Eurasian Journal of Educational Research 106 (2023) 211-231

Eurasian Journal of Educational Research
www.ejer.com.tr

Lexical Richness in Thai Textbooks Published in China

Minsu Kong[1*], Siripen Ungsitipoonporn[2], Pattama Patpong[3]

ARTICLEINFO

Article History:
Received: 24 April 2023
Received in revised form: 22 November 2023
Accepted: 19 December 2023
DOI: 0.14689/ejer.2023.106.013

Keywords
Lexical Density, Type-Token Ratio (Ttr), Standardized Type/Token Ratio (Sttr), Thai Language Textbooks, Vocabulary Richness.

ABSTRACT

Purpose: This study explores the lexical density found in Thai language textbooks utilised in Chinese educational environments, with the aim of deepening our comprehension of language acquisition. This study utilises corpus linguistics techniques to examine four sets of textbooks, each consisting of four volumes. The objective is to identify similarities and differences in the usage of vocabulary. **Method:** This study employs corpus linguistics techniques to comprehensively examine Thai language textbooks, using the type-token ratio (TTR) as a quantitative measure. Preliminary observations indicate a correlation between decreased content levels, restricted vocabulary, and an increased type-token ratio (TTR). A non-linear relationship was observed when comparing type-token ratios (TTRs) across volumes at the same level. To address these anomalies, a stoplist and standardised TTR (STTR) analysis were introduced. **Results:** The analysis demonstrates that higher-level texts display more intricate language structures compared to lower-level texts, which contradicts initial assumptions. The refinement of these findings is achieved using a stoplist and STTR analysis. Although there may be segmentation errors, the total number of tokens is an important factor in analysing TTR. This emphasises the importance of considering token count when examining lexical patterns in textbooks. **Implications for Research and Practice:** The study's findings have significant implications for both language educators and learners. This research provides a foundation for future investigations into effective language instruction and course design methods by examining lexical patterns in Thai language textbooks. The discovery of an unexpected non-linear relationship highlights the significance of improving assumptions and methodologies in language education research. This study promotes a nuanced approach to evaluating vocabulary, highlighting the importance of flexible pedagogical strategies that can be adjusted based on the observed complexities of the vocabulary in educational materials.

[1] Associate Professor, Foreign Languages Department, Kunming university, Yunnan province, China 650214 (PHD student, Research Institute for Languages and Cultures of Asia, Mahidol University, Nakhon Pathom, Thailand, 73170). Email: kongminsu@163.com, ORCID: https://orcid.org/0009-0006-4758-1617
[2] Associate Professor, Research Institute for Languages and Cultures of Asia, Mahidol University, NakhonPathom, Thailand 73170. Email: siripen.ung@mahidol.edu, ORCID: https://orcid.org/0000-0002-7685-8498
[3] Assistant Professor, Linguistics Department, Research Institute for Languages and Cultures of Asia, Mahidol University, Nakhon Pathom, Thailand, 73170. email: ppattama@yahoo.com
ORCID: https://orcid.org/0000-0001-9916-4205
* Corresponding Author Email: kongminsu@163.com

1. Introduction

China and Thailand have maintained diplomatic relations since 1975. Thai language education in China has a long history, with its origins dating back to 1579 at Si Yi Guan (an official institute of translators in the Ming dynasty).[1] During that period, access to education was limited to the male offspring of government employees (Liu, 2015). Peking University established the Department of Eastern Languages, known as Dong Yu Xi, in 1962, following the establishment of the new China. At its inception, the department offered a major in Thai and enrolled a modest number of 13 students. According to government statistics, there were more than 5000 Thai language learners in the Yunnan Province in 2022. Additionally, 18 universities in the province provided bachelor's degree programmes in Thai language studies (Read, 2000).

Yunnan province holds a strategic position in China, serving as a gateway to both South and Southeast Asia. This advantageous location has positioned Yunnan as a key player in China's efforts to engage with the global community. Proficiency in the Thai language significantly impacts the implementation of the Belt and Road Initiative and facilitates exchanges between Yunnan and Thailand across multiple domains, including economy, trade, culture, and education (Aroonmanakun, 2007).

Textbooks of varying difficulty levels are chosen for Thai language learners, ranging from beginners to advanced learners. Certain universities utilise multiple textbooks for teaching Thai courses, covering a range of proficiency levels, starting from fundamental concepts to more advanced topics (Davies & Elder, 2004). Vocabulary instruction is essential for all learners, regardless of the textbook being used. This study aims to analyse Thai textbooks published in China, specifically focusing on the vocabulary used in commonly used textbooks (Schreuder & Weltens, 1993). The analysis will utilise corpus analysis techniques to investigate the distribution of word frequencies across all textbooks and at different levels (Archer, 2016).

In 2022, statistics indicate that 50 universities in China, including 18 universities in Yunnan Province, provided Thai language courses.[2] The researcher conducted an online questionnaire survey to investigate the selection of Thai textbooks among Chinese universities that have recently introduced Thai courses (Gries, 2012). The survey focused on the period from 2018 to 2022 and specifically examined four sets of textbooks, each consisting of four volumes.

Vocabulary and textbooks are crucial components in input- and input-based instructional approaches. This research focuses on the vocabulary size and hierarchical structure of textbooks published in China for Thai learners (Thai National Corpus, 2007). Figure 1 demonstrates that teachers' experience and subjective judgement alone are insufficient for determining the answers. Individual differences can lead to the neglect of certain problems when people rely solely on their personal experiences. This study utilises corpus analysis to provide practical methods for data analysis (McEnery, Xiao, & Tono, 2006).

The research question in this study is.

"How big is the vocabulary in the textbooks for Thai learners in Chinese universities based on the analysis of the relationship between the number of types and tokens in different levels of the texts and textbooks?"

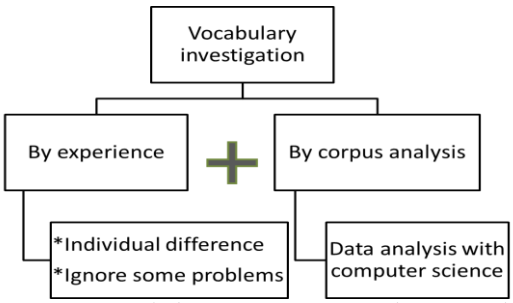

***Figure 1**: Vocabulary Investigation in This Research.*

There is no "right" or "best" way to teach vocabulary. The optimal approach in each scenario varies based on factors such as student characteristics, target vocabulary, school system, curriculum, and other relevant considerations. Schmitt (2008) suggests that certain principles should be considered when designing a vocabulary component for a language course. This study aims to analyse the word richness and frequency distribution characteristics of textbooks by conducting a corpus analysis of conversations and reading passages within the textbooks (Baayen, 2001).

The study encompasses four sets of Thai textbooks (including 16 volumes). There are few recently published textbooks, consisting of only one or two volumes that are rarely used in Thai language learning courses offered by universities (Nation, 2022). The researcher will focus on studying the commonly used textbooks in four volumes at the same level, excluding newly created and less-used textbooks.

