreviations can be displayed in English, such as BBL (Brain-based Learning), DLTV (Distance Learning Television)
- Repeated words presented by the Thai repeating symbol (ๆ) were transcribed as two words, e.g. มาก ๆ as มาก มาก (really)
- Borrowed words from other languages were transcribed as words in Thai, such as คอมพิวเตอร์ (computer) and มัลติเพิ่ลช้อยส์ (multiple choices).

Level 2: N-gram technique

The segmented data processed in Level 1 was processed by AntConc using an n-gram to form word clusters.

The N-gram function is one type of data transformation employed in several studies in Thai NLP (e.g., Ousirimaneechai & Sinthupinyo, 2018; Thumrongluck & Mongkolnavin, 2011). This technique slices a text into small groups, called grams, which are consistent in length. For example, the text is sliced into 2-length grams called bigram or 2-grams (Cavnar & Trenkle, 1994). In the context of corpus linguistics, an n-gram is a sequence of a number of items and sometimes refers to as multiword units. Generating n-grams could help a researcher make unnoticed units more salient (Sketch Engine, 2016).

In this study, n-grams can identify chunks of text which are associated with the focus of the study – in this case, the assessment practices used by school teachers. The clusters of words generated by the n-gram technique provide a meaningful and insightful point of view for the researchers due to their overall context of each single word. For instance, some words which were not tokenized properly in a specific context, such as คะแนนเก็บ (continuous assessment) can be tokenized by LexTo as two separate words – คะแนน (score) and เก็บ (to keep) which create a different meaning. In this case, the n-gram technique could enhance the accuracy of tokenized items processed by LexTo (Angsumalee et al., 2016). Although tri-grams have been reported as best for Thai written texts (Thangthai & Jaruskulchai (2003) as cited in Angsumalee et al., 2016), four-grams were likely to be the best number for this study's preliminary analysis. Steps for generating n-grams are presented below.

1. Upload segmented text prepared in Level 1 to AntConc.
2. Perform 2- to 4-grams to seek for the target results that best address the project focus or the results which indicate certain information about the corpus.

3.4 Procedures

1. The text was a transcript of Thai language interview data written consecutively without spaces between word units. This sample text was segmented employing three different segmentation methods. The segmented texts were saved as three separated text files (corpora) as follows: 1) a corpus segmented by LexTo, 2) a corpus segmented by TLex, and 3) a corpus segmented by 2-level segmentation.
2. The three segmented corpora were implemented as input texts for AntConc in order to generate three-word frequency lists.
3. The three-word frequency lists of texts prepared by the three methods were compared. To reduce rare words which are less likely to present key concepts, only the units with a minimum frequency of five were selected.

These procedures are presented in Figure 1.

