

Unearthing Collocations of Keywords in News Coverage on the Ancient Town of Si Thep, Thailand’s Seventh UNESCO World Heritage Site: A Corpus-based Study

NATEETHORN NARKPROM

Faculty of Education, Phetchabun Rajabhat University, Thailand

Author email: nateethorn.nar@pcru.ac.th

Article information	Abstract
Article history: Received: 7 Apr 2024 Last revised: 11 Apr 2026 Accepted: 12 Apr 2026 Available online: 21 Apr 2026	<i>The ancient town of Si Thep in Phetchabun, Thailand, was declared a UNESCO World Heritage Site in 2023, boosting tourism and raising the demand for effective English communication in the area. This study’s objectives are to analyze keywords and their collocational patterns in a specialized corpus of news coverage on Si Thep (hereafter: the ST corpus), comprising 30,311 words, gathered from 76 English language news articles, published between April 2015 and February 2024. The results reveal three grammatical categories of keywords: nouns (N = 81), adjectives (N = 12), and verbs (N = 7). The keywords in the ST corpus are integrated into 172 types of lexical collocations, grammatically classified as Adjective + Noun (43.60%), Noun + Noun (37.79%), Noun + Verb (7.56%), Noun + of Noun (6.40%), Verb + Noun (4.07%), and Verb + Adverb (0.58%). Notably, the keywords “local”, “cultural”, “ancient”, “UNESCO”, and “Dvaravati” are associated with the highest number of collocates in the ST corpus. Various pairs of keywords and collocates also represent nine topics of the ancient town of Si Thep, as reported in the news, e.g., Si Thep Overwhelmed by Overtourism, Si Thep Threatened by Oil Drilling. This study may offer alternatives for teaching collocations of specialized vocabulary and developing instructional materials, reflective of authentic language use, as well as methodological guidelines for English teachers in other locations designated as UNESCO World Heritage Sites.</i>
Keywords:	
Collocations	
Corpus	
Keywords	
UNESCO world heritage site	

INTRODUCTION

The district of Si Thep in Phetchabun province, Thailand, has captured media attention in recent years as its main attraction, the Si Thep Historical Park, was officially declared as a World Heritage Site by the United Nations Educational, Scientific and Cultural Organization (hereafter: UNESCO) in September 2023 under the formally assigned name “The Ancient Town of Si Thep and its Associated Dvaravati Monuments” (UNESCO World Heritage Convention, 2023). Formerly unrecognized to foreign visitors and underappreciated by Thais, this small district has since emerged as one of the enticing tourist sites in Thailand, unprecedentedly boosting the local economy through mass tourism and generating financial income for its residents in the process (“Historical Park Grapples with Crowds”, 2023). Located roughly

350 kilometers north of Bangkok, where both domestic and international flights are operated daily, the Si Thep district is now evolving into a must-visit destination for sightseers, reaping the benefits of its newfound reputation and reclaiming its past glory as a historically “significant community” and “commercial hub for trading with external communities” (The Office of National Resources and Environmental Policy and Planning and The Fine Arts Department of Thailand, 2022, p. 61).

As a means of global communication, English plays an undisputed role in various industries worldwide, namely in tourism (Fujita, 2019). Thailand, for instance, is one of those countries whose domestic economy substantially depends on tourism as the number of inbound foreign visitors surges yearly, creating a high demand for effective English communication skills among workforces in the tourism industry. Nevertheless, local tourism personnel in Thailand often struggle to communicate competently with international visitors, citing a lack of grammatical knowledge and specialized vocabulary relevant to the tourist spots where they work as one of the main challenges (Namtapi, 2022).

Collocation is undeniably one of the keys to successful and fluent English communication for non-native speakers (Jaén, 2009; Wray, 2002). It is defined by Nesselhauf (2005, p. 25) as “a type of word combination in a certain grammatical pattern”. Unfortunately, collocation is also a problematic area that Thai learners and users of English struggle to overcome due to 1) inadequate instruction of collocations in Thai education (Boonyarattanasoontorn et al., 2020); and 2) negative transfer from L1 Thai (Boonraksa & Naisena, 2022).

Collocation is, in addition, one of the primary focuses in corpus linguistics, as the application of corpus software enables researchers to identify collocations from a large collection of textual data using empirical quantitative evidence, rather than human intuition, which is considered “a poor guide to collocation” (McEnery et al., 2006, p. 83). To address the specific needs of Thai learners and users of English, several recent studies have utilized the capability of corpus software to identify keywords or groups of vocabulary “which occur with unusual frequency in a given text” (Scott, 1997, p. 236) and extract their statistically associated collocates from specialized corpora. These specialized corpora were compiled from authentic texts in specific fields, including *insurance* (Khamphairoh & Tangpijaikul, 2012), *nursing* (Trinant & Yodkamlue, 2019), *tourism* (Trinant & Kijpoonphol, 2021), *logistics* (Suraprajit, 2022), *business* (Sukman et al., 2022; Sukman & Tangkiengsirisin, 2024), and *physical education and sport science* (Thaksanan & Kamsa-ard, 2025).

Although these aforementioned studies significantly contribute to addressing the specific needs of Thai learners and users in those particular disciplines and can serve as a foundation for developing instructional lessons, their corpus-informed findings might not entirely apply to individuals within more localized or context-specific settings. This is further reinforced by the fact that most commercial textbooks on English for Specific Purposes (i.e., *tourism*) are designed for generic, rather than local audiences (Namtapi, 2022).

Hence, the ultimate aim of this research is to fill the gap left by previous studies through an investigation of the collocations of keywords in a specialized corpus of news coverage on the

ancient town of Si Thep and its recent designation as a UNESCO World Heritage Site, thereby adding further insights to the existing body of knowledge. The results of this study can be utilized as a basis for vocabulary teaching materials authentically created for and culturally related to Thai learners or users of English who engage in tourism activities in the Si Thep district of Phetchabun province.

Additionally, Thailand has numerous historical sites nationwide. Along with the ancient town of Si Thep in Phetchabun, there are four other UNESCO World Cultural Heritage Sites: the Historic City of Ayutthaya, the Historic Town of Sukhothai, the Ban Chiang Archaeological Site, and, most recently, as of July 2024, Phu Phra Bat in Udon Thani (“Phu Phra Bat Named World Heritage Site”, 2024), which flourished during the Dvaravati period, similar to the ancient town of Si Thep. As a result, this study may offer methodological guidelines and pedagogical insights for teachers in broader contexts of English language teaching who wish to develop instructional materials on collocations of specialized vocabulary based on corpus-informed findings, representative of their local heritages and customized for their students. The research questions are as follows:

1. What keywords appear in news coverage on the ancient town of Si Thep, Thailand’s seventh UNESCO World Heritage Site?
2. What are the most common types of collocation of the keywords?
3. Which keywords have the highest number of collocates?
4. In what topics on the ancient town of Si Thep do the collocations of keywords occur?

LITERATURE REVIEW

Keywords

According to Nation (2013), words can be divided into four categories: high-frequency, academic, technical, and low-frequency vocabulary. The high-frequency and academic vocabularies are also referred to as the New General Service List (hereafter: NGSL) and the Academic Word List (hereafter: AWL), respectively. The NGSL comprises 2,368 words, covering 90% of general English (Browne, 2014), while the AWL contains 570 word families that frequently appear in academic texts, providing 10% coverage (Coxhead, 2000). Excluded from the NGSL and AWL, technical vocabulary consists of subject-specific words, constituting approximately 20 to 30% of coverage in specialized texts (Chung & Nation, 2003). Low-frequency vocabulary comprises non-academic or non-technical words, accounting for only 5% of academic texts.

Khamphairoh and Tangpijaikul (2012) assert that the importance of words should not be based solely on their frequency of use. This is further supported by the notion that technical vocabularies relevant to an individual’s specific needs might be considered “low-frequency words” (Nation, 2013, p. 20). The concept of keywords in corpus linguistics values these intermingled aspects by taking into account not only frequency, but also keyness, or “a measure of saliency” (Baker, 2006, p. 125). The saliency of keywords, or keyness, is

traditionally measured by comparing the frequency of the same words in the target corpus and reference corpus, along with the running words of the two corpora, using statistical measurements such as Log-likelihood, hereafter: LL (Szudarski, 2023). Each word with a probability value, statistical threshold of keyword, ranging from less than 0.05 to 0.0001 is considered a significant keyword worth reporting (Baker, 2006). Other context-specific thresholds, potentially influencing the keyword results, include the type of reference corpus (i.e., *specialized* versus *general*) and various cut-off rankings, such as the top 25, 50, 100, and 200 keywords (Pojanapunya & Lieungnapar, 2017).