The approaches are quantitative as they rely on statistical methodologies. The objective is to uncover the underlying word distribution at various levels and categories of texts using random data. According to some researchers, more research is necessary to determine the best pedagogical strategy for teaching high-frequency vocabulary (Allen, 1983). Conducting word frequency analysis can inform the establishment of realistic vocabulary size goals and the development of effective strategies for teaching high-frequency words in language programmes and learner materials (Li, 1992). This approach can be valuable in classroom research aimed at assessing the effectiveness of these strategies (Schmitt & Schmitt, 2014).

2. Literature Review

2.1 Vocabulary knowledge

Vocabulary extends beyond individual words and encompasses a broader scope than mere linguistic units. Recent studies on vocabulary focus on lexis, which is derived from the Greek word for "word" and refers to the complete set of words in a language (Barcroft, Sunderman, & Schmitt, 2011). Vocabulary refers to the collection of words in a language, encompassing both individual words and multi-word expressions that convey specific meanings like single words. Vocabulary acquisition is crucial for individuals learning a new language. Vocabulary plays a crucial role in language instruction and acquisition, as it enables students to comprehend others and articulate their thoughts effectively. People

can communicate effectively using valid words and expressions, even without adhering to strict grammar rules. Assisting students in acquiring effective vocabulary knowledge and enhancing their vocabulary learning strategies is crucial for promoting effective communication and comprehension skills.

When people acquire a new language, the quantity of vocabulary they possess influences their ability to effectively communicate in various contexts. When educators or learners are prompted to discuss "the primary challenges in acquiring a foreign language," the topic of "vocabulary" consistently emerges as a prominent response. Many language learners often express frustration with the challenges they face in learning a foreign language due to a limited vocabulary. Consider the field of teaching English as a foreign language (TEFL) (Folse, 2004). Teachers are instructed to deliver comprehensible input in order to engage students in authentic native speech, rather than teaching words in isolation. Students are advised to read for "gist" and listen for "essential," disregarding any perplexing words (Schmitt & Schmitt, 2020). The primary issue arises when students encounter a significant number of unfamiliar words, leading to increased frustration for both teachers and students. This is particularly concerning as it hampers the acquisition of comprehensible input and valuable reading and listening skills.

To acquire the "mastery of words," it is imperative to engage in the study of words. Lexis and grammar are distinct entities, with grammar arising from patterns found in the lexicon (Lewis, 1997). According to Sinclair (2004), a lexical item refers to an extended unit of meaning. A word or phrase, along with its collocates, semantic prosody, semantic preference, and colligation, can be categorised as either obligatory or optional, as depicted in Figure 2.

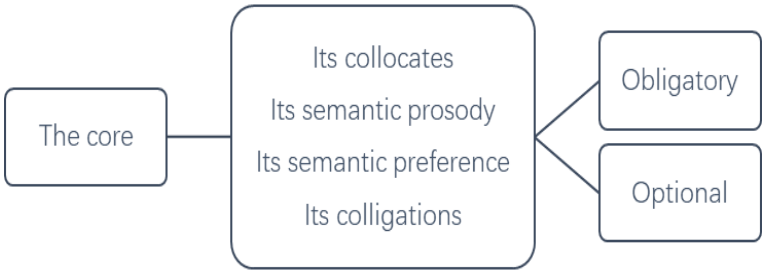

***Figure 2**: Extended Unit of Lexical Meaning.*

Language-oriented vocabulary development is crucial for the majority of second language (L2) learners in a language course. Based on research on L2 vocabulary acquisition and the researcher's experience, an effective approach to supporting Chinese Thai learners in acquiring vocabulary is to teach words that have been analysed for their frequency, range, and collocation in the target language (Klein, 1986). Next, attempt to distribute the chosen words as one acquires knowledge. Substantial input is essential for language learning, and textbooks play a vital role in providing the necessary vocabulary for input and input-based teaching (Hu, 2005). By utilising scientific analysis to select words from a corpus, teachers and students can effectively plan language teaching and learning activities, fostering creativity in the process (Laufer & Ravenhorst-Kalovski, 2010).

3. Corpus Linguistics

Prior to the emergence of computers, language samples were manually gathered on small paper slips. Computers have enabled the creation of corpora. The inaugural computer-based corpus, known as the Brown (2001) corpus, was established in 1961 and consisted of approximately 1 million words (Sharoff et al., 2013). Currently, generalised corpora consist of hundreds of millions of words, and the field of corpus linguistics is making significant advancements in the realm of second language research and education (Wilkinson, 2011). According to Leech (1992), this approach is referred to as "computer compass linguistics," which is a novel philosophical approach. Gries (2009) and others view it solely as a methodology, neither more nor less. Taylor (2013) provides a comprehensive overview of the extensive size and diversity of the field, further expanding on the discussion of this issue. Gries (2009) argues that corpus linguistics extends beyond being a mere methodology. It is reasonable to assert that a significant portion of corpus linguistics research primarily focuses on descriptive or applied aspects rather than extensively incorporating theoretical language frameworks (Nation & Waring, 1997). This is because corpus data can be utilised by linguists from various theoretical backgrounds. Gries acknowledges that a specific type of language theory aligns well with corpus linguistic methods (Gries, 2010). Usage-based cognitive theories of language are compatible with corpus linguistics in various ways. Linguists perceive corpus linguistics either as a research tool or methodology, or as a distinct discipline or theory. According to Kübler and Zinsmeister (2015) "the answer to the question whether corpus linguistics is a theory, or a tool is simply that it can be both''. It depends on how corpus linguistics is applied". "Corpus linguistics is concerned not just with describing patterns of form", says Cheng (2011), "but also with how form and meaning are inseparable".

Corpus linguistic analysis relies on assessing frequencies, which are considered a fundamental explanatory mechanism in cognitively motivated approaches. Frequency, along with its cognitive counterpart, cognitive entrenchment, is one of several central mechanisms evaluated in this analysis (Ellis, 2002).

"In corpus linguistics quantitative and qualitative methods are extensively used in combination (Gries, 2014). It is also characteristic of corpus linguistics to begin with quantitative findings, and work toward qualitative ones.…Generally it is desirable to subject quantitative results to qualitative scrutiny--attempting to explain why a particular frequency pattern occurs, for example……" (Traugott, 2012).

Alternatively, Sinclair (2005) presents many primary classifications derived from linguistic corpora, such as a solitary text, an archival collection, and specifically the World Wide Web. Corpus types can be classified based on the medium in which the data is contained. The user's text refers to several forms of media, including written text (such as text documents, historical manuscripts, and the World Wide Web), audio recordings, audio-visual content, transcriptions of spoken texts based on audio recordings, and other similar formats. Corpus types can be differentiated based on their content or source. The terms "synchronic" and "historical" refer to several types of corpora, such as national corpora, learner corpora, academic discourse, children's language, interviews, monitor corpora, multilingual corpora, and web-based corpora (Gries, 2010).