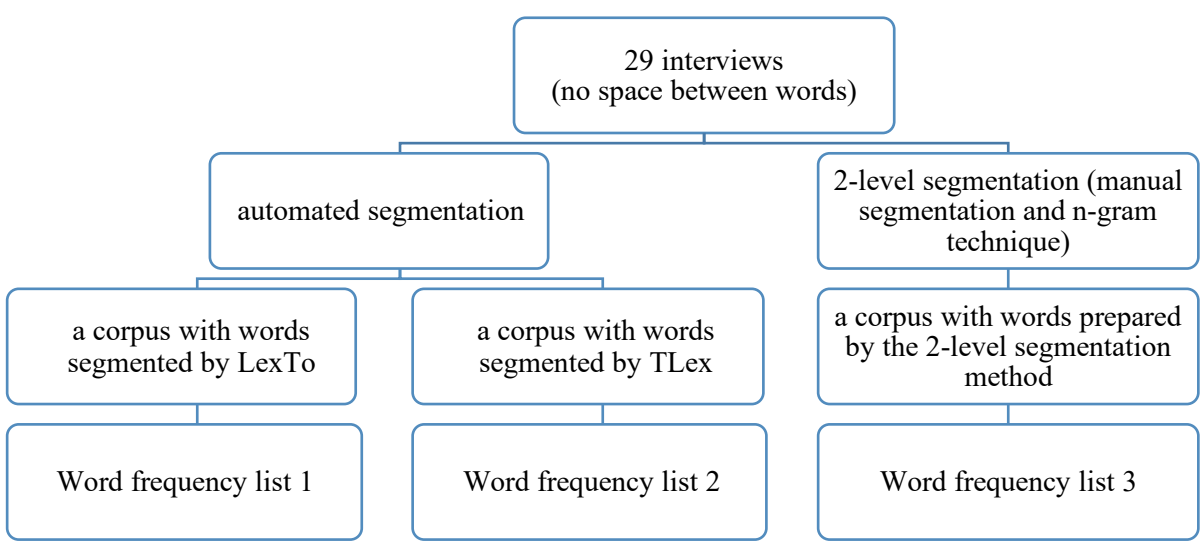

Figure 1. Procedures to generate word frequency lists of three corpora segmented by LexTo, TLex, and 2-level segmentation.

3.5 Data analysis

The comparative outputs of the three methods were investigated both quantitatively and qualitatively.

1. The number of segmented units (types and tokens) processed by three methods were compared. The results will show whether the number of words is manageable for further analysis. We aim to find the method which provides a more manageable number of words since these words will be further investigated, such as by analyzing the concordance lines, to see how each word unit is mentioned in the interviews as well as its context, or to identify themes of teachers' thoughts which emerge from the interviews.
2. Forms (types) of the top 50 segmented units were compared across the three lists. The word forms could help researchers decide if the method returns useful segmented units for further analysis of a specialized corpus.
3. Ranks of the target words throughout the three lists were studied to identify if the expected words are ranked higher or lower in each list. Based on the research purposes and aspects being investigated, a glossary of terms is employed as a benchmark for evaluating the quality of segmented units with regards to how well the segmented units match the expectation. Researchers often focus on the top ranked words, e.g., the top 20, 50, or 100 words. Thus, it is more likely that high-ranked words are considered to be more important.

4. Results

Outputs created by the three segmentation methods are presented in three main aspects: the number of segmented units, the forms of segmented units, and the ranks of words contained in a glossary in the three lists.

4.1 The number of segmented units

The number of types not only suggests the number of potential words that could be further studied, it also reflects the length of words. A fewer number of segmented words in one list would suggest that the segmented units might be longer when compared with those in the other lists. The 2-level segmentation method produced the smallest numbers of units (344 types) suggesting that it contained the longest units compared with the other two methods. For example, after the three methods have been applied to segment "ข้อสอบปลายภาคส่วนใหญ่ก็จะเป็นมัลติเพิลช้อยส์" (Most final exams are multiple choice.), the word "มัลติเพิลช้อยส์" (multiple choice) was segmented into 2 short words (มัลติเพิล (multiple) and ช้อยส์ (choice)) by LexTo, 3 short words (มัลติ (multi), เพิล (multiple), and ช้อยส์ (choice)) by TLex, and 1 longest word by the 2-level segmentation (multiple choice (มัลติเพิลช้อยส์)), as presented below.

Segmented by LexTo: ข้อสอบ | ปลาย | ภาคส่วน | ใหญ่ | ก็ | จะเป็น | *มัลติเพิล* | *ช้อยส์* |
Segmented by TLex: ข้อสอบ | ปลาย | ภาค | ส่วน | ใหญ่ | ก็ | จะ | เป็น | *มัลติ* | *เพิล* | *ช้อยส์*
Segmented by 2-level segmentation: ข้อสอบปลายภาค | ส่วนใหญ่ | ก็ | จะ | เป็น | *มัลติเพิลช้อยส์*

Table 1. Number of word types and tokens

Methods	Types of the methods	Wordlists	Types	Tokens
Method 1 (LexTo)	Automated	Word frequency list 1	6,020	144,097
Method 2 (TLex)	Automated	Word frequency list 2	4,772	165,122
Method 3 (2-level segmentation)	4-grams	Word frequency list 3	344	2,778

In Table 1, the number of types and tokens which occur at least five times in the corpus was counted and presented. The results show that the self-designed method returned 344 types which is more manageable than the units prepared by LexTo and TLex.

