A Corpus-Based Wordlist of Government Official English Examinations in Thailand

Darunee Yotimart[1*]

[1]General Education Program, Faculty of Sports Communication, Thailand National Sports University, Thailand

[*]Corresponding author's E-mail: ajarnlidatnsu@gmail.com

Received: January 30, 2025
Revised: March 06, 2025
Accepted: March 10, 2025
DOI: http://doi.org/

Abstract

This study sought to identify the most prevalent content terms in government official examinations for English topics and to analyze the ratio of word families in the General Service List (GSL) and Academic Word List (AWL) inside the Government Official Examination Word List (GOEWL). The English subject government official examinations were aggregated to form the Government Official Examination Corpus. Coxhead's frequency criterion for the AWL was applied in selecting words for the GOEWL. All word families that appeared once or several times in the examinations were incorporated into the wordlist. AntWordProfiler software was employed to analyze and produce the GOEWL, comprising 3,085 word families. The findings indicated that 83.38 percent of the GOEWL's word families were part of the first 1,000 of the GSL, 6.22 percent belonged to the second 1,000 of the GSL, 2.41 percent were from the AWL, and the remaining 7.99 percent did not correspond to any wordlists. Students and candidates preparing for government official examinations can build a strong vocabulary through the GOEWL. Educators

may adapt the vocabulary list to suit teaching materials, making it more accessible and relevant for students, particularly at the higher education level.

Keywords: academic word list (AWL), corpus-based analysis, general service list (GSL), government official examination

Introduction

A significant number of Thai people perceive English primarily as a tool for gaining employment rather than for communication (Apichat & Fatimah, 2022). The paramount objective of English language study for most Thais aspiring to enter the government sector is to successfully pass an English examination. Examinations for national government posts in Thailand encompass roles for government officials, police officers, military personnel, government educators, and various other public service positions. Nonetheless, it seems that there is an absence of a book or lexicon that specifically consolidates the vocabulary for each distinct examination. Providing a vocabulary list to assist candidates in preparing for higher education examinations in Thailand is crucial (Zhang & Sukying, 2021). In the official examination conducted by the government, the scoring system allocates a value to the English subject (Wudthayagorn, 2022). Consequently, it is an essential criterion for entry into government agency and official employment. A corpus-based analysis facilitates the study and identification of patterns and messages concealed within various texts, which may not be easily visible in an individual text (Worarattapong & Phoocharoensil, 2023). Thus, the official government-administered test is the most appropriate source for creating a government examination vocabulary list through a corpus-based analysis.

Research Objectives

1. To identify the most frequent vocabulary used in the government official examinations in Thailand

2. To analyze the percentage proportion of the General Word List (GSL) and the Academic Word List (AWL) that covers the vocabulary in the government official examination Word List (GOEWL)

Literature Review

Vocabulary and reading comprehension are essential components of the English test in Thailand. To successfully pass the examination in Thailand, students must acquire a sufficient vocabulary (Cherngchawano & Jaturapitakkul, 2014). It is essential to support students in enhancing their vocabulary to optimize text comprehension and effectively address the vocabulary in examinations. Classifying vocabulary may assist students in organizing their study and enhancing their test preparation. Vocabulary can generally be categorized based on many criteria, such as function words, content words, or parts of speech. The criteria are the frequency of occurrence. Nation and Webb (2011) classified words into four categories based on their frequency of occurrence. The primary purpose of word classification is to provide a foundation for the planning of teaching and learning, as distinct categories of vocabulary require varied instructional and learning approaches. Instructors or course designers must specify the specific terms and quantity that students should be provided with. Simultaneously, course planners must assess if some words should be excluded from students' focus or consideration (Schmitt & Schmitt, 2020).