Sinclair (2004) noted that "anyone studying a text is likely to need to know how often each different word form occurs in it." Frequency-sorted word lists are a fundamental component of the conventional approach to leveraging corpora. Tribble and Jones (1997) proposed an approach for utilising texts in the language classroom, suggesting that a word list sorted by frequency is the most efficient initial step in comprehending a text (O'keeffe, McCarthy, & Carter, 2007). A frequency-sorted word list documents the frequency of each word in a text, offering valuable insights about the presence or absence of terms in texts.

The corpus analysis approach, as outlined by Liang (2016) in Figure 3, utilises Chinese characters and English words as corpora. This process serves as the primary research foundation for the current study.

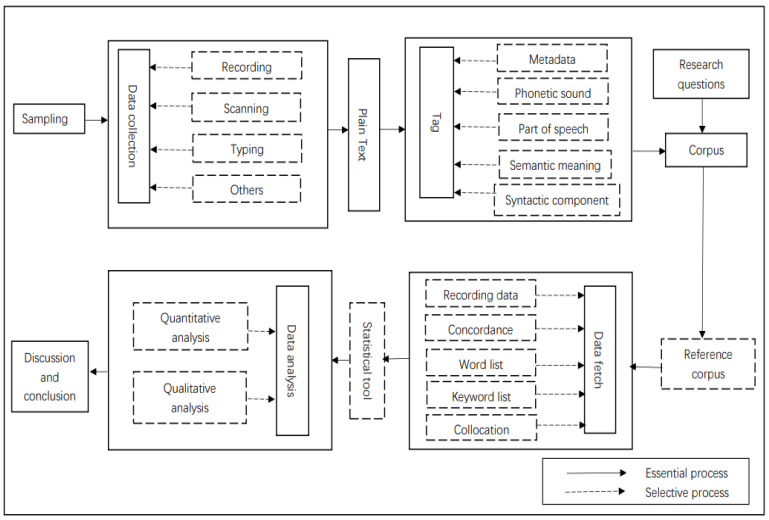

Figure 3: *Process of Corpus Analysis* (Liang, 2016).

Utilising corpus analysis in the process of learning the Thai language will yield effective lexicon resources for acquiring Thai vocabulary. According to a study by Nation (2016), before making a decision, teachers and material authors should weigh the advantages and disadvantages of teaching or including particular vocabulary items. He asserts that word lists are fundamental to the design of effective vocabulary courses, the creation of graded listening and reading materials, vocabulary research, and vocabulary assessments.

4. Research Methodology

This study collects data from four sets of Thai textbooks (each consisting of four equivalent volumes) that are currently being used for Thai language instruction in China. The corpora comprise readily available open materials. The purpose of this study is to utilise textbooks as a pedagogic corpus for analysing the language employed in written textbooks. The corpus is defined as a tool for researching and comparing word frequency and distribution within each textbook and volume. By utilising freeware such as AntConc for corpus analysis, one can determine the answers by computing the frequency of each word in comparable statistics.

The file names of each textbook in this research are abbreviated based on their Chinese Pinyin phonetic alphabet, as indicated in Table 1. To ensure a consistent and symmetrical abbreviation for each set of ordered textbooks, a format consisting of four letters has been established. When naming the first set of textbooks, the researcher took the first two letters "ty" and added the two letters abbreviated from its publishing place, "Beijing," to create the abbreviation "tybj (sansarn, 2023)." To create the name for the fourth set, the researcher selected the initial four letters from the lengthy Chinese Pinyin and formed it as "dxty".

Table 1

The File's Name of Each Textbook.

No.	Textbook in Chinese title	Translated title (informal)	The name read by Chinese Pinyin	Vol.	File
1	《泰语》	Thai	tài yǔ	one two three four	tybj-a tybj-b tybj-c tybj-d
2	《基础泰语》	Basic Thai	ji chǔ tài yǔ	one two three four	jcty-a jcty-b jcty-c jcty-d
3	《泰语教程》	Thai Course	tài yǔ jiào chéng	one two three four	tyjc-a tyjc-b tyjc-c tyjc-d
4	《大学泰语综合教程》	College Thai	dà xué tài yǔ zong hé jiào chéng	one two three four	dxty-a dxty-b dxty-c dxty-d

When there is a conversation, the file's name will include the letter "c" with a number in turns, such as "tybj-a01-01c, tyjc-b02-02c, jcty-a10-01c, jcty-a10-02c ". Similarly, if the file is from a passage, its name will be entitled by the letter "p" with a serial number, like "tybj-c11-01p, dxty-d17-01p, tyjc-d09-02p". The Figure 4 is a sample of marking a file's name.

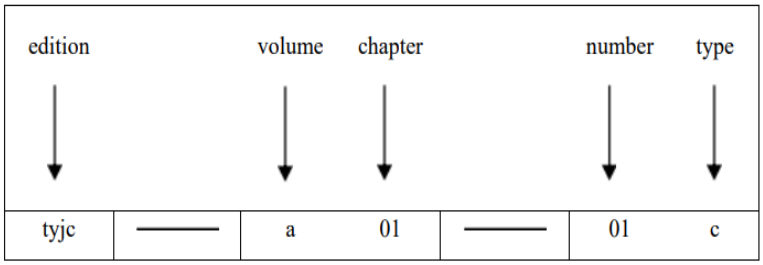

Figure 4: *The File's Name of Each Textbook.*

The initial step in conducting corpus analysis involves creating a database. This process encompasses data collection, converting the data into plain text, segmenting the text, and revising the data. The process is depicted in the subsequent flowchart (Figure 5).

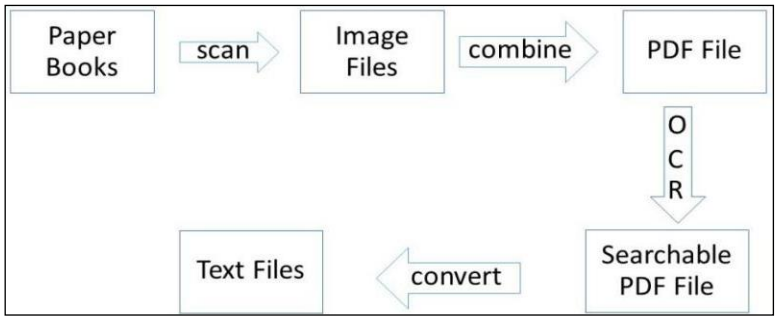

Figure 5*: Data Collection Process.*

Once data is collected and stored on a computer, it is necessary to convert the raw text data into plain text. A clean text should be free of non-word elements, excessive spaces, meaningless symbols, and formatting errors. The following figures show the data collection process from Raw text, → Messy text, → Clean text.