4.2 Forms of segmented units

In order to investigate whether there is any distinction in the top units in the three lists which could facilitate researchers in better understanding the texts, the top 50 units of the three lists were compared. Only those related to the main focus of the study are presented in Table 2.

Table 2. Words ranked in the top 50 which are related to the sample texts.

LexTo			TLex			2-level segmentation		
Dictionary based (Longest matching algorithm)			Machine learning (Conditional Random Fields algorithm)			Manual segmentation and n-gram		
Rank	**Words**	**Freq**	**Rank**	**Words**	**Freq**	**Rank**	**4-grams**	**Freq**
7	เด็ก (student)	2191	15	เด็ก (student)	2247	9	การ เรียน การ สอน (learning and teaching)	23
25	อาจารย์ (teacher)	1007	24	เรียน (study)	1518	12	วัด และ ประเมิน ผล (evaluation and assessment)	21
26	คะแนน (score)	943	30	คะแนน (score)	1128	17	วัด ผล ประเมิน ผล (evaluation and assessment)	19
37	สอน (teach)	658	33	อาจารย์ (teacher)	1015	19	การ วัด และประเมิน (evaluation and assessment)	18

45	ครู (teacher)	598	40	ครู (teacher)	876	24	เอ บี ซี ดี (a b c d)	16
48	โรงเรียน (school)	527	48	วัด (measure)	719	27	การ วัด ผล ประเมิน (evaluation and assessment)	15
49	ข้อสอบ (test)	521				28	การ ออก ข้อ สอบ (designing test)	15
						31	ฟัง พูด อ่าน เขียน (listen, speak, read, write)	14
						42	การ วัด ประเมิน ผล (evaluation and assessment)	11
						49	การ ทำ ข้อ สอบ (take the test)	10

While lists 1 and 2 show a few words related to the study of assessment practices used in Thai schools, those words are considered general words in English teaching, e.g., เด็ก (student), อาจารย์ (teacher), คะแนน (score), โรงเรียน (school), ข้อสอบ (test), and วัด (measure). Words which are more closely related to the research objective can be found in List 3, e.g., การเรียนการสอน (learning and teaching), วัดและประเมินผล (evaluation and assessment), การออกข้อสอบ (designing test), ฟังพูดอ่านเขียน (listen, speak, read, write), and การทำข้อสอบ (take the test). More informative contexts through the analysis of the concordance lines containing these words can be easily inspected.

4.3 Ranks of the expected words

A glossary of terms created by the researchers based on the research purposes and aspects being investigated was utilized as a benchmark for evaluating the quality of segmented units with regards to how well they corresponded with expectations.

When items are ranked based on frequency, the rank implies the accessibility of the word to the researchers. Words with higher ranks will be at the top of the list and therefore easier to identify. So, in this section, the glossary of related words was used as a benchmark for identifying useful words in these lists. Ranks and frequencies of some words found in the texts are presented in Table 3.

Table 3. Examples of five words related to the focus of the interviews in the three lists