Three types of keywords include proper nouns; lexical words (i.e., *nouns, verbs, adjectives, adverbs*), and grammatical words (i.e., *articles, prepositions, pronouns*). Lexical words, along with high-frequency proper nouns, are typically most useful as they indicate the content of the texts, representing the “aboutness” of the target corpus (Scott, 2015, p. 236). In addition to providing an effective method for characterizing a text, corpus software capabilities also enable researchers to investigate the relationships between keywords and their nearby collocations by using the keywords as the node word.

Collocation

Collectively, collocation is conceptualized as the co-occurrence of words, unlikely by chance. In a broader sense, Baker (2006) refers to collocation as “the phenomenon of certain words frequently occurring next to or near each other” (p. 96). In the same vein, however, statistically, collocations are defined by Lewis (2000) as “the way in which words co-occur in natural text in statistically significant ways” (p. 132), by O’Dell and McCarthy (2008) as “pairs of words that occur regularly together, with a high degree of probability” (p. 5), and by Bennett (2010) as “the statistical tendency of words to co-occur” (p. 8). In corpus linguistics, statistical methods such as LL and Mutual Information (hereafter: MI) are applied to evaluate the significance and strength of the collocation between two words. The MI score assesses “the chance of two words co-occurring, considering their frequencies of co-occurrence with all other words in the corpus” (Lui & Lei, 2018, p. 6), while LL value is the “tests of statistical significance... employed to identify collocations” (Szudarski, 2023, p. 35).

Although there have been several categorizations of collocations, this study adopts an adaptation of the lexical collocations framework established by Benson et al. (2010) for two major reasons. First, it provides a comprehensive categorization of lexical collocations, often encountered by non-native speakers of English (Shamsudin et al., 2013). Second, it aligns with the study’s objectives, which are to first identify lexical keywords and then investigate their lexical collocates (i.e., *nouns, verbs, adjectives, and adverbs*), in other words, moving structurally from an isolated word to a word-cluster. Unlike several past studies (Sukman et al., 2022; Sukman & Tangkiengsirisin, 2024; Thaksanan & Kamsa-ard, 2025; Trinant & Kijpoonphol, 2021; Trinant & Yodkamlue, 2019), the present study includes one more type of lexical collocation, also proposed by Benson et al. (2010), namely *Noun + of Noun*, into the theoretical framework, thereby totalling seven types of lexical combinations, as subsequently presented in Table 1.

Table 1
Adaptation of Benson et al.'s (2010) framework of lexical collocations (pp. 32–34)

Types	Examples
1. Verb + Noun	compose music, pose a question, exterminate vermin
2. Adjective + Noun	a formidable challenge, a crushing defeat, an implacable foe
3. Noun + Noun	house arrest, jet engine, land reform, aptitude test
4. Noun + Verb	bees buzz, blood circulates, bombs explode
5. Noun + of Noun	a pride of lions, a bit of advice, an act of violence
6. Adverb + Adjective	strictly accurate, sound asleep, keenly aware
7. Verb + Adverb	affect deeply, apologize humbly, appreciate sincerely

Relevant studies

Investigating the collocations of lexical keywords found in a specialized corpus of insurance research articles, Khamphairoh and Tangpijaikul (2012) found that *insurance*, *risk*, and *insurer* appeared in the top three of the 100 keyword list. In descending order of grammatical distribution, 80% were nouns (e.g., *market*), 7% adjectives (e.g., *financial*), 4% nouns/adjectives (e.g., *variable*), 2% nouns/verbs (e.g., *purchase*), and 2% verbs (e.g., *insure*). The most dominant keywords were semantically related to money and its values (e.g., *cost*, *fund*, *asset*) and measures or regulations (e.g., *contract*, *claim*, *average*). The collocations of the keyword *insurance* included *life*, *health*, and *purchase*, forming two-word clusters: *life insurance*, *health insurance*, and *purchase insurance*. Those of *risk* included *low*, *high*, and *morality*, establishing two-word combinations: *low risk*, *high risk*, and *morality risk*, while those of *insurer* involved *life*, *stock*, and *international*, constructing two-word clusters: *life insurer*, *stock insurer*, and *international insurer*.

Trinant and Yodkamlue (2019) studied keywords and their collocates in a specialized corpus of nursing research articles, discovering that, in descending order of distribution, the majority of keywords were grammatically categorized as 64% nouns (e.g., *medication*), 22% adjectives (e.g., *clinical*), 13% verbs (e.g., *conducted*), and 2% adverbs (e.g., *positively*). The prevalent combination types of keywords and their collocates emerged from the data were 41% Noun + Noun (e.g., *care provider*, *health care*), 28% Adjective + Noun (e.g., *mental health*, *clinical practice*), and 11% Noun + Verb (e.g., *patients were admitted*, *children had been diagnosed*).

Examining the collocations of lexical keywords in a specialized corpus of tourism research articles, Trinant and Kijpoonphol (2021) discovered that most keywords grammatically belong to nouns (63%; e.g., *hospitality*), adjectives (18%; e.g., *international*), verbs (16%; e.g., *revisit*), and adverbs (6%; e.g., *frequently*). The dominant collocational types were 46% Noun + Noun (e.g., *tourism industry*, *service failure*), 26% Adjective + Noun (e.g., *outdoor recreation*, *sustainable tourism*), and 9% Noun + Verb (e.g., *employees need*, *tourists seek*), respectively.

Suraprajit (2022) studied keywords and their collocates in a specialized corpus of logistics magazines. The most frequent keywords were categorized as nominal (e.g., *cargo*, *customers*, *services*), verbal (e.g., *said*, *offer*, *provide*), and adjectival (e.g., *cargo*, *logistics*, *global*) keywords. The prevalent types of grammatical collocations were Noun + Preposition (e.g., *cargo in*), Adjective + Preposition (e.g., *important to*), and Verb + Preposition (e.g., *said with*),

while those of lexical collocations included Verb + Noun (e.g., *provide updates*), Noun + Verb (e.g., *cargo is*), and Adjective + Noun (e.g., *new facility*).

Sukman et al. (2022) examined the collocations of lexical keywords in a specialized corpus of business news released during the years 2020 to 2021, finding that, in descending order of distribution, the majority of keywords appeared as 64% nouns (e.g., *prices*), 8% verbs/nouns (e.g., *sanctions*), 7% adjectives (e.g., *global*), 4% adjectives/nouns (e.g., *executive*), 2% adverbs (e.g., *soaringly*), and 1% verbs (e.g., *explains*). The most common collocational types were 40% Adjective + Noun (e.g., *online retailer, pandemic era*), 35% Noun + Noun (e.g., *energy bills, lockdown restrictions*), and 20% Noun + Verb (e.g., *pandemic continues, prices surged*), respectively.

Exploring the lexical keywords and their collocates in a specialized corpus of business news published between 2021 and 2022, Sukman and Tangkiengsirisin (2024) discovered that most of the keywords belong to the grammatical categories of nouns (83%; e.g., *companies*), adjectives (13%; e.g., *digital*), and verbs (5%; e.g., *started*). The most frequent combinational types were 54% Noun + Noun (e.g., *pandemic restrictions, supply shortages*), 27% Adjective + Noun (e.g., *hybrid workers, digital assets*), and 20% Noun + Verb (e.g., *economy rebounded, restrictions were lifted*), 4% Verb + Noun (e.g., *hit record, increase productivity*), and 1% Verb + Adverb (e.g., *working remotely, hit hard*).

Most recently, Thaksanan and Kamsa-ard (2025) investigated keywords and their collocations in a specialized corpus of research journals in the fields of physical education and sport science. The findings showed that the most common keywords were nouns (64%; e.g., *athletes*), verbs (21%; e.g., *practices*), and adjectives (14%; e.g., *physical*). The proportions of collocational patterns were as follows: 50% Noun + Noun (e.g., *fitness levels, body mass*); 28% Adjective + Noun (e.g., *physical literacy, training intensity*); and 17% Noun + Verb (e.g., *participants provided, study highlights*).