Segmentation will be performed using dedicated software and online tools. In this case, the Thai dictionary will serve as a reference for the classification of word types. A frequency list consists of words that are present in a corpus or a subset of a corpus, along with their corresponding occurrence frequencies (Anthony, 2019). The composition of these lists relies on the input data from a corpus, and the words mentioned in corpus linguistics must be identifiable by a computer. Before running the programme for analysis, it is necessary to identify word tokens in each file at the computer's level of word recognition. This is because the computer can only process strings of characters (Aroonmanakun & Rivepiboon, 2004). In such instances, it is necessary to clearly establish the definition of a word as a sequence of letters or characters enclosed by spaces or symbols such as a slash "/" or a vertical line "|." This procedure is referred to as tokenization.

Thai words differ significantly from English in terms of spacing, as a word is not typically followed by a space unless it is essential for clarity or to indicate a pause. The process of precisely specifying the meaning of a word to a computer requires further clarification. In addition to utilising a programme for Thai tokenization, manual tokenization is also necessary. Performing manual data revision is crucial when errors are present in the input data. The process is intricate and requires a significant amount of time. The tokenization (also known as segmentation) process in this study mostly utilises LexTo (a Thai Lexeme Tokenizer Programme and a free application on the website). Tokenization can be performed either through the online interface or by using a downloaded desktop application. Nevertheless, the procedure remains only partially automated due to the occurrence of unforeseen errors, such as inaccurate spelling, variations in spelling within loaded words, and wrong spacing in the proper names of individuals, locations, businesses, goods, and other entities.

This study utilises a basic **concordancer programme** for corpus analysis. AntConc is freely available software that can be downloaded from the Internet. It provides all the necessary features for corpus analysis, including the ability to upload files to create a corpus, retrieve word forms, view concordance lines, and access the wider context. AntConc is a freely available concordance programme created by Prof. Laurence Anthony,

who serves as the Director of the Centre for English Language Education at Waseda University in Japan. There are versions available for Windows, Mac, and Linux. The program can be downloaded from the following website.

*http://www.antlab.sci.waseda.ac.jp/AntConc._index.html

AntConc can only recognise files that are encoded. Replicate the textual content within .txt files and store them using one of the subsequent encoding formats: ANSI, Unicode, UTF-8, UTF-16. English texts may be saved as ANSI or UTF-8 files. With practical tests, Unicode is suggested that Unicode for Thai.

Using analysis conducted with AntConc, generating a targeted word list based on the established research questions is both efficient and expeditious. This study employed a word frequency list comprising collected data to investigate the correlation between type and token across various proficiency levels and textbooks.

Lexical density is determined by the mean number of unique words and the proportion of total words in each text. The collected data will be analysed quantitatively and qualitatively to identify differences in type-token ratio and lexical variety across all volumes and textbooks (Mahlberg, 2006).

The TTR, or type-token ratio (Templin, 1957), is a metric used to assess the level of vocabulary diversity in a given set of input data. The type-token ratio is a useful metric for assessing the diversity of vocabulary in a given text. It has the potential to track variations in lexical diversity. The lexical diversity of a language segment can be calculated by dividing the total number of unique words (types) by the total number of words (tokens), as depicted in Figure 6. A high TTR signifies a significant amount of lexical diversity, whereas a low TTR suggests the opposite (Aroonmanakun & Rivepiboon, 2004).

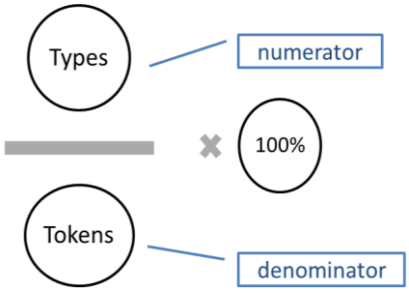

Figure 6: *Type-Token Ratio.*

The type-token ratio, in theory, accounts for the vocabulary diversity relative to the sample size. Increasing the sample size is necessary, as it is reasonable to assume that a larger sample would yield a higher probability of encountering a greater variety of words. A TTR indicates less repetition in vocabulary usage. The 'ideal' type-token ratio (TTR) of a speech sample with 50 different words is calculated as 1 (100%) by dividing 50 by 50. In contrast, repeating the same word 50 times in the sample results in a frequency of 1/50, or 0.02 (2%). TTR thus appears to be a useful indicator of lexical diversity that is independent of sample size. It has been utilised in numerous studies as a potential diagnostic tool for language impairment.

TTR, calculated from a consistent number of tokens, can serve as a valuable research tool in corpus analysis. All words in this study were counted as tokens. This method is suitable for various research purposes in this study, such as comparing specific Thai texts in terms of size or competence. When acquiring a second language, students need to learn each morphological word separately and consider it as a distinct and unique entity (Chiu, 2013).

The type-token ratio can be employed to track variations in vocabulary usage across different study periods, like lexical density. This study focuses on individuals who are beginners in learning the Thai language and advanced learners who have experienced strokes resulting in word retrieval and naming difficulties. This research aims to investigate the underlying factors that contribute to the relationship between type frequencies and sample sizes in vocabulary, whether it is considered normal or abnormal.

What are the consequences of using testing samples with varying sizes? Can individuals assess their lexical richness using the type-token ratio (TTR)? Yes, the answer is affirmative, provided that it adheres to the standard type-token ratio (STTR). Normalisation is employed to standardise the results. The default method for calculating the standardised type/token ratio (STTR) is to analyse a wordlist in each text file, with a frequency count conducted every 1000 words (Thomas, 2005). The ratio is computed for the first 1,000 words, the subsequent 1,000 words, and so on until the conclusion of the text or corpus. A running average is calculated by determining the average type/token ratio for consecutive 1,000-word segments of text. The provided formula is capable of calculating the STTR (Richards, 1987).

$$\mathbf{STTR = (TTR^{1st\ 1000} + TTR^{2nd\ 1000} + TTR^{3rd\ 1000} + TTR^{n...th\ 1000}) / n}$$

Consequently, corpus-linguistic research has become increasingly thorough and accurate. Quantifying linguistic data enhances the study of language as a multifactorial subject and facilitates the connection between corpus-based and experimental findings (Chomsky, 2000).

5. Result

5.1 Initial TTR comparison

During the initial session, the researcher conducted a general comparison of textbooks and volumes. This included a total of four volumes from four sets of textbooks. Table 2 presents TTRs for four sets of textbooks, allowing for a general comparison of word richness.

Table 2

TTR Comparison Among Textbooks.

Textbooks	Word types	Word tokens	TTR (%)
jcty	6773	46451	14.58
tyjc	9186	100494	9.14
dxty	10461	99015	10.57

Based on the data presented in the table, textbook-jcty exhibits the highest vocabulary diversity with a ratio of 14.58%. Out of the four collected textbooks, the textbook-jcty text exhibits the highest lexical density, indicating that it possesses the most diverse vocabulary. When comparing the word token counts in the respective columns, there is a notable disparity observed among the textbooks. Textbook-jcty has the smallest vocabulary size, while textbook-tyjc has the highest number of tokens. Textbook-dxty and tybj follow with intermediate token counts.