Word	LexTo			TLex			2-level segmentation		
	Rank	Word	(f)	Rank	Word	(f)	Rank	Word	(f)
ประเมิน (assess)	165	ประเมินผล	147	109	ประเมิน (assess)	311	12	วัด และ ประเมิน ผล	21
	217	ประเมิน	104				17	วัด ผล ประเมิน ผล	19
	553	การประเมินผล	32				19	การ วัด และ ประเมิน	18
	613	การประเมิน (assess)	28				27	การ วัด ผล ประเมิน	15
							42	การ วัด ประเมิน ผล (evaluation and assessment)	11
ข้อสอบ (test)	49	ข้อสอบ (test)	521	59	ข้อสอบ (test)	556	49	การ ทำ ข้อ สอบ (doing a test)	10
	321	ออกข้อสอบ (design a test)	70	842	ตัวข้อสอบ (test itself)	10	28	การ ออก ข้อ สอบ (designing test)	15
				1346	ข้อสอบถาม (question)	4	110	ทำ ข้อ สอบ ไม่ (do a test, no)	7
				1823	ข้อสอบคุณ (your test)	2	159	จะ ทำ ข้อ สอบ (will do a test)	6
							244	ข้อ สอบ เดียว กัน (the same test)	5
ชี้ (point)	302	ชี้วัด (indicate)	75	281	ชี้ (point)	83	109	ตัว ชี้ วัด ตัว (indicator)	7
	1686	ตัวบ่งชี้ (indicator)	6				54	ตาม ตัว ชี้ วัด (due to the indicator)	10
	2153	บ่งชี้ (point out)	4				162	ตัว ชี้ วัด ที่ (indicator that)	6
สาระ (content)	275	สาระ (content)	84	308	สาระ (content)	72	71	สาระ การ เรียน รู้ (subject matter)	9
				2120	สาระวิชา (subject)	2	176	หัว หน้า กลุ่ม สาระ (head of department)	6
							326	แต่ ละ กลุ่ม สาระ (each department)	5
หน่วย (unit)	262	หน่วย (unit)	87	246	หน่วย (unit)	97	72	หน่วย การ เรียน รู้ (learning unit)	9

Overall, the five words presented in Table 3 appeared at the lower ranks (i.e., higher numbers) in Lists 1 and 2 compared with List 3. This means that useful words are more easily noticed in List 3. For example, ประเมิน (evaluate) can be found highly ranked at rank 12 on List 3 but found much lower in the list at ranks 165 and 109 on Lists 1 and 2 which were segmented by LexTo and TLex, respectively.

Some words can be found at similar ranks in all three lists. For example, ข้อสอบ (test) is found in rank 49 (LexTo output), 59 (TLex output), and 29 and 48 in list 3. However, the frequency of this word in lists 1 (521) and 2 (556) are very different from list 3 (10 and 15). The frequency also represents the number of concordance lines of each word which could facilitate researchers when interpreting the data. Therefore, it could be summarized that the number of words as well as the frequency of words which affect the number of concordance lines to be examined are more manageable on List 3 rather than the other two lists.

5. Recommendations for segmenting Thai words

This study compared the outputs of word frequency lists of the Thai corpus of interviews which was segmented by three different methods. The comparison shows the 2-level segmentation method worked best since the outputs in terms of the number of words are the most manageable and the most relevant to the focus of the study. The manageable number of word types will facilitate analysts to conceptualise the content of the corpus when concordance lines are analyzed for more details. The n-gram technique also allows the formation of words which enhance the degree of relevance to specific concepts in the corpus and could address the focus of the study. These words were also found in the top ranks on the word list.

The word frequency lists generated from the corpora segmented by the automated methods, namely LexTo and TLex, contained many more word types. Although word types in the list prepared by LexTo were fewer than TLex, the word forms returned as the outputs were not distinctive when focusing on the first 50 words. Six and seven words were found most relevant to the focus of the study on assessment. To identify important concepts (e.g., the perspectives of the teacher interviewees in this study), the analysts must study further by examining more than 50 words.

When choosing any of the existing segmentation software tools, the researchers should consider the approaches and methodology used for designing the software, e.g., RB technique, DB technique, and MLB technique (Tapsai et al., 2021). Although in this study the two programs which employed DB and MLB techniques show little difference in their outputs, in terms of word forms specific to the corpus of interviews on assessment, considering an input dictionary and a training corpus used by a segmentation program as the basis for segmentation is still recommended. This is because specific words contained in a specific corpus affect a segmentation program's capability, in other words, the proportion of segmented words and unknown words identified by each program.

Ideally, it is recommended that analysts choose a program whose input data is similar in genre to the specific corpus being studied. This would produce more satisfactory results for further detailed analysis of a corpus. However, because the number of segmentation programs which are available for general users are limited, the analysts should examine the segmentation methods each available program implements and then do a segmentation process with sample data before making a decision on which of those programs to use for their study.

The results of the comparisons across three methods provide some guidelines for researchers who collect data in Thai and wish to use corpus-based analysis as follows.