The majority of the aforementioned studies applied a similar methodological criterion for the identification of keywords, but their options for identifying collocates varied. To be specific, a cut-off point of 100 keywords was determined and the British National Corpus (BNC) was utilized as a reference corpus (Khamphairoh & Tangpijaikul, 2012; Sukman et al., 2022; Trinant & Kijpoonphol, 2021). Although all of these studies employed MI score to statistically measure the association strength between keywords and collocates, the window-span options ranged from 3L/3R (Sukman & Tangkiengsirisin, 2024; Thaksanan & Kamsa-ard, 2025; Trinant & Kijpoonphol, 2021; Trinant & Yodkamlue, 2019) to 4L/4R (Khamphairoh & Tangpijaikul, 2012; Sukman et al., 2022). Consequently, this difference in methodology might explain the different types and varying proportions of collocations observed across some of these studies. However, some of the methodological choices, such as a cut-off point of 100 keywords and a collocation span of 3L/3R, provide methodological justification for the present study. This will be further elaborated in the following section.

METHODOLOGY

Data collection for the target corpus

The specialized corpus of news coverage on the ancient town of Si Thep (hereafter: ST corpus) was compiled. News articles, reports, and editorial pieces, published online between April 2015 and February 2024 by established English-medium news outlets in Thailand, were manually collected in plain text format (i.e., *.txt*), forming the target corpus of 30,311 running words. It is worth acknowledging that the size of the ST corpus is small, practically due to the limited availability of the data at the time of this research. Nevertheless, methodology-wise, it is considered orthodox and acceptable for specialized and context-specific corpora, which typically range from 20,000 to 200,000 running words (Vaughan & Clancy, 2013). Additionally, not only is a small, specialized, and domain-specific corpus representative of the texts, but it also preserves an intricate connection between language authenticity and context for researchers (Koester, 2010; Vaughan & Clancy, 2013). The detailed description of the corpus is provided in the following table.

Table 2
Description of the ST corpus

News Outlets	No. of Articles	Running Words
Bangkok Post	27	12,932
The Nation	11	4,130
Thai PBS WORLD	9	4,145
The Thaiger	9	2,742
THAI.NEWS	7	2,913
Pattaya Mail Online	7	1,210
Mass Communication Organization of Thailand	3	730
Bangkok Big Ears	2	1,095
Khaosod English	1	414
Total	76	30,311

Research procedure

A corpus-based approach was adopted for this study to identify the keywords and their collocates in the ST corpus, utilizing the built-in functionalities (i.e., *Keyword* and *Collocate*) of the corpus software *AntConc* 4.2.4 (Anthony, 2023). The application of *AntConc* is justified by its established reputation in the landscape of empirical corpus-based research (Khamphairoh & Tangpijaikul, 2012; Sukman et al., 2022; Sukman & Tangkiengsirisin, 2024; Suraprajit, 2022; Thaksanan & Kamsa-ard, 2025; Trinant & Kijpoonphol, 2021; Trinant & Yodkamlue, 2019), availability for free use, and practicality in fulfilling the objectives of the present study.

Identification of keywords

The identification of keywords in the ST corpus was conducted using the Keyword tool provided by *AntConc*. In doing so, the following criteria were predefined: a cut-off point of $p < 0.0001$ for top-100 keywords (15.13 with Bonferroni correction); a minimum frequency of ≥ 5 ; and a minimum range of ≥ 3 . To put it another way, words that meet the calculated keyness values of at least 15.13, indicating the highest level of confidence that a word is

key, and occur at least 5 times in at least 3 texts across the target corpus are considered keywords for this study. Subsequently, the ST corpus was compared against the reference corpus AmE06. Natively included in *AntConc* 4.2.4, the AmE06 corpus is a research-oriented reference corpus of written publications in general American English, totalling 1,017,879 running words (Potts & Baker, 2012). Initially, *AntConc* identified a total of 246 keywords that met the applied criteria. The list was then manually analyzed by the researcher and experts in the field using a “rating scale approach” (Chung & Nation, 2003) to formulate a top-100 keyword list that is manageable, salient, relevant, and representative of the ST corpus. The cut-off point of 100 keywords is among the most commonly used rankings (e.g., *top-25*, *50*, and *200*) for keyword analysis, as it is seen as sizable and sufficient for classifying topics, while simultaneously representing “a variety of word types” in the target corpus (Pojanapunya & Lieungnapar, 2017, p. 117). During this process of selectivity, irrelevant keywords (e.g., *grammatical keywords*) were excluded from further analysis. As a result, the finalized list not only includes lexical keywords but also encompasses high-frequency proper nouns and abbreviations, which are essential for representing the “aboutness” (Scott, 2015, p. 236) and are considered “a key element” (Khamphairoh & Tangpijaikul, 2012, p. 182) of a specialized corpus.

Identification of collocations

Utilizing the Collocate tool within *AntConc*, each keyword from the revised top-100 keyword list was used as the node to identify its associated collocates. To achieve this, the following criteria were applied: the cut-off point of $p < 0.05$ for LL value (≥ 3.84 with Bonferroni correction); a word span from 3L to 3R; a minimum frequency of ≥ 3 ; and a minimum range of ≥ 3 . That is to say, for a word to be considered a collocate of a keyword, it must show a statistically significant association ($LL \geq 3.84$) and be situated within either three words to the left or three words to the right of the node (i.e., *keyword*), forming grammatical combinations based on this study’s adaptation of Benson et al.’s (2010) framework. Additionally, the collocate must occur at least three times in at least three texts throughout the ST corpus. The cut-off point ($p < 0.05$) for LL value was chosen as the threshold due to its conventionality for identifying “statistically significant and linguistically interesting” collocates in specialized corpora (Szudarski, 2023, p. 35). Moreover, in alignment with this study’s research design, the LL is more suitable for extracting combinations of lexical words (i.e., *verbs*, *nouns*, *adjectives*, *adverbs*) or “when bigrams containing grammatical words are ignored” (Evert, 2005, as cited in Lyse & Andersen, 2012, p. 97). The LL value also performs more efficiently “for medium- to low-frequency data” (Evert, 2009, p. 37), thus making it more appropriate than the MI score, given the relatively small size of the ST corpus. Finally, the smaller span size of L3/R3 is more efficient for detecting collocates than that of L5/R5, which, as per Bartsch and Evert (2014), often shows a drop in precision due to its larger context.

RESULTS

The top-100 keywords emerging from the ST corpus are presented along with their ranks and frequencies in descending order of keyness values in Table 3.