Table 3

TTR Comparison Among Volumes.

Volume	Level	Type	Token	TTR (%)
a	Low	2081	13589	15.31
b	Mid	7861	70857	11.09
c	Mid-high	10316	104531	9.87
d	High	12500	132978	9.40

Table 3 presents a comparison of TTRs across various text levels in four textbooks, ranging from volume-a to volume-d. The TTR analysis results in the table above provide the average lexical density for different volume levels, regardless of the length of each text. The entry-level volumes exhibit the highest TTR. Providing explanations for each level's content can be beneficial in enhancing understanding. The primary focus of volume-a in Thai language learning is phonetics and word spelling rules, as it primarily consists of syllables, words, and phrases. The words at this level primarily focus on pronunciation and spelling, resulting in a greater range of words compared to other volumes. This is because the beginner-level words are listed without accompanying passages for practicing and enhancing pronunciation skills. Words with diverse spelling rules are chosen, resulting in a lower frequency of word repetition compared to those found in conversations or texts.

Table 4 compares the TTR among the same volumes in different textbooks. Each textbook consists of four volumes, ranging from basic low to advanced high level.

Table 4

TTR Comparison Among Textbooks.

Volumes	Textbooks	Word types	Word tokens	TTRs (%)
a	tybj	1224	5203	23.52
	jcty	474	2490	19.04
	tyjc	956	4505	21.22
	dxty	638	1391	45.87
b	tybj	2649	12623	20.99
	jcty	1803	8108	22.24
	tyjc	2558	14353	17.82
	dxty	5747	35701	16.1
c	tybj	4891	28100	17.41
	jcty	2296	14278	16.08
	tyjc	4611	33178	13.9
	dxty	4695	28975	16.2
d	tybj	4469	30069	14.86
	jcty	4436	21503	20.63
	tyjc	6645	48458	13.71
	dxty	5205	32948	15.8

Among the four textbooks examined, dxty-a exhibits the highest TTR at 45.87%. This indicates that the texts within this particular volume of the textbook are composed of 638 unique words, which are reused to account for 45.87% of the total word count of 1391.

tyjc-c and tyjc-d show the lowest TTR as 13.9% and 13.71%, respectively, corresponding to type:token of 4611:33178 and 6645:48458, which means the words in these two volumes are reused nearly or more in 13.9% and 13.71% amount of words in all texts.

Based on the TTR comparison result presented in the table above, it can be observed that dxty-b, dxty-c, and dxty-d exhibit similar TTR values. This indicates that the lexical density of these three volumes within the textbook-dxty is distributed in a manner that tends to achieve balance.

When the volumes are organised in ascending order and their TTRs are graphed from left to right, the relationship between volumes and textbooks becomes evident. If there is a negative correlation between low-level and TTR and a positive correlation between high-level and TTR, it is expected that the graph line will exhibit a gradual upward slope. Contrary to expectations, Figure 7 demonstrates that this is not true. The previous paragraph has already provided an explanation for why low-level volume A exhibits a high TTR. This is because the content of texts at this level relies heavily on words that are pronounced. There is no discernible correlation between the volume level and the presence of textbooks. It would be an inaccurate assertion.

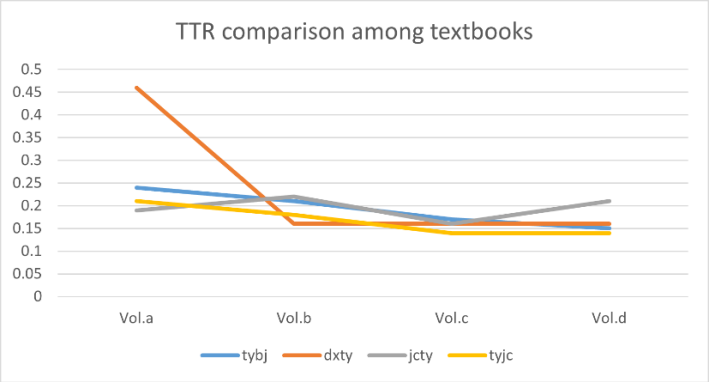

Figure 7: *TTR Comparison Among Textbooks.*

The graph above illustrates the comparison of TTRs obtained from various token counts. The curves of the textbook's lexical density from volume-a to volume-d exhibit a gradual decline, indicating a decrease in the rate of lexical density as the level increases. This trend contradicts the expected pattern. The TTRs graph lines for the textbooks of dxty and tyjc exhibit a comparable pattern of downward slope at lower levels and a straight trajectory at subsequent levels. The textbook-jcty exhibits fluctuating TTR, which complicates the identification of volume correlations. The variability in TTR tendency is attributed to the variability in token count within each volume. Comparing texts of different lengths or token counts is akin to comparing dissimilar things. The raw TTRS provide information solely on the vocabulary size of the collected data, specifically the percentage of unique words in relation to the total number of words.

5.2 STTR Comparison

The observations depicted in the graphs can be attributed to statistical significance. The sample size has an impact on the TTRs. Sample size predicts TTR. This also demonstrates the impact of contextual variables on a small sample size, specifically the differences between samples from volume A and samples from volume. Assume that the data is plotted in a manner like the initial set. In such instances, the outcome will exhibit a consistent slope either upwards or downwards towards the right, indicating a correlation between the volume level and the TTR of the textbooks.

Examining TTR could be more efficient by counting tokens from texts of varying lengths. In addition to the issue of high word reoccurrence rates in texts, there is also a problem with varying sample sizes in testing. Using a wordlist that includes TTR can provide a more effective approach. It will be applied to calculate standard TTRs (hereafter, this text will be abbreviated as STTR) to normalise the results.

To obtain more accurate results, the researcher conducted a search for texts containing the STTR measure to investigate potential variations in vocabulary richness.

Table 5 displays the methodology employed to calculate the STTR for each textbook. The table includes the total number of types and tokens found in the entire volume. The purpose of this analysis is to investigate whether there are any discrepancies in vocabulary richness outcomes when comparing previous TTRs with the newly calculated STTRs in this particular section.

Table 5

STTR Comparison Among Textbooks.

Textbooks	Types	Tokens	TTRs (%)	STTRs (%)
tybj	8328	75995	10.96	60.48
jcty	6773	46451	14.58	61.82
tyjc	9186	100494	9.14	60.32
dxty	10461	99015	10.57	59.48

The comparison of STTRs and ordinary TTRs reveals that the STTRs in four textbooks, which include the full content of texts, align with the findings of previous TTRs, as depicted in Figure 8. Textbook-jcty maintains the highest lexical density rate. The STTR testing methods have led to a reduction in the gaps between the TTRs of the four textbooks. The graph shows that the STTRs curve is inverted and closely aligned with the TTRs curve, exhibiting a nearly straight line with comparable numerical values.