This study argues for the 2-level segmentation as the most appropriate method for segmenting a specific corpus, under the condition that the number of word segmentation programs which are available for general users are limited. However, since it is time-consuming, especially in the manual segmentation in the first stage, analysts are recommended to consider corpus size and time as the two main factors whether the manual segmentation is reasonable for their study. Based on these two factors, the recommendations given below are for those who will use and will not use the 2-level segmentation.

First, researchers who decide to use the 2-level segmentation can follow the following steps.

1. Segment Thai texts into smallest units. In doing this, they should clearly define what a smallest unit means in the study. Guidelines for a manual segmentation should also be well-prepared. The definition and the guidelines will help the manual segmentation remain consistent.
2. Edit the segmented corpus using a glossary of terms related to the focus of the study as the basis for segmentation (optional).
   3.Generate 3- and 4-grams of the segmented corpus and do further detailed analysis of concordances of the 3- and 4-gram units.

Second, for researchers who decide not to use the 2-level segmentation, they may aim to use existing and available automated segmentation programs. They should also have some background on the approaches used as the basis for the program to segment word units. They might consider following the steps below.

1. Segment the Thai corpus using some existing segmentation programs. If possible, the researchers should do a trial segmentation by segmenting sample data employing more than one program. They can generate a frequency list of segmented words by those different programs, then decide which of them returns word types that effectively address the focus of their study.
2. Segment a corpus by using the program which provides satisfactory outputs.
3. Edit the segmented corpus using a glossary of terms related to the focus of the study as the basis for segmentation (optional).
4. At this stage, there are two options to use the list of segmented words for further analysis.
   Option 1: Take many words into consideration, e.g., the top 250 or top 500 for further detailed analysis.
   Option 2: Generate 3- and 4-grams of the segmented corpus and do further detailed analysis. By choosing this option, the researchers might be able to study more details of fewer words than those for the first option, e.g., the top 50 or top 100 words.

By this second choice for those who decide not to do a manual segmentation, they can also create a segmented corpus with the units appropriate for the purpose of the study. If they find the segmented outputs from stage 2 satisfactory, this second choice is recommended as it can be done with less time and effort than the choice which requires manual segmentation. However, as far as we have concerned, segmented units produced by each of few programs available as a freeware do not address the focus of the study well enough. In that case, the 2-level segmentation is the right choice.

As mentioned at the beginning, there is no perfect method or program that can yield the best results for all studies, but the one which returns the most satisfactory results serving the objective of the study is what each researcher is seeking. It is hoped that this study provides insights into the criteria which can be used to select the most appropriate method.

References

Angsumalee, B., Sotthisopha, N. & Vateekul, P. (2016). A Framework to Detect Unqualified Restaurant Reviews. *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*, 115 – 120.

Anthony, L. (2019). AntConc (Version 3.5.8). https://www.laurenceanthony.net/software/antconc/

Apisuwankun, P. & Mongkolnavin, J. (2013). Opinion Strength Identification in Customer Review Summarizing System Using Association Rule Technique. *In Proceedings of the International Conference on E-Technologies and Business on the Web (EBW2013)*, 16-21.

Aroonmanakun, W. (2002). Collocation and Thai word segmentation. *In Proceedings of the Fifth Symposium on Natural Language Processing & the Fifth Orientation COCOSDA Workshop*, Bangkok, Thailand, 68 – 75.

Aroonmanakun, W. (2007). Thoughts on word and sentence segmentation in Thai. *In Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15*, 85–90.

Aroonmanakun, W. (2018). การตัดคำภาษาไทย : ตัดคำอย่างไร [Thai Word Segmentation: How to segment]. https://awirote.medium.com/การตัดคำภาษาไทย-ตัดคำอย่างไร-86b007fb648c

Bennett, G. R. (2010). *Using Corpora in the Language Learning Classroom: Corpus linguistics for teachers*. University of Michigan Press.

Bheganan, P., Nayak, R. & Xu, Y. (2009). Thai Word Segmentation with Hidden Markov Model and Decision Tree. *Pacific-Asia Conference on Knowledge Discovery and Data Mining Springer*, 74-85.