Table 3
The top-100 keywords in the ST corpus

Rank	Keyword	Frequency	Keyness	Rank	Keyword	Frequency	Keyness
1	Si Thep	459	3,259.54	51	monument	24	122.89
2	park	319	1,748.32	52	prehistoric	19	117.31
3	ancient	269	1,509.56	53	world	99	115.36
4	historical	294	1,484.48	54	tentative	21	115.14
5	UNESCO	164	1,163.07	55	historic	26	113.71
6	site	201	986.17	56	laterite	16	113.39
7	Phetchabun	135	957.28	57	locals	26	112.40
8	heritage	146	910.23	58	sacred	24	109.15
9	Thailand	131	873.93	59	located	28	108.66
10	World Heritage Site	103	730.26	60	listing	21	108.30
11	Dvaravati	93	659.33	61	cave	25	105.52
12	Khao Klang Nok*	88	623.87	62	Buddha	18	103.78
13	province	95	579.14	63	conservation	20	100.11
14	archaeological	81	574.23	64	zone	26	100.04
15	Thai	67	474.95	65	recognition	26	96.27
16	town	123	439.82	66	climbing	18	93.73
17	cultural	105	408.71	67	influx	18	87.89
18	sites	83	378.53	68	dating	20	87.47
19	Fine Arts Department	52	368.59	69	treasures	15	86.98
20	visitors	62	349.56	70	meeting	33	86.79
21	stupa	46	326.05	71	recognised	12	85.04
22	ruins	52	310.88	72	World Heritage List	12	85.04
23	Khmer	43	304.78	73	moats	12	85.04
24	Khao Thamorrat*	37	262.25	74	Khao Klang Nai*	12	85.04
25	World Heritage Sites	35	248.07	75	km	15	84.11
26	drilling	36	234.18	76	authorities	27	83.81
27	artefacts	33	233.89	77	project	37	78.52
28	FAD	33	233.89	78	unearthed	11	77.96
29	local	86	232.33	79	accommodation	11	77.96
30	tourists	40	219.94	80	ice-cream	11	77.96
31	monuments	31	219.72	81	concerns	28	74.90
32	Buddhist	33	208.78	82	status	29	72.58
33	exploration	38	204.65	83	WHC	10	70.87
34	tourist	35	190.26	84	Shiva	10	70.87
35	listed	38	187.98	85	petroleum	17	70.61
36	Hindu	29	181.40	86	kingdom	20	70.05
37	committee	65	172.30	87	centuries	20	70.05
38	tourism	28	160.78	88	delegation	9	63.78
39	kilometers	28	160.78	89	parks	14	58.70
40	district	44	156.09	90	protect	23	57.86
41	excavation	22	155.92	91	hub	10	57.00
42	buffer	26	153.74	92	culture	31	56.83
43	oil	52	148.59	93	significance	18	56.73
44	archaeologists	20	141.75	94	Mahayana	8	56.70
45	meters	29	138.74	95	concessions	8	56.70
46	area	54	138.66	96	archaeologist	8	56.70
47	stupas	19	134.66	97	session	16	56.52
48	city	74	131.36	98	nomination	14	56.42
49	list	42	129.20	99	provincial	11	56.23
50	sculptures	25	127.09	100	style	28	54.97

Note. *Three major archaeological sites of the ancient town of Si Thep: Khao Klang Nai, translated as “the treasury mountain in the inner town”; Khao Klang Nok, translated as “the treasury mountain in the outer town”; and Khao Thamorrat, translated as “the mountain of the valuable gems” (The Office of National Resources and Environmental Policy and Planning & The Fine Arts Department of Thailand, 2022, p. ii).

As illustrated in Table 3, the finalized list, containing keywords with a cut-off point of $p < 0.0001$ for keyness (15.13 with Bonferroni correction), is mostly dominated by nouns ($N = 81$; e.g., *Si Thep*, *park*, *UNESCO*), followed by adjectives ($N = 12$; e.g., *ancient*, *historical*, *cultural*), and verbs ($N = 7$; e.g., *listed*, *protect*, *located*). The keyword *Si Thep* exhibits the highest keyness value of 3,259.54, while *style* has the lowest keyness value of 54.97. It should be noted that certain multi-word sequences representing organization names (e.g., *Fine Arts Department*, abbreviated as *FAD*) and UNESCO's official designations (e.g., *World Heritage Site* as *WHS*; *World Heritage List* as *WHL*) are treated as single keywords, as they collectively convey a cohesive semantic unit or unified meaning in context rather than separate meanings.

Table 4
Types of lexical collocations of keywords in the ST corpus

Types	No. of Types	%	Examples
1. Adjective + Noun	75	43.60	<i>prehistoric times</i> , <i>cultural treasure</i>
2. Noun + Noun	65	37.79	<i>laterite stupa</i> , <i>Buddha sculptures</i>
3. Noun + Verb	13	7.56	<i>archaeologists believe</i> , <i>UNESCO recognised</i>
4. Noun + of Noun	11	6.40	<i>number of visitors</i> , <i>town of Si Thep</i>
5. Verb + Noun	7	4.07	<i>boost tourism</i> , <i>raised concerns</i>
6. Verb + Adverb	1	0.58	<i>dating back</i>
Total	172	100	

Note. See Appendix for all lexical collocations of the keywords

Table 4 presents six types and a total of 172 lexical collocations found in the ST corpus. Each collocate, meeting the applied criteria of an LL value ≥ 3.84 ($p < 0.05$), as well as a minimum frequency and range of ≥ 3 , is situated within span of 3L/3R of the node (i.e., *keyword*).

The two most common types of grammatical combination are Adjective + Noun ($N = 75$; see Appendix) and Noun + Noun ($N = 65$; see Appendix), representing 43.60% and 37.79% of the total, respectively. The occurrences of the Adjective + Noun combination (e.g., *prehistoric times*) are exemplified in the concordance lines (1)–(3), while those of Noun + Noun (e.g., *laterite stupa*) are shown in lines (4)–(6).

	Left context	Node	Right context
(1)	was discovered since	prehistoric	<i>times</i> and the local
(2)	turned towards the north. In	prehistoric	<i>times</i> , cemeteries were
(3)	of human settlement from	prehistoric	<i>times</i> . This city flourished
(4)	collapsed brick and <i>laterite</i>	stupa	about 64 metres long on
(5)	formidable brick and <i>laterite</i>	stupa	that stands 20 meters high
(6)	is a collapsed brick and <i>laterite</i>	stupa	, extending 64 meters on

The third and fourth most frequent types of lexical combination, discovered in the ST corpus, are Noun + Verb ($N = 13$; see Appendix) and Noun + of Noun ($N = 11$; see Appendix), constituting 7.56% and 6.40% of the total, respectively. The concordance lines (7)–(9) and (10)–(12) illustrate the pattern of the Noun + Verb (e.g., *archaeologists believe*) and that of the Noun + of Noun (e.g., *number of visitors*).

	Left context	Node	Right context
(7)	2014 and 2017, and	archaeologists	<i>believe</i> the ancient stupa –
(8)	between 2014 and 2017.	Archaeologists	<i>believe</i> the main stupa of
(9)	square meters at the base.	Archaeologists	<i>believe</i> it was constructed
(10)	, he said. “The <i>number of</i>	visitors	has climbed from 300 to
(11)	two years. As the <i>number of</i>	visitors	surged, authorities have
(12)	due to the large <i>number of</i>	visitors	. One Facebook user posted

The least occurring types of collocation pairs emerging from the ST corpus are Verb + Noun ($N = 7$; see Appendix) and Verb + Adverb ($N = 1$; see Appendix), accounting for 4.07% and 0.58% of the total, respectively. The repetitive pattern of the Verb + Noun (e.g., *boost tourism*) and that of the Verb + Adverb (e.g., *dating back*) are shown in the concordance lines (13)–(15) and (16)–(18) below.

	Left context	Node	Right context
(13)	Heritage Site, it could <i>boost</i>	tourism	and economic opportunities
(14)	World Heritage Site will <i>boost</i>	tourism	and the local economy in the
(15)	to help second-tier cities <i>boost</i>	tourism	in the lower northern region.
(16)	to monuments, and artefacts	dating	<i>back</i> to the Dvaravati period
(17)	These extraordinary artworks,	dating	<i>back</i> to the 8th century,
(18)	, pottery and other items	dating	<i>back</i> up to 2,000 years ago.

From this section onward, five keywords in the ST corpus with the most numbers of collocate will be presented, consisting of three adjectival keywords: *local*, *cultural*, and *ancient*, as well as two nominal ones: *UNESCO* and *Dvaravati*.

Among the top-100 keywords found in the ST corpus, *local*, ranked 29th in the list, is the adjectival keyword that has the highest number of collocates, constituting 10 variations of Adjective + Noun combination, as illustrated in the concordance lines (19)–(28).

	Left context	Node	Right context
(19)	Thep Historical Park have hired	local	<i>residents</i> to assist the
(20)	more material support from	local	government <i>agencies</i> , such as
(21)	inclusive, with the FAD and	local	<i>communities</i> having a say.
(22)	relics and stupas, revered by	local	<i>Buddhists</i> , are located there.
(23)	souvenirs, various amenities,	local	<i>guides</i> , and tour services related
(24)	is making concerns of the	local	<i>community</i> a top priority.
(25)	, which created jobs and helped	local	<i>people</i> gain an understanding
(26)	to work with Phetchabun and	local	<i>authorities</i> to better manage
(27)	UNESCO site could boost the	local	<i>economy</i> —but it could also put
(28)	these areas while boosting the	local	economy and <i>tourism</i> .

Ranked 17th in the list, *cultural* is also an adjectival keyword and has the second-highest number of collocates in the ST corpus, representing 9 pairs of Adjective + Noun sequence, as shown in the concordance lines (29)–(37).