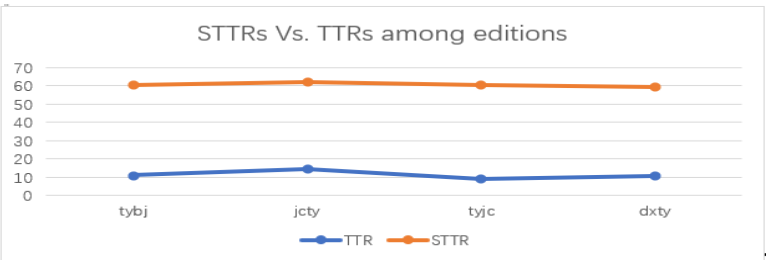

Figure 8: *Sttrs and Ttrs Comparison Between Textbooks.*

Higher puncture ratios are associated with lower repetitive lexical density, if the text level remains constant. This correlation can be further confirmed by analysing additional graphs that calculate the token count of each text for complete data.

As the token count increases, it is necessary to exclude volumes a and b from the subsequent sections of the TTRs study due to the predominantly word and phrase content in volume a. Conversely, the texts within the volume exhibit a combination of dialogues and passages. There are notable variations in the number of tokens. Volumes c and d consist solely of passages that can be used for efficient text comparisons. The following analysis selected only passages in volumes c and d as the testing samples.

When STTR was applied to compare volume-c and volume in the same textbook, the testing samples are shown in Table 9.

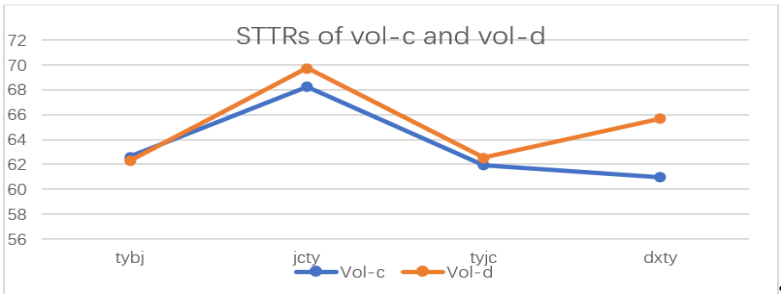

Figure 9: *STTR For Vol. C Against Vol.D in Four Textbooks.*

The graph above illustrates that the STTRs of texts in volumes d were consistently higher than those in volumes c across four textbooks. The correlation between the assumed outcome and STTR was further supported by the token count of STTR calculated at each volume in the mid-high and high levels.

Preliminary data analysis was inconclusive, as it did not find any correlation between lower- and higher-level texts when the texts were not of ideal equivalent length. By taking into account two important factors, namely token count and content, the STTR method can yield more accurate results. Research has demonstrated a positive correlation between the level of texts and their lexical complexity. Among the four textbooks examined, the textbook just demonstrates the greatest vocabulary variety at the mid-high and high levels.

5.3 TTR Comparison After Applying a Stop List.

The calculation of TTRs and Standardised STTRs provides a measure of lexical diversity in a given text or utterance. However, the assessment of lexical complexity requires more than just considering the type-token count. Function words (including conjunctions, prepositions, modal particles, and interjections) are frequently observed in the wordlist. They occur with a significant frequency in both function and content texts. These words increase the number of tokens and **tend to** decrease the TTR (Wang, 2011). Another test conducted in this research involved the application of a **stop list** consisting of Thai function words in TTR comparisons. The TTR will be calculated solely using the content words in this section.

There is an additional explanation provided in the testing samples. In addition to function words, a stop list of English words is also utilised in this section's comparisons to assess the lexical richness of Thai words more effectively in texts.

Figure 10 will compare textbooks and volumes. A selection of texts with similar lengths will be chosen for pairwise comparison to provide further clarification. If the results are consistent with previous findings, the answer can be determined through the subsequent analysis.

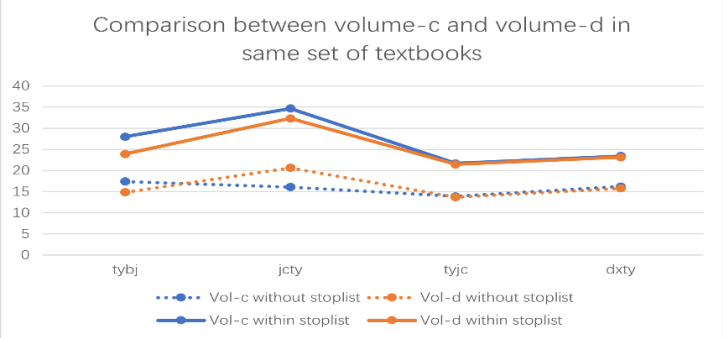

Figure 10: *TTR Comparison Between Vol. C And Vol.D In the Same Textbook.*

The graph above does not demonstrate the anticipated increase in lexical richness at higher levels. Similarly, it reveals that texts in higher levels of volume-d exhibit lower lexical variety compared to volume-c. The most notable difference between the two sets of statistics lies in the turning point of volume-c in textbooks. Specifically, after applying the stoplist, the curve of volume-c without the stoplist shifts to follow the same upward trend as the line of volume.

5.4 TTR Comparison Between Texts with the Same Length

There are four pairs of texts that have the highest similarity in terms of their length (quantity of tokens). The main passage is denoted as -01p, while additional passages are denoted as -02p, -03p, -04p, and -05p in sequential order. In this section of the research, passages with similar marks (e.g., -01p or -02p) were chosen from relative volume-c and volume in four textbooks. The purpose was to investigate whether the results would align with the previous comparison, which did not involve a stoplist.

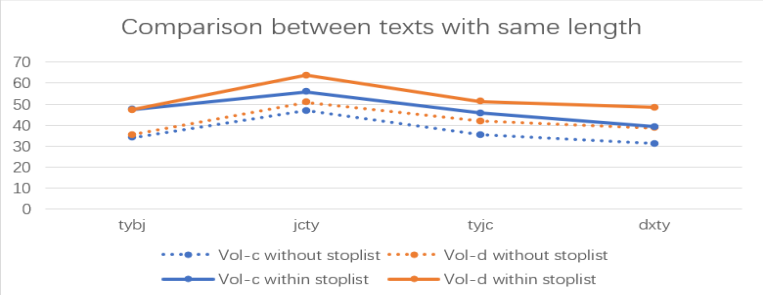

Figure 11: *TTR Comparison Between Texts with The Same Length in The Same Textbook.*

Figure 11 illustrates the comparison outcome of texts with similar attributes in terms of volume-c, volume before, and volume after the implementation of the stoplist. The relative lines in the text exhibit a similar pattern to those observed in STTR. Consequently, textbook-jcty continues to display the highest level of lexical complexity among the four textbooks. Text volumes were generally higher in all textbooks compared to volume-c. Texts at higher levels are expected to have a higher TTR and greater lexical complexity when tested with equivalent samples. This correlation finding was further validated in the subsequent analysis of this section.