Cavnar, W. B. & Trenkle, J. M. (1994). N-gram-based text categorization. *Ann Arbor Mi, 48113 (2)*, 161-175.

Chamyapornpong, S. (1983). *A Thai Syllable Separation Algorithm*. Master thesis, Asian Institute of Technology, Bangkok.

Chanta, S. & Porrawatpreyakorn, N. (2013). A Web News Information Classification and Retrieval System using Multilayer Perceptron Neural Network, *Information Technology Journal, 9*(2), 15-19.

Chantaraj, P. & Rungrattanaubol, J. (2020). Applied Information Extraction technique for extracting the king name who build a temple in Lanna Kingdom from historical documents. *Information Technology Journal, 16*(1), 24-33.

Egbert, J., Larsson, T., & Biber, D. (2020). Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User (Elements in Corpus Linguistics). *Cambridge University Press*. doi:10.1017/9781108888790.

Haruechaiyasak, C. & Kongyoung, S. (2009). TLex: Thai lexeme analyser based on the conditional random fields. *In: InterBEST 2009 Thai Word Segmentation Workshop*, 13–17.

Haruechaiyasak, C., Kongyoung, S., & Dailey, M. (2008). A comparative study on Thai word segmentation approaches. 2018 *5th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology, 1*, 125-128. doi: 10.1109/ECTICON.2008.4600388.

Hasko, V. (2013). Qualitative corpus analysis. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 4758 – 4764). Blackwell.

Hengsanankun, T., & Namburi, A. (2020). Improving Thai Word Segmentation using HMM: A Case Study of Sentiment Analysis. *2020 24th International Computer Science and Engineering Conference (ICSEC)*, 1-6.

Hirankan, P., Suchato, A. & Punyabukkana, P. (2013). Detection of Wordplay Generated by Reproduction of Letters in Social Media Texts, *10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 6-10.

Huntley, S. J., Mahlberg, M., Wiegand, V., van Gennip, Y., Yang, H., Dean, R. S., & Brennan, M. L. (2018). Analysing the opinions of UK veterinarians on practice-based research using corpus linguistic and mathematical methods. *Preventive Veterinary Medicine,* 150, 60-69.

Jucksriporn C., & Sornil O. (2011). A Minimum Cluster-based Trigram Statistical Model for Thai Syllabification. *Proc. International Conference on Intelligent Text Processing and Computational Linguistics, Berlin, Heidelberg*, 493-505.

Khowrurk, P. & Kongkachandra, R. (2020). A Machine Learning Approach for the Classification of Methamphetamine Dealers on Twitter in Thailand. *15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1-5. doi: 10.1109/iSAI-NLP51646.2020.9376817.

Kruengkrai, C., Sornlertlamvanich, V., & Isahara, H. (2006). A Conditional Random Field Framework for Thai Morphological Analysis. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Kurdkit, S., Aun-a-nan, A. & Meesad, P. (2015). การจำแนกความน่าเชื่อถือของเนื้อหาในเว็บไซต์ภาษาไทยด้านมะเร็งโดยใช้ CancerDic+ [Classification of Reliable Content on Cancer Thai Website using CancerDic+]. *Journal of Information Science and Technology, 5*(2), 34-43.

Scott, M. (2020). *WordSmith Tools version 8*, Stroud: Lexical Analysis Software.

Sketch Engine. (2021). *n-gram*. https://www.sketchengine.eu/my_keywords/n-gram/

Lu, X. (2020). Corpus Linguistics and the Description of English. *Journal of English Linguistics, 48*(1).97-104. https://doi.org/10.1177/0075424219896604

Mahatthanachai C., Malaivongs K., Tantranont N., & Boonchieng E. (2015). Development of Thai Word Segmentation Technique for Solving Problems with Unknown Words. *2015 International Computer Science and Engineering Conference (ICSEC)*. IEEE, 1–6.

Marchi, A., & Taylor, C. (2009). If on A Winter's Night Two Researchers...: A Challenge to Assumptions of Soundness of Interpretation. CADAAD *Journal [Critical Approaches to Discourse Analysis across Disciplines], 3*(1), 1-20.