High-frequency words are fundamental English terms prevalent in daily discourse and all forms of writing as word list is advantageous for language learners

since it facilitates the acquisition of new vocabulary. The General Service List of English terms (GSL), established by West in 1953, comprises a standard compilation of 2,000 high-frequency word families, encompassing both function and content words. Each of the 2,000 words serves as a headword denoting a word family that is only vaguely delineated by West. Approximately 80% of the words in the text are high-frequency terms. Furthermore, the Academic Word List (AWL) analyzed by Coxhead (2000) is a widely recognized and frequently utilized resource in the field of vocabulary study. These terms are frequently seen in diverse scholarly books but are absent in mainstream English. They constitute approximately 9% of the lexical items in an academic text. The list comprises 570 word families, including headwords together with their inflected and derived forms. There are around 3,100 distinct word forms in total. The list was generated by analysing more than 3,500,000 words of text. The words chosen for the AWL are those that often appear across several academic disciplines, including the arts (such as history, psychology, and sociology), business (such as economics, marketing, and management), law, and the sciences (such as biology, computer science, and mathematics).

Furthermore, technical terminology which refers to jargon specific to a certain discipline and varies considerably across different fields is another group of vocabulary classification. Typically, students acquire these terms as they study the specialised subject area. Definitions of terms within this category are available in a specialised technical dictionary pertaining to that particular field. Another vocabulary group that falls outside the aforementioned three categories is the low-frequency words. They encompass highly specialised terminology from several fields, as well as lexicon that is infrequently seen in everyday discourse. Proper nouns are included within this category. This collection of terms is expected to comprise approximately

5% of the total words in an academic document. However, the well-known word lists, the GSL and AWL, have been compared by various researchers in various texts and examination, each serving a specific purpose, in order to collect and construct an additionally specific word list. All of them are often designed to serve as a foundation for language instruction or the development of educational resources.

In terms of the word list of Thailand national examination, Cherngchawano and Jaturapitakkul (2014) investigated university admission tests in Thailand that represent the country's education system. To pursue higher education, all students typically must undertake the University Admission Tests administered by the National Institute of Educational Testing Service (NIETS). This research aims to perform a word analysis of the distribution and linguistic characteristics of Thailand University Admission Tests. Fifteen papers including 55,161 running words were examined within the context of two word lists: the General Service List (GSL) and the Academic Word List (AWL). The findings indicated that the coverage of the General Service List (GSL) and the Academic Word List (AWL) is 85.05% and 4.58%, respectively. The incorporation of the GSL and the AWL encompasses 89.63% of the texts. The coverage and reading comprehension at a 4,000-word level facilitates reasonable understanding of 94.82% of the materials. Both the GSL and the AWL are recommended as valuable resources for students to prepare for the test and to engage in higher university studies.

Similarly, Chanasattru and Tangkiengsirisin (2017) examine the distribution and coverage of vocabulary in the New General Service List (NGSL) and the Academic Word List (AWL) among social science research publications. Sixty-four open access English social science research articles published between 2013 and 2015 in the ScienceDirect General category were selected and collected into the

Social Science Corpus (SSC). The AntWordProfiler 1.4.0 was employed to determine the frequency and coverage percentage of words from the two lexical lists. Word families from levels 1 and 2 of the NGSL were employed in over 70 percent of instances, but level three word families constituted over 60 percent of the total SSC. Likewise, 99.65 percent of the AWL word families were identified. The NGSL word families constituted over 70 percent of the coverage, whereas the AWL word families comprised over 14 percent, indicating substantial representation from both word lists. The top 10 NGSL word families corresponded to the subject areas of the journals from which they originated, but the top 10 AWL word families were utilized more frequently and associated with social science study domains. The discovery of extensive distributions and coverage confirmed that the NGSL and the AWL substantially enhance vocabulary instruction in equipping students for reading and writing social science research articles.

Khany and Kalantari (2021) identified 658 academic word families with the greatest frequency in the corpus, termed the Accounting Academic Word List (AAWL). These 658 word families constituted 10.16% of the whole corpus. Subsequent investigation revealed that of the identified high-frequency word families, only 354 corresponded with those enumerated in the Academic Word List (AWL). Furthermore, the 50 most often used terms in the list comprised 3.98% of the whole corpus. The aforementioned terms were included in six distinct word lists across several fields, exhibiting varying frequencies, serving as a foundation for the creation of a composite word list.