	Left context	Node	Right context
(29)	city of Si Thep as a world	cultural	<i>heritage</i> during the 45 th session
(30)	Sites in Thailand comprise three	cultural	<i>sites</i> – Historic Town of
(31)	be Thailand’s fourth in the	cultural	<i>category</i> , he said.
(32)	a prominent hub for trade and	cultural	<i>exchange</i> . During its golden era,

	Left context	Node	Right context
(33)	communities stand to gain from	cultural	<i>tourism</i> . The Phetchabun
(34)	to delve into the rich history,	cultural	<i>significance</i> , and global
(35)	encompassing both natural and	cultural	<i>World Heritage Sites</i> , totaling 50
(36)	in Phetchabun province as a	cultural	<i>World Heritage Site</i> , bringing
(37)	Si Thep Historical Park is a	cultural	<i>treasure</i> for the whole country.

The last adjectival keyword *ancient*, ranked 3rd in the list, reveals the third-highest number of collocates—forming 8 types of Adjective + Noun clusters, as seen in the concordance lines (38)–(45), excerpted from the ST corpus.

	Left context	Node	Right context
(38)	just 20 kilometers west of the	ancient	<i>town</i> of Si Thep.
(39)	value, and significance of the	ancient	<i>city</i> of Si Thep, both to the
(40)	years ago. The matter of the	ancient	<i>ruins</i> was thrust into the
(41)	has temporarily closed off the	ancient	<i>monument</i> , allowing visitors
(42)	the explosive blasts. Many	ancient	<i>stupas</i> on that mountain have
(43)	48 Buddhist and Hindu	ancient	<i>monuments</i> . The Buddhist
(44)	one of the most intact	ancient	<i>sites</i> in Thailand. The Cabinet
(45)	rounded corners, housing 64	ancient	monuments and <i>structures</i> ,

The 5th ranked keyword in the ST corpus, *UNESCO*, is also a nominal constituent in 9 types of Noun + Noun patterns and in 1 type of Noun + Verb pattern, as exemplified in the concordance lines (46)–(55).

	Left context	Node	Right context
(46)	Si Thep's recognition as a new	UNESCO	<i>World Heritage Site</i> , the Fine
(47)	about to change. Following	UNESCO	's <i>recognition</i> , the
(48)	province, has been on the	UNESCO	<i>World Heritage List</i> since
(49)	added to the prestigious list of	UNESCO	<i>World Heritage Sites</i> in the
(50)	currently being considered for	UNESCO	<i>listing</i> . They include:
(51)	Park will threaten bid for	UNESCO	<i>status</i> . A protest against oil
(52)	it remains to be seen how the	UNESCO	<i>designation</i> will affect
(53)	and view the ruins as sacred.	UNESCO	's <i>declaration</i> of Si Thep
(54)	was added to the tentative	UNESCO	<i>list</i> in 2019, measures aimed
(55)	and Culture Center. Recently,	UNESCO	officially <i>recognised</i> the

In the same vein, *Dvaravati*, ranked 11th in the list, is another nominal keyword in the ST corpus with the highest number of collocates for nominal words, juxtaposing with 9 types of noun in Noun + Noun sequence, as represented in the concordance lines (56)–(64).

	Left context	Node	Right context
(56)	cultural characteristics of the	Dvaravati	<i>period</i> in terms of urban
(57)	structures are tied to the	Dvaravati	<i>culture</i> (7 th –11 th centuries),
(58)	be the largest temple from the	Dvaravati	<i>era</i> to be found in Thailand.
(59)	homage to local wisdom of the	Dvaravati	<i>civilization</i> , which existed
(60)	in the low-relief sculptures in	Dvaravati	<i>style</i> lining a small cave at its
(61)	Buddhism-based culture of the	Dvaravati	<i>kingdom</i> and Khmer culture
(62)	of Si Thep and its associated	Dvaravati	<i>monuments</i> to be inscribed
(63)	the distinctive aesthetics of	Dvaravati	<i>art</i> . Among the unearthed
(64)	the religious diversity of the	Dvaravati	<i>Empire</i> , which spanned the

Notably, *local*, *cultural*, *ancient*, *UNESCO*, and *Dvaravati*, are ranked in the top 30 of the list of keywords in the ST corpus, suggesting a correlation between the higher rank of the keywords and the larger number of collocates they have.

To better grasp how the top-100 keywords and their collocates are salient features of the ST corpus—reflective of the topics or “the big picture” (Baker et al., 2013, p. 35), the researcher qualitatively analyzed their occurrences in context using the Keyword in Context (KWIC) tool in AntConc. The analysis reveals nine topics of the news coverage on the ancient town of Si Thep, as shown in Table 5, as follows.

Table 5
Topics categorized by keywords and collocates in the ST corpus

Topics	Keywords and Collocates
1. Si Thep Honoured by UNESCO for Rich Culture	<i>UNESCO recognition, cultural heritage, cultural significance</i>
2. Si Thep Attracts Tourists	<i>cultural tourism, number of visitors, visitors explore</i>
3. Si Thep Overwhelmed by Overtourism	<i>sparked concerns, tourist overcrowding, unexpected influx</i>
4. Si Thep Art Influenced by Buddhism, Khmer, and Dvaravati	<i>Buddhist stupa, Dvaravati art, Khmer art</i>
5. Si Thep as a Historic Community	<i>ancient city, Dvaravati civilization, prehistoric times</i>
6. Si Thep Linked to Khao Klang Nok and Khao Thamorrat	<i>Khao Klang Nok site, Khao Klang Nok ruins, Khao Thamorrat site</i>
7. Si Thep Threatened by Oil Drilling	<i>drilling project, oil drilling, petroleum exploration</i>
8. Si Thep Conservation Includes Buffer Zones	<i>buffer zone, conservation area, conservation efforts</i>
9. Si Thep Hosts Excavation	<i>archaeologists unearthed, first excavation, team of archaeologists</i>

Table 5 illustrates nine specific topics related to the ancient town of Si Thep in the target corpus, supported by examples of the identified keywords and collocates. The emerging topics include Si Thep’s recognition by UNESCO, positive and negative outcomes of such recognition, historical descriptions, threats and conservation plans, as well as past and present archaeological activities. The first topic—*Si Thep Honoured by UNESCO for Rich Culture*—reports UNESCO’s decision to formally include Si Thep in its list of World Heritage Sites based on Si Thep’s vibrant and unique culture. This topic involves keywords and collocates such as *UNESCO recognition*, *cultural heritage*, and *cultural significance*, as exemplified in (65).

- (65) The UNESCO World Heritage Committee has recently recognized *Si Thep* Historical Park as a *cultural heritage* site.

The second topic, *Si Thep Attracts Tourists*, highlights the impact of UNESCO status in drawing throngs of visitors to Si Thep. Present within this topic are keywords and collocates such as *cultural tourism*, *number of visitors*, and *visitors explore*, as illustrated in (66).

- (66) The *Si Thep* Historical Park fanpage posted on Sunday that the park’s two parking lots were full due to the large *number of visitors*.

The other side of Si Thep's recent UNESCO designation is presented in the third topic, *Si Thep Overwhelmed by Overtourism*, noting its inability to host and handle the escalating volume of sightseers, causing worries within the community. This topic is reflected upon keywords and collocates, including *sparked concerns*, *tourist overcrowding*, and *unexpected influx*, as shown in (67).

- (67) An unexpected surge in tourists at *Si Thep* Historical Park, recently declared a World Heritage Site by UNESCO on September 19, *sparked concerns* among locals in Phetchabun.

Topic four—*Si Thep Art Influenced by Buddhism, Khmer, and Dvaravati*—focuses on the art and architectural structures of Si Thep being shaped by the religious and cultural intermingling between Buddhism, Khmer, and Dvaravati. This cross-cultural impact is represented through keywords and collocates, for instance, *Buddhist stupa*, *Dvaravati art*, and *Khmer art*, as demonstrated in (68).

- (68) *Si Thep* Historical Park is a significant archaeological site in Phetchabun Province, influenced by a blend of cultural and artistic traditions such as *Dvaravati art* and *Khmer art*.