Nararatwong, R., Kertkeidkachorn, N., Cooharojananone, N., & Okada, H. (2018). Improving Thai Word and Sentence Segmentation Using Linguistic Knowledge. *IEICE Transactions on Information and Systems, 101-D*, 3218-3225.

NECTEC. (2016). *LexTo Thai Lexeme Tokenizer*. http://www.sansarn.com/lexto/

Ousirimaneechai, N. & Sinthupinyo, S. (2018). Extraction of Trend Keywords and Stop Words from Thai Facebook Pages Using Character n-Grams. *International Journal of Machine Learning and Computing, 8*(6), 589-594.

Paripremkul, K., & Sornil, O. (2021). Segmenting Words in Thai Language Using Minimum Text Units and Conditional Random Field. *Journal of Advances in Information Technology, 12*, 135-141.

Poowarawan, Y. (1986). Dictionary-based Thai Syllable Separation. *In Proceedings of the Ninth Electronics Engineering Conference*, 409–418.

Rayson, P. (2009). *Wmatrix: a web-based corpus processing environment*, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Tanantong, T., Kreangkriwanich, S., & Laosen, N. (2020). Extraction of Trend Keywords from Thai Twitters using N-Gram Word Combination. *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 320-323.

Tapsai, C., Unger, H & Meesad, P. (2021). *Thai National Language Processing: Word Segmentation, Semantic Analysis, and Application*. In J. Kacprzyk (Ed.) Springer.

Teddiman, L. (2009). Contextuality and Beyond: Investigating an Online Diary Corpus. *Proceedings of the International AAAI Conference on Web and Social Media, 3*(1), 331-333.

Thairatananond, Y. (1981). *Towards the Design of a Thai Text Syllable Analyzer*. Master Thesis of Science, Asian Institute of Technology.

Thangthai, A., & Jaruskulchai, C. (2003). Text Summarization using Singular Value Decomposition for Thai. *In Proceedings of the 7th National Computer Science and Engineering Conference, Chonburi, Thailand, October 28-30.*

Thangthai, K., Chotimongkol, A., & Wutiwiwatchai, C. (2013). A Hybrid Language Model for Open-vocabulary Thai LVCSR. *INTERSPEECH*.

Thumrongluck, T. and Mongkolnavin, J. (2011). การพัฒนาระบบสรุปบทวิจารณ์สินค้าภาษาไทยโดยผู้บริโภคแบบอัตโนมัติ [The Development of an Automated System for Summarizing Product Reviews of Thai Consumer].*Chulalongkorn Business Review. 33*(2): 40-62.

Touileb, S. & Salway, A. (2014). Constructions: A New Unit of Analysis for Corpus-based Discourse Analysis. *28th Pacific Asia Conference on Language, Information and Computation*, 634-643. https://aclanthology.org/Y14-1072.

Unlee, P. & Seresangtakul, P. (2016). Thai to Isarn Dialect Machine Translation Using Rule-based and Example-based. *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1-5. doi: 10.1109/JCSSE.2016.7748892.

Vichianchai, V. (2014). The Comparison of Thai Word Segmentation with Thai Writing Structures and Syllable Structures. *J Sci Technol MSU, 33*(5), 503-509.

Watson Todd, R., Tepsuriwong, S., Trakulkasemsuk, W., Jaturapitakkul, N., Pojanapunya, P., & Chanchula, N. (2020). โครงการการสำรวจแนวปฏิบัติด้านการวัดและประเมินผลรายวิชาภาษาอังกฤษที่ใช้ในโรงเรียนไทยในปัจจุบันโดยเน้นแนวปฏิบัติที่ส่งผลกระทบเชิงบวก [A survey of current English language assessment practices at schools throughout Thailand focusing on practices with positive washback]. Report submitted to the National Research Council of Thailand.

Weetman, K., Dale, J., Scott, E. et al. (2020). Adult Patient Perspectives on Receiving Hospital Discharge Letters: A Corpus Analysis of Patient Interviews. BMC Health Serv Res 20, 537, https://doi.org/10.1186/s12913-020-05250-1