In terms of the software program, Crawford and Csomay (2015) suggested two essential programs for analyzing language using corpus data. Both AntWordProfiler and AntConc are owned by Laurence Anthony, a renowned corpus

linguist employed at Waseda University in Japan. He has created several corpus tools and software applications, which he has made freely accessible. Anthony (2022) states that AntWordProfiler is a complimentary instrument used for examining data in corpus-based research, namely for vocabulary profiling and the creation of word lists. Researchers have the flexibility to use an unlimited number of texts for analysis with this program. AntWordProfiler has several features for creating wordlists, including the ability to categorize words by word families or word kinds and sort the word list by range or frequency. In addition, it offers researchers access to two well-known wordlists: the General Service List (GSL) by West (1953) and the Academic Word List (AWL) by Coxhead (2000). This allows researchers to exclude items from the GSL and AWL while constructing their own wordlist.

Additional information on the government official examination's criteria, test patterns, and word choices should be found. Over the course of several testing cycles, a substantial volume of data should have been collected and examined to get a precise outcome. The requirements outline the exact kind of vocabulary or grammatical structures that will be required and used for the future test. Based on the aforementioned facts, it elucidates the rationale for the significance of this research. The significance of the government official test for Thai examiners and the reasons why it is the exclusive focus in this research based on a corpus.

Methodology

This study used a corpus-based research design.

Software for Analysis: To examine the coverage of GSL and AWL items in the government official examination in Thailand, Coxhead's (2018) word family and that of West (1953) had been used as the analytical unit. The AntWordProfiler

(Anthony, 2014) is now the premier program for lexical profiling of texts, offering enhanced data analysis along with many extra features beneficial for scholars (Xodabande & Xodabande, 2020). The program used for lexical profiling of psychology research papers in this study was AntWordProfiler (2014), a freely accessible tool designed for measuring the vocabulary level and textual complexity. The GSL and AWL are the preloaded word lists included with AntWordProfiler. The software analyzes the texts inputted into the computer by comparing them to a collection of vocabulary level lists. It then outputs comprehensive vocabulary statistics and detailed frequency information on the corpus (Kongcharoen, 2023).

Compiling the Government Official Examination Corpus (GOEC): The Government Official Examination Corpus (GOEC) was derived from authentic government official exams. This research utilized a total of 10 government official tests conducted between the years 2014 and 2024. A selection of the examinations might be accessed on the official website of the Office of the Civil Service Commission. The Office of the Civil Service Commission was the entity responsible for delivering educational services related to testing and measuring. The official website limitedly offered a simulated examination. In order to create a comprehensive word list, it was necessary to include all government official examinations. Thus, alternate sources such as websites and pdf file were additional options for acquiring the data for other position types. This research used both authentic and simulated exams that were digitized in the form of images or .pdf files. The files were re-uploaded and translated into the plain text format, namely a .txt file, for use in the corpus program. All visual aids, such as photos, graphs, and charts, that were not directly related to the exam questions were omitted from the .txt file. The

count of word tokens and word kinds in each year's examination type are shown in the table 1.

Table 1

The Numbers of Tokens and Types of the Government Official Examination Used

Examination Type	Years	WordType	Word Token
Policeman and Soldier	2017 - 2023	1536	5265
Local Administrator	2017 - 2021	548	1865
Government Teacher	2022 - 2023	3591	24962
Civil Service commission	2020 - 2021	774	2154

Word Selection Criteria: AntWordProfiler 1.4.0w, developed by Laurence Anthony (2014), served as the primary corpus tool for generating the word list in this research. The frequency requirements established by Coxhead for the Academic Word List (AWL) were used in conjunction with the Government Official Examination Corpus (GOEC). The term range was not primarily regarded as the principal criteria before the others, since this wordlist was developed to notify examiners of the most prevalent terms; nevertheless, the actual tests from certain years have been inaccessible for download or retrieval. Consequently, the term range has less significance. This research used Coxhead's frequency criteria, stipulating that each word family must appear a minimum of 100 times in the whole Academic Word List (AWL). The rationale behind the frequency criteria, grounded in empirical analysis, emphasizes the importance of selecting words that are both frequent and widely applicable (Coxhead, 2000). The Government Official Examination Corpus

(GOEC) comprises just 34,246 tokens running words, categorizing it as a notably tiny corpus. Following the use of Coxhead's criteria, terms that appeared just once in the corpus were required to be included into the word list. Consequently, the minimum frequency of each word is once. This computation made it feasible to use Coxhead's word frequency criteria.