The fifth topic, *Si Thep as a Historic Community*, emphasizes Si Thep's role as an early settlement, having thrived through several bygone eras. This topic is conveyed by keywords and collocates such as *ancient city*, *Dvaravati civilization*, and *prehistoric times*, as in (69).

- (69) *Si Thep* Historical Park is an *ancient city* that shows evidence of the *Dvaravati ancient civilization*.

The combinations of proper noun keywords and collocates (e.g., *Khao Klang Nok site*, *Khao Klang Nok ruins*, *Khao Thamorrat site*) suggest that Si Thep is not an independent entity but rather a center of a multifaceted and integrative compound, interconnected with other strategically vital areas. This observation gives rise to the sixth topic: *Si Thep Linked to Khao Klang Nok and Khao Thamorrat*, as exemplified in (70).

- (70) *Si Thep* Historical Park will be nominated as a cultural heritage site consisting of the three structures of *Si Thep* ancient city, *Khao Klang Nok* archaeological site and *Khao Thamorrat* archaeological site.

Topic seven, *Si Thep Threatened by Oil Drilling*, is probably the most notorious issue reported in the news. Not only is Si Thep in the limelight for its cultural and historical values, but the area is also well known for its petroleum supplies. As a result, Si Thep has been targeted by the oil drilling industry, aiming to exploit these natural resources. This controversy is evidenced by keywords and collocates such as *drilling project*, *oil drilling*, and *petroleum exploration*, as demonstrated in (71).

- (71) Conflicts over the pros and cons of an *oil drilling* project near the *Si Thep* Historical Park in Thailand's lower northern province of Phetchabun, have flared up again.

Accompanied by keywords and collocates, including *buffer zone*, *conservation area*, and *conservation efforts*, the eighth topic is *Si Thep Conservation Includes Buffer Zones*, as shown in (72). To mitigate the threats posed by the petroleum industry, which could adversely affect Si Thep's status as a UNESCO World Heritage Site, authorities have developed conservation plans to establish buffer zones between the heritage site and the oil drilling areas.

- (72) With recognition as a Unesco World Heritage Site possibly hanging in the balance, the government's Fine Arts Department has begun delineating a *buffer zone* at *Si Thep* Historical Park in Phetchabun to try and protect it from proposed oil drilling.

The final topic—*Si Thep Hosts Excavation*—reflects another significant activity within the area. Archaeological excavations in Si Thep, conducted and supervised by a team from the government, lead to discoveries of historical evidence vital to the continual revelation of Si Thep's past. This topic is highlighted by keywords and collocates like *archaeologists unearthed*, *team of archaeologists*, and *first excavation*, as in (73).

- (73) A *team of archaeologists* of the Fine Arts Department found more than 100 ancient Buddhist and Hindu monuments such as Khao Khlang Nai, Prang *Si Thep* and Prang Song Phi Nong as well as dozens of ancient reservoirs of different sizes.

The subsequent section will discuss the findings of the present study in relation to those of relevant previous research.

DISCUSSIONS

The investigation of the top-100 keywords in the ST corpus reveals three grammatical categories of lexical keywords, namely, nouns, adjectives, and verbs, in descending order of occurrence. The dominance of nominal keywords, followed by adjectival and verbal ones, in this study, is consistent with previous analyses of lexical keywords in other specialized corpora of *insurance* (Khamphairoh & Tangpijaikul, 2012), *nursing* (Trinant & Yodkamlue, 2019), *tourism* (Trinant & Kijpoonphol, 2021), and *business* (Sukman & Tangkiengsirisin, 2024). In this study, the keywords *Si Thep*, *World Heritage Site*, *Dvaravati*, *UNESCO*, and *FAD* are all ranked among the top 30 most salient keywords, as indicated by their keyness values. This phenomenon is rather expected, as all these keywords are considered focal points and main agents of the news coverage under study. The findings of this study, however, may reaffirm the essential role of proper nouns and abbreviations in representing the "aboutness" (Scott, 2015, p. 236) and serving as "a key element" (Khamphairoh & Tangpijaikul, 2012, p. 182) in a specialized corpus. This stands in contrast to previous studies on specialized corpora (Sukman et al., 2022; Sukman & Tangkiengsirisin, 2024; Suraprajit, 2022; Thaksanan & Kamsa-ard, 2025; Trinant & Kijpoonphol, 2021; Trinant & Yodkamlue, 2019), which completely marginalized proper nouns and abbreviations from keyword analysis.

The analysis of collocations of keywords extracted from the ST corpus shows that Adjective + Noun and Noun + Noun are the two most common types of collocational sequences in

descending order of distribution. The result of Adjective + Noun being the most frequent collocation pattern, followed by Noun + Noun, in this study is in line with that of Sukman et al. (2022) but contradicts those of Sukman and Tangkiengsirisin (2024), Thaksanan and Kamsa-ard (2025), Trinant and Kijpoonphol (2021), and Trinant and Yodkamlue (2019) who all discovered that the number of Noun + Noun combinations was higher than that of Adjective + Noun.

The predominance of Adjective + Noun clusters in the ST corpus could be influenced by the significant presence of adjectives in news reporting, either to shape the tone or engage readers with context-dependent adjectives (Bin Kadri, 2023). For instance, Adjective + Noun combinations such as *cultural significance*, *cultural treasure*, and *historical significance* are likely used to set a positive tone toward UNESCO's declaration of the ancient town of Si Thep as a World Heritage Site, while Adjective + Noun sequences such as *Khmer arts*, *pyramidal ruins*, *grand stupa*, and *prehistoric settlement* all contain context-driven descriptive adjectives, aiming to sensationalize and immerse readers into their mental depiction of the ancient town as a primitive compound with a blend of rich culture and history.

Another point worth discussing, regarding the higher number of Noun + Noun clusters in other specialized corpora (Sukman & Tangkiengsirisin, 2024; Thaksanan & Kamsa-ard, 2025; Trinant & Kijpoonphol, 2021; Trinant & Yodkamlue, 2019), is the difference between the text genres of those specialized corpora and the ST corpus. The specialized corpora compiled by Thaksanan and Kamsa-ard (2025), Trinant and Yodkamlue (2019), and Trinant and Kijpoonphol (2021) consist of research articles in academic genres. According to Swales (2008, p. v), the lexical combinations in academic writing are mostly "noun-centric", thus, resulting in a higher number of Noun + Noun sequences. Likewise, although the specialized corpus created by Sukman and Tangkiengsirisin (2024) consists of news reports, similar to the ST corpus, its content is business-related, thereby leaning more towards objectivity and neutrality. The authors of business news are less obliged to focus on sensationalism, establishing the tone, or captivating readers through frequent use of adjectives. This, consequently, leads to a higher number of Noun + Noun and a lower number of Adjective + Noun clusters compared to those in the ST corpus.

In addition to genre differences, the methodological choices opted for by each study could influence the findings regarding the proportions of identified collocational patterns. Previous studies, including those by Sukman and Tangkiengsirisin (2024), Thaksanan and Kamsa-ard (2025), Trinant and Kijpoonphol (2021), and Trinant and Yodkamlue (2019), have consistently found that Noun + Noun combinations are the most prevalent collocational type in their corpora, using the MI score to gauge the strength of association between keywords and their collocates. In contrast, this study employed the LL value, ultimately resulting in a higher dominance of Adjective + Noun combinations. According to Smith (2009), the MI score and LL value are technically dissimilar in how they calculate collocational association; the former does not take the total frequencies of each collocate into consideration, while the latter does. In other words, MI measures "effect size", while LL assesses "significance" of keywords and collocates (Evert, 2009, p. 31). This potentially leads to different results in collocation identification.

Identifying keywords and collocations from the specialized ST corpus, gathered from a genre of journalistic writing, should provide teachers and students with a manageable list of more sophisticated lexical combinations of key vocabulary and collocates that reflect authentic language use. For instance, in this study, the keyword *ancient* is more formal than its commonly used equivalent *old*. Such examples can help raising students' awareness on opting for suitable word choices across language contexts. Regarding collocation, *ancient* specifically combines with nouns denoting structures and buildings that have exceeded a lifetime or lasted for generations (e.g., *ancient town*, *ancient monument*, *ancient ruin*), while *old* can generally co-occur with terms for living beings (e.g., *people*, *animals*, *plants*). This serves as an illustrative example of the concept of English collocations and highlights the subtle nuances in meaning and semantic preferences of the keyword *ancient* and its synonymous counterpart *old*.