In response to the lack of expected growth in vocabulary richness using typical comparison methods, the researcher attempted two additional methods for further comparisons. One of these methods involved the application of STTR analysis. Another option is to use a stoplist while using AntConc to eliminate function words from all texts and only compare content words in texts of similar lengths.

6. Conclusion

The average number of content words and the proportion of content word tokens in a text are what determine vocabulary richness, or lexical density. The TTR analysis result only provides the average lexical density for different text lengths. The initial analysis reveals that low-level textbooks exhibit the highest TTR among the volumes examined in the first section. Volumes-a has the largest vocabulary, consisting of individual syllables, words, and phrases used for pronunciation practice. The words at the beginner level are listed without any accompanying conversations or passages. Words in the English language are subject to different spelling rules, resulting in a lower frequency of occurrence compared to words found in mid-high-level textbooks.

When comparing TTR among the same levels, the lexical density decreases as the level increases, contrary to the expected outcome. The unpredictable TTR tendency arises from the varying token count in each volume. Comparing texts of different lengths or token counts is akin to comparing dissimilar things. The raw TTRs provide information solely on the vocabulary size of the collected data, specifically the number of unique word types concerning the total number of word types.

To ensure accurate comparisons of TTRs across texts, the study employed three methods for additional analysis. Initially, the approach involved selecting pairs of texts that had similar lengths and solely comparing the reading passages between volume-c and volume. The second approach involved comparing it with a standardised TTR. The last step involved comparing the results after applying a stoplist of Thai function words.

Therefore, when using the same set of testing samples, texts at higher levels are expected to have a higher TTR, indicating greater lexical complexity. The comparisons made by STTR, after applying a stoplist to remove non-content words, align with the test results. The TTR of texts at higher levels was consistently higher than that of texts in lower-level volumes across all textbooks. The correlation result was further validated in the analysis conducted in this section.

7. Limitation

7.1 Problems with Segmentation

During the data collection phase, unforeseen challenges were encountered in the segmentation process. Proper names and abbreviations were commonly fragmented into multiple words or letters. To address segmentation errors encountered in this study, the researcher employed a method of manual self-checking and verification with Thai native speakers. Nevertheless, the output list may still contain unforeseen errors that could be attributed to a computer application or website service. Despite implementing manual double-checking measures, it is inevitable to encounter errors during the process of data collection.

7.2 Problem with TTR in the Above Research

The significance of considering the total token count is evident in the analysis conducted in Chapter IV. The type/token ratio (TTR) varies significantly based on the length of the text or corpus under examination. A 1,000-word article typically has a Type-Token Ratio (TTR) of 40%, while a shorter article may have a TTR of 70%. A larger corpus of 4 million words is likely to have a TTR of approximately 2%. Type/token information, while provided in a Wordlist statistics display, is often considered insignificant. The conventional TTR is useful for analysing a corpus consisting of numerous text segments that are of equal size (Breyer, 2011). In practical research scenarios, particularly when the focus is on the text rather than the language, it is common to encounter texts of varying lengths. In such cases, the conventional TTR may offer limited assistance.

7.3 Suggestions for Future Research

The research commenced by quantifying the number of types and tokens in texts, employing a systematic approach that involved analysing each text individually as well as considering volumes and textbooks as separate units of analysis. This method provides the most efficient means of obtaining vocabulary information. Additionally, lexical richness can be assessed by comparing TTR through statistical analysis involving word count. The chapter on high-frequency word analysis also employed word frequency counts to determine the study's results. In addition to keyword analysis, the keywords were searched using computational methods (Charalampidou, 2021).

Word frequency analysis is a complex task that requires a theory-specific approach. The choice of counting and normalisation basis should be suitable for the specific language feature being studied and should align with the theoretical framework employed in the research. Various features necessitate distinct metrics and statistical analyses. The use of visual representations of counts can facilitate the process of analysis.

In addition to the application of corpus linguistics methodology, artificial analysis played a crucial role in this study. The study of language encompasses various aspects such as phonetics, semantics, syntax, pragmatics, and related fields in linguistics. These areas of study involve understanding the sounds and meanings of words, sentence structure, and overall comprehension of language. The corpus analysis used in this study allowed for a comprehensive and varied selection of relevant texts that are likely to be representative of

texts in general. Future research should aim to replicate these findings by employing word lists instead of larger corpora. In order to appropriately investigate the null hypothesis, it is necessary for corpora to include genre metadata.

Teachers should adopt a comprehensive approach to vocabulary instruction and assume responsibility for both explicit and incidental vocabulary development. Recognising the gradual process of vocabulary acquisition is crucial, as is recognising the importance of a comprehensive and sustained vocabulary learning programme that emphasises high-frequency words. There is no definitive "best" teaching method, but teachers can effectively enhance learning outcomes by prioritising sustained engagement with language.

Thai learners in China, particularly Thai primary students, are required to acquire proficiency in a new language within a span of 3–4 years, primarily using textbooks. This study involved constructing a database consisting of 16 textbook volumes. The purpose of this database is to offer guidance for educators and students in the process of selecting appropriate texts.

8. Acknowledgements

I express my gratitude to all individuals who have provided assistance during the course of my research and thesis writing. The support of my family and advisors at the Research Institute for Languages and Cultures of Asia at Mahidol University has been instrumental in the success of this thesis. I express my sincere gratitude to my supervisor, Dr. Isara Choosri, for his valuable lectures and guidance, which greatly benefited me. I would like to express our gratitude to Professor Somsonge Burusphat, Associate Professor Siripen Ungsitipoonporn, and Assistant Professor Pattama Patpong for their valuable assistance, insightful guidance, and extensive expertise. I am deeply grateful to my family for providing me with both emotional support and financial assistance.

Bibliography

Allen, V. F. (1983). *Techniques in Teaching Vocabulary*. ERIC. https://eric.ed.gov/?id=ED242213

Anthony, L. (2019). *Laurence Anthony's AntConc*. Laurence Anthony. https://www.laurenceanthony.net/software/antconc

Archer, D. (2016). Does frequency really matter? In *What's in a Word-list?* (pp. 1-15). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315547411-1

Aroonmanakun, W. (2007). Creating the Thai national corpus. *MANUSYA: Journal of Humanities, 10*(3), 4-17. https://doi.org/10.1163/26659077-01003001

Aroonmanakun, W., & Rivepiboon, W. (2004). A unified model of Thai word segmentation and romanization. In *Proceedings of The 18th Pacific Asia Conference on Language, Information and Computation* (pp. 205-214). Logico-Linguistic Society of Japan. https://doi.org/http://hdl.handle.net/2065/574

Baayen, R. H. (2001). *Word frequency distributions*. Springer Dordrecht. https://doi.org/10.1007/978-94-010-0844-0

Barcroft, J., Sunderman, G., & Schmitt, N. (2011). Lexis. In *The Routledge handbook of applied linguistics* (pp. 571-583). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203835654-46