PEDAGOGICAL IMPLICATIONS

Ability to recognize and use lexical chunks is of paramount importance in the acquisition of English vocabulary (Lewis, 1997). This study's analysis of the lexical collocations of keywords in the ST corpus could provide ESP instructors with practical applications for designing lexical collocation exercises, authentically tailored for Thai learners or users of English, involved in tourism-related ventures in the ancient town of Si Thep. The development of collocation exercises using authentic corpus-derived data not only provides students with exposure to word combinations with strong collocation but also to weaker ones that could be useful for written and oral communication in English (O'Dell & McCarthy, 2008). Adapted from Khamphairoh and Tangpijaikul (2012), two types of collocation exercises, designed based on this study's corpus-informed findings, are as follows.

Isolated Collocation Matching

- | | |
|--------------|--------------|
| 1. ancient | district |
| 2. Dvaravati | ruins |
| 3. Si Thep | civilization |

First, isolated collocation matching is aimed at raising the awareness of learners and users of English regarding the fundamental concept of lexical collocations, implying that two specific words are more likely to occur together in certain grammatical combinations than others.

Contextualized Collocation Matching

significance	excavations	date
archaeological	back	historical

1. Si Thep Historical Park, spread over an area of 867 hectares within the Si Thep district, houses three cultural sites: the ancient town of Si Thep, Khao Klang Nok, and Khao Thamorrat Cave. These ancient monuments to the Dvaravati period, a kingdom that held sway from the sixth to the 11th century.

2. Not merely a badge of honor, UNESCO recognition was a testament to the timeless beauty and that Si Thep and its monuments have held, silently whispering tales of the bygone Dvaravati period to those who walk their ancient paths.
3. The limestone and rhyolite cave, named Khao Thamorratt, in Si Thep is believed to be a hollowed mountain, which served as an important monastery during the Dvaravati period, and a variety of artefacts from the late Bronze Age were uncovered during the there.

Second, contextualized collocation matching is harder to complete than the isolated one, as this very exercise requires learners or users of English to be aware of collocations and utilize cognitive reading skills, such as making inferences from context clues, in order to place the correct keyword-collocate combination in the blanks. Noticeably, the two exercises are arranged in increasing levels of difficulty. Learners and users of English can start by practicing the former before progressing to the latter. Vocabulary exercises are more beneficial for learners when they are systematically developed for repeatability and structured in ascending levels of difficulty, as part of instructional scaffolding (Khamphairoh & Tangpijaikul, 2012).

CONCLUSION

The ST corpus reveals three grammatical categories of keywords, nouns, adjectives, and verbs in descending order of occurrence. The keywords in the ST corpus are integral parts of 172 types of lexical collocations, grammatically classified as Adjective + Noun, Noun + Noun, Noun + Verb, Noun + of Noun, Verb + Noun, and Verb + Adverb, arranged from highest to lowest proportion. The adjectival keywords *local*, *cultural*, and *ancient* constitute 27 variations of Adjective + Noun combinations, while the nominal keywords *UNESCO* and *Dvaravati* are associated with 18 types of nouns in Noun + Noun sequences. The combinations of keywords and collocates are also apparent across nine topics about the ancient town of Si Thep, indicating the role these sequences contribute to achieving different communicative goals.

This study may be subject to several limitations. Firstly, the size of the ST corpus is relatively small due to the limited availability of data during the research period. Consequently, certain types of lexical collocations, such as Adverb + Adjective combinations, may be entirely absent from this study. Therefore, employing a larger corpus could enhance the likelihood of identifying less common types of lexical collocations. Secondly, as this study solely relied on corpus-data from Thailand-based news outlets, the findings may not be generalizable to the broader English-language discourse surrounding the ancient town of Si Thep. If interested, future researchers could address this shortcoming by gathering corpus data from news articles published by foreign news agencies. However, to the researcher's best knowledge, news coverage on the ancient town of Si Thep in international media has been scarcer compared to Thailand-based media. To address this challenge, transcriptions of English commentary from online videos related to the ancient town of Si Thep, narrated by native English speakers, could serve as an alternative authentic source for the specialized corpus.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Assoc. Prof. Dr. Supakorn Phoocharoensil of the Language Institute of Thammasat University for his academic guidance during the initial drafting of this paper, and to Ajarn Oliver Oonlamai of the Faculty of Humanities at Phetchabun Rajabhat University for proofreading the final draft.

THE AUTHOR

Nateethorn Narkprom, PhD, is a lecturer of English in the Faculty of Education at Phetchabun Rajabhat University, Thailand. His primary research passion lies in corpus linguistics.

nateethorn.nar@pcru.ac.th

REFERENCES

- Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software>
- Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.
- Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511920103>
- Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press ELT. <https://doi.org/10.3998/mpub.371534>
- Benson, M., Benson, E., & Ilson, R. (2010). *The BBI combinatory dictionary of English: Your guide to collocations and grammar* (3rd ed.). John Benjamins. <https://doi.org/10.1075/z.bbi>
- Bin Kadri, H. K. (2023). The sentiment analysis on news reporting: The sentiment of adjectives in news articles. *E-Jurnal Bahasa Dan Linguistik (e-JBL)*, 5(2), 68–74. <https://doi.org/10.53840/ejbl.v5i2.154>
- Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. In A. Abel & L. Lemnitzer (Eds.), *Network strategies, access structures, and automatic extraction of lexicographical information* (2nd Work Report of the Academic Network Internet Lexicography, OPAL–Online publizierte Arbeiten zur Linguistik, pp. 48–61). Leibniz Institute for the German Language.
- Boonraksa, T., & Naisena, S. (2022). A study on English collocation errors of Thai EFL students. *English Language Teaching*, 15(1), 164–177. <https://doi.org/10.5539/elt.v15n1p164>
- Boonyarattanasoontorn, P., Sirawich, S., & Chomploen, C. (2020). The effect of a collocation teaching innovation on students' use of collocations. *Journal of Studies in the English Language*, 15(2), 98–129. <https://so04.tci-thaijo.org/index.php/jsel/article/view/240158>
- Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, 3(2), 1–10. <http://dx.doi.org/10.7820/vli.v03.2.2187-2759>
- Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, 2, 103–116.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations* [Unpublished doctoral dissertation]. Institute for Natural Language Processing, University of Stuttgart.

- Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook*. (pp. 1212–1248). Mouton de Gruyter. <https://doi.org/10.1515/9783110213881.2.1212>
- Fujita, R. (2019). English for tourism and hospitality. In H. Terauchi, J. Noguchi, & A. Tajino (Eds.), *Towards a new paradigm for English Language Teaching: English for Specific Purposes in Asia and beyond* (1st ed., pp. 172–180). Routledge. <https://doi.org/10.4324/9780429423963>
- Historical park grapples with crowds. (2023, September 21). *Bangkok Post*. <https://www.bangkokpost.com/thailand/general/2652870/historical-park-grapples-with-crowds>
- Jaén, M. M. (2009). A corpus-driven design of a test for assessing the ESL collocational competence of university students. *International Journal of English Studies*, 7(2), 127–148. <https://revistas.um.es/ijes/article/view/49031>
- Khamphairoh, T., & Tangpijailul, M. (2012). Collocations of keywords found in insurance research articles: A corpus-based analysis. *Manutsayasat Wichakan*, 19(2), 166–188. <https://so04.tci-thaijo.org/index.php/abc/article/view/54124>
- Koester, A. (2010). Building small specialised corpora. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66–79). Routledge.
- Lei, L., & Liu, D. (2018). The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics*, 23(2), 216–243. <https://doi.org/10.1075/ijcl.16135.lei>
- Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice*. Language Teaching Publications.
- Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach*. Language Teaching Publications.
- Lyse, G. I., & Andersen, G. (2012). Collocations and statistical analysis of N-grams. Exploring newspaper language: Using the web to create and investigate a large corpus of modern Norwegian. *Studies in Corpus Linguistics*, 49, 79–109. <https://doi.org/10.1075/scl.49>
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.
- Namtapi, I. (2022). Needs analysis of English for specific purposes for tourism personnel in Ayutthaya. *LEARN Journal: Language Education and Acquisition Research Network*, 15(1), 409–439. <https://so04.tci-thaijo.org/index.php/LEARN/article/view/256730>
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
- Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins. <https://doi.org/10.1075/scl.14>
- O’Dell, F., & McCarthy, M. (2008). *English collocations in advanced use*. Cambridge University Press.
- Phu Phra Bat named World Heritage site. (2024, July 27). *Bangkok Post*. <https://www.bangkokpost.com/thailand/general/2836758/phu-phra-bat-named-world-heritage-site>
- Potts, A., & Baker, P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics*, 17(3), 295–324. <https://doi.org/10.1075/ijcl.17.3.01pot>
- Pojanapunya, P., & Lieungnapar, A. (2017). *The quality of choices determines the quantity of keywords*. In *Proceedings of the Doing Research in Applied Linguistics 3 and 19th English in South-East Asia Conference 2017 (DRAL 3/19th ESEA)* (pp. 115–124). School of Liberal Arts, King Mongkut’s University of Technology Thonburi. http://sola.kmutt.ac.th/dral2017/proceedings/3-Papers/115-124_The%20quality%20of%20choices%20determines%20the%20quantity%20of%20Key%20words_Punjaborn%20Pojanapunya%20and%20Angvarrah%20Lieungnapar.pdf
- Scott, M. (1997). PC analysis of key words - and key key words. *System*, 25(2), 233–245. [https://doi.org/10.1016/S0346-251X\(97\)00011-0](https://doi.org/10.1016/S0346-251X(97)00011-0)
- Scott, M. (2015). *WordSmith tools manual (Version 6.0)*. Lexical Analysis Software Ltd. <https://lexically.net/downloads/version6/wordsmith6.pdf>