Breyer, Y. A. (2011). *Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges. English Corpus Linguistics. Volume 13*. ERIC. https://eric.ed.gov/?id=ED530947

Brown, H. D. (2001). *Principles of language learning and teaching*. Beijing: Foreign Language Teaching and Research Press. https://www.longmanhomeusa.com/catalog/products/principles-of-language-learning-and-teaching

Charalampidou, P. (2021). The use of corpora in an interdisciplinary approach to localization. In *Proceedings of the Translation and Interpreting Technology Online Conference* (pp. 216-226). INCOMA Ltd. https://aclanthology.org/2021.triton-1.25

Cheng, W. (2011). *Exploring corpus linguistics: Language in action*. Routledge. https://doi.org/10.4324/9780203802632

Chiu, Y.-H. (2013). Computer-assisted second language vocabulary instruction: A meta-analysis. *British Journal of Educational Technology, 44*(2), E52-E56. https://doi.org/10.1111/j.1467-8535.2012.01342.x

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press. https://doi.org/10.1017/CBO9780511811937

Davies, A., & Elder, C. (2004). *The handbook of applied linguistics*. Wiley Online Library. https://doi.org/10.1002/9780470757000

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143-188. https://doi.org/10.1017/S0272263102002024

Folse, K. S. (2004). The underestimated importance of vocabulary in the foreign language classroom. *CLEAR News, 8*(2), 1-6. https://tungumalatorg.is/katla/files/2012/11/Keith-Folse.pdf

Gries, S. T. (2009). Corpus-Based Methods in Analysis of Second Language Acquisition Data. In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 406-431). Routledge. https://doi.org/10.4324/9780203938560-25

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In *Corpus-linguistic applications* (pp. 197-212). Brill. https://doi.org/10.1163/9789042028012_014

Gries, S. T. (2012). Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", 36*(3), 477-510. https://doi.org/10.1075/sl.36.3.02gri

Gries, S. T. (2014). Corpus and quantitative methods. In *The Bloomsbury companion to cognitive linguistics* (pp. 279-300). Bloomsbury Publishing. https://stgries.info/research/2014_STG_CorpAndQuantMeth_CompToCogLing.pdf

Hu, G. (2005). Building a strong contingent of secondary English-as-a-foreign-language teachers in China: Problems and policies. *International Journal of Educational Reform, 14*(4), 454-486. https://doi.org/10.1177/105678790501400408

Klein, W. (1986). *Second language acquisition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815058

Kübler, S., & Zinsmeister, H. (2015). *Corpus linguistics and linguistically annotated corpora*. Bloomsbury Publishing. https://www.bloomsbury.com/us/corpus-linguistics-and-linguistically-annotated-corpora-9781441164476

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30. http://hdl.handle.net/10125/66648

Leech, G. (1992). Corpora and theories of linguistic performance. In S. Jan (Ed.), *Directions in Corpus Linguistics* (pp. 105-126). De Gruyter Mouton. https://doi.org/10.1515/9783110867275.105

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*. Hove UK: Language Teaching Publications. https://www.tesl-ej.org/ej09/r10.html

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory, 38*(6), 1842-1845. https://doi.org/10.1109/18.165464

Liang, M. C. (2016). *What is Corpus Linguistics*. Shanghai: Shanghai Foreign Language Education Press.

Liu, Y. (2015). Foreign language education planning in China since 1949: A recurrent instrumentalist discourse. *Working Papers in Educational Linguistics (WPEL), 30*(1), 65-85. https://repository.upenn.edu/handle/20.500.14332/49566

Mahlberg, M. (2006). Lexical cohesion: Corpus linguistic theory and its application in English language teaching. *International Journal of Corpus Linguistics, 11*(3), 363-383. https://doi.org/10.1075/ijcl.11.3.08mah

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge. https://www.routledge.com/Corpus-Based-Language-Studies-An-Advanced-Resource-Book/McEnery-Xiao-Tono/p/book/9780415286237

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208

Nation, I. S. P. (2022). The Cambridge Applied Linguistics Series. In *Learning Vocabulary in Another Language* (pp. ii-iv). Cambridge University Press. https://doi.org/10.1017/9781009093873

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy, 14*(1), 6-19. https://www.lextutor.ca/research/nation_waring_97.html

O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press. https://doi.org/10.1017/CBO9780511497650

Read, J. A. (2000). *Assessing vocabulary*. Cambridge University Press. http://www.cambridge.org/9780521627412

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Languages, 14*(2), 201-209. https://doi.org/10.1017/S0305000900012885

sansarn. (2023). *LexTo : Lexto (Thai word cutting program, Thai word division)*. http://www.sansarn.com/lexto

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329-363. https://doi.org/10.1177/1362168808089921

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching1. *Language Teaching, 47*(4), 484-503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press. https://doi.org/10.1017/9781108569057

Schreuder, R., & Weltens, B. (1993). *The bilingual lexicon*. John Benjamins Publishing. https://doi.org/10.1075/sibil.6

Sharoff, S., Rapp, R., Zweigenbaum, P., & Fung, P. (2013). *Building and using comparable corpora*. Springer. https://doi.org/10.1007/978-3-642-20128-8

Sinclair, J. (2004). The search for units of meaning. In *Trust the Text: Language, Corpus and Discourse* (pp. 1000-1032). Routledge. https://doi.org/10.4324/9780203594070-6

Sinclair, J. (2005). Corpus and Text – Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). Oxbow Books. https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora, 8*(1), 81-113. https://doi.org/10.3366/cor.2013.0035

Templin, M. C. (1957). *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press. https://www.upress.umn.edu/book-division/books/certain-language-skills-in-children

Thai National Corpus. (2007). *Thai National Corpus Project: Principles and reasons*. https://www.arts.chula.ac.th/ling/tnc

Thomas, D. (2005). *Type-token Ratios in One Teacher's Classroom Talk: An Investigation of Lexical Complexity*. United Kingdom: University of Birmingham. https://www.birmingham.ac.uk/documents/college-artslaw/cels/essays/languageteaching/daxthomas2005a.pdf

Traugott, E. C. (2012). Geoffrey Leech, Marianne Hundt, Christian Mair and Nicholas Smith, Change in contemporary English: A grammatical study. Cambridge: Cambridge University Press, 2009. Pp. xxviii+ 341. ISBN 978-0-521-86722-I. *English Language & Linguistics, 16*(1), 183-193. https://doi.org/10.1017/S1360674311000359

Tribble, C., & Jones, G. (1997). *Concordances in the classroom: A resource guide for teachers*. Athelstan.

Wang, X. (2011). The bilingual lexicon: models and implications. In R. K. Mishra & N. Srinivasan (Eds.), *Language and cognition interface: state of the art* (Vol. 5, pp. 199-220). LINCOM publishers. http://dx.doi.org/10.13140/2.1.1422.6886

Wilkinson, M. (2011). WordSmith Tools: The best corpus analysis program for translators? *Translation Journal, 15*(3). http://www.bokorlang.com/journal/57corpus.htm