- Smith, N. (2009, February 14). *Collocations: “Delicious” and “tasty”: How do they behave with other words?* Corpus Linguistics: A Practical Web-based Course. https://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l07_4.htm
- Shamsudin, S., Sadoughvanini, S., & Zaid, Y. H. (2013). Iranian EFL learners’ collocational errors in speaking skill. *Procedia–Social and Behavioral Sciences*, 70, 1295–1302. <https://doi.org/10.1016/j.sbspro.2013.01.190>
- Sukman, K., & Tangkiengsirisin, S. (2024). Lexical collocations of keywords used in business news: A corpus-based study. *European Journal of English Language Teaching*, 9(1), 23–37. <http://dx.doi.org/10.46827/ejel.v9i1.5202>
- Sukman, K., Triwatwanon, W., Munkongdee, T., & Chumnumnawin, N. (2022). A corpus-based study of lexical collocations of keywords found in online business news articles. *European Journal of English Language Teaching*, 7(3), 59–72. <http://dx.doi.org/10.46827/ejel.v7i3.4275>
- Suraprajit, P. (2022). English words and collocations found on logistics magazines: A corpus-based study. *Advances in Language and Literary Studies*, 13(1), 41–48. <https://doi.org/10.7575/aialc.all.v.13n.1.p.41>
- Swales, J. M. (2008). Foreword. In D. Belcher & A. Hirvela (Eds.), *The oral-literate connection: Perspectives on L2 speaking, writing, and other media interactions* (pp. v–viii). University of Michigan Press ELT. <https://doi.org/10.3998/mpub.231182>
- Szudarski, P. (2023). *Collocations, corpora and language learning*. Cambridge University Press. <https://doi.org/10.1017/9781108992602>
- Thaksanan, P., & Kamsa-ard, T. (2025). Lexical collocation analysis in a corpus of research articles in physical education and sport science. *The New English Teacher*, 19(1), 51–70. <https://assumptionjournal.au.edu/index.php/newEnglishTeacher/article/view/8149/4272>
- The Office of National Resources and Environmental Policy and Planning, & The Fine Arts Department of Thailand. (2022). *The ancient town of Si Thep: World Heritage nomination submitted by the Government of the Kingdom of Thailand AD 2022* (Nomination text). <https://whc.unesco.org/document/192676>
- Trinant, K., & Kijpoonphol, W. (2021). Lexical collocations in a sample corpus of tourism research articles (SCTRA). *NKRAFA Journal of Humanities and Social Sciences*, 9, 94–108. <https://so04.tci-thaijo.org/index.php/KANNICHA/article/view/249524>
- Trinant, K., & Yodkamlue, B. (2019). Lexical collocations in a Sample Corpus of Nursing Research Articles (SCNRA). *Suranaree Journal of Social Science*, 13(1), 45–72. <https://doi.org/10.55766/OBXC3400>
- UNESCO World Heritage Convention. (2023). *Decision: 45 COM 8B.41 The ancient town of Si Thep and its associated Dvaravati monuments (Thailand)*. <https://whc.unesco.org/archive/2023/whc23-45com-19-en.pdf>
- Vaughan, E., & Clancy, B. (2013). Small corpora and pragmatics. In J. Romero-Trillo (Ed.), *The yearbook of corpus linguistics and pragmatics 2013: New domains and epistemologies* (Vol. 1, pp. 53–73). Springer. https://doi.org/10.1007/978-94-007-6250-3_4
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Appendix

All collocations of the keywords in the ST corpus categorized by types

Types	Lexical Collocations
Adjective + Noun	ancient city, ancient monument, ancient monuments, ancient ruins, ancient sites, ancient structures, ancient stupas, ancient town, archaeological excavations, archaeological site, archaeological sites, associated monuments, Buddhist monuments, Buddhist sculptures, Buddhist stupa, cultural category, cultural exchange, cultural heritage, cultural significance, cultural sites, cultural tourism, cultural treasures, cultural World Heritage Site, cultural World Heritage Sites, designated area, extended session, first excavation, foreign visitors, fourth site, grand stupa, Hindu monuments, historic city, historic town, historic towns, historical park, historical parks, historical significance, historical site, inner town, Khmer art, Khmer arts, Khmer style, largest stupa, local agencies, local authorities, local Buddhists, local communities, local community, local economy, local guides, local people, local residents, local tourism, main stupa, national committee, national parks, natural sites, northern province, other World Heritage Sites, outer town, prehistoric settlement, prehistoric times, prehistoric village, provincial executive, provincial committees, pyramidal ruins, red laterite, tentative list, Thai artefacts, Thai identity, three sites, two sites, unexpected influx, Thai authorities, Thai delegation
Noun + Noun	brick stupa, Buddha heads, Buddha images, Buddha sculptures, buffer zone, buffer zones, Buriram province, district chief, city wall, conservation area, conservation efforts, drilling project, Dvaravati civilisation, Dvaravati empire, Dvaravati era, Dvaravati kingdom, Dvaravati period, Dvaravati art, Dvaravati culture, Dvaravati monuments, Dvaravati style, exploration company, exploration plan, heritage center, heritage committee, heritage protection, heritage sites, heritage status, Khao Klang Nok monument, Khao Klang Nok ruins, Khao Klang Nok site, Khao Klang Nok stupa, Khao Thamorrat cave, Khao Thamorrat caves, Khao Thamorrat mountain, Khmer empire, laterite bricks, laterite stupa, Mahayana Buddhism, oil drilling, oil exploration, oil project, limestone cave, petroleum exploration, Phetchabun province, Si Thep district, Si Thep city, square kilometres, relief sculptures, tourist accommodation, tourist attractions, tourist center, tourism influx, tourist overcrowding, Udonthani province, UNESCO declaration, UNESCO designation, UNESCO list, UNESCO listing, UNESCO recognition, UNESCO World Heritage List, UNESCO World Heritage Site, UNESCO World Heritage Sites, world heritage, World Heritage Site status
Noun + Verb	archaeologists believe, archaeologists unearthed, artefacts (were) found, authorities discovered, committee considered, excavation began, Fine Arts Department began, park contains, province (was) prepared, town (is) surrounded, UNESCO recognised, visitors climb, visitors explore
Noun + of Noun	town of Si Thep, city of Si Thep, province of Phetchabun, number of visitors, surge of visitors, influx of visitors, images of ruins, list of World Heritage Sites, number of World Heritage Sites, team of archaeologists, influx of tourists
Verb + Noun	boost tourism, climbing Khao Klang Nok, expressed concerns, granted concessions, raised concerns, sparked concerns, support project
Verb + Adverb	dating back