Upon completing the criterion establishment, the AntWordProfiler was used to generate a wordlist from government official exams, organizing the word families just by frequency criteria, excluding their range. While the AntWordProfiler has the capability to eliminate vocabulary from renowned wordlists, such the General Service List (GSL) of 1953 and the Academic Word List (AWL) of 2000, this functionality was not applicable to the Government Official Examination wordlist. Given the varied uses and limited quantity of operational terms, their removal was deemed inappropriate.

Function Words and Unrelated Words Removal: Once the first wordlist was generated, all function words had to be manually eliminated from the list. Function words have a diminished significance in the word list due to many factors. Firstly, students are familiar with and have extensively used function words such as propositions and conjunctions during their years of study in secondary education (Ward, 2009). It was anticipated that they would possess the ability to obtain and use them proficiently. Furthermore, the wordlist may be dominated by function words rather than content words. The word list did not include function words such as proper nouns, pronouns, modal verbs, prepositions, conjunctions, abbreviations, numerals, and non-words. The revised edition of this word list only consisted of content words.

Subsequently, each content word was examined using Laurence Anthony's AntConc 3.5.7 (2018). AntConc is a program mostly used to extract concordance lines from each corpus. Each term that met the frequency requirements in the preliminary list was entered into AntConc to analyze its use. Certain content words may also operate as function words. For instance, the word "like" may function as both a verb and a preposition. Given that the preceding approach requires the exclusion of all function words, it is imperative that the frequency of "like" as a preposition is not included in the wordlist.

The comprehensive version of the government official examination wordlist was completed after the manual removal of irrelevant words. The vocabulary in this wordlist was organized from greatest to lowest frequency. All remaining words in the AntWordProfiler were selected for inclusion in the list. Consequently, throughout the arrangement process, it was advisable to provide the most often occurring words first, as this will enable students to identify which terms are prevalent in government official tests. Word members were shown with their frequency to enable students to access the word items included in the examinations. The final edition of the Government Official Examination Word List (GOEWL) had been finished with these concluding steps. Upon the completion of the Government Official Examination Word List (GOEWL), the word families within the GOEWL were aligned with other renowned word lists to assess the difficulty of each word family. The GOEWL data were examined to determine the number of word families associated with each renowned wordlist and to assess the similarities and differences between the GOEWL and other research investigations.

Results

The Government Official Examination Word List (GOEWL) was entirely developed by compiling terms from the English subject examinations. The GOEWL contained 3,085 word families. Every term in the GOEWL was correlated with three significant word lists: the initial 1,000 high-frequency terms from the General Service List (GSL), the subsequent 1,000 high-frequency words from GSL, and the Academic Word List compiled by Coxhead. Within the 3,085 word families in GOEWL, 895 are part of the first 1,000 frequency words in GSL, 531 belong to the second 1,000 words of GSL, and the remaining 270 word families are categorized in the AWL. Ultimately, there exist 1,389 word families that are not included in any renowned wordlist. Excluding function terms, the examples of word families included in each of the renowned word lists are presented in Table 2.

Table 2

The Example of Word Families Covered in Famous Wordlists

The 1st 1,000 GSL	The 2nd 1,000 GSL	AWL	Not belong to previous wordlists
people	passage	job	professor
students	sorry	volunteers	underlined
water	thank	select	apartment
time	clothes	areas	soccer
other	weather	process	printer
take	tomorrow	stress	drug
new	correct	items	discount

After removing function words, the ten most frequent content words in GOEWL were: work, like, people, student, desire, best, make, good, know, other, and

use. All ten of these terms were part of the first 1,000 high-frequency words in the GSL. These terms were fundamental and frequently employed in various settings. The analysis of the word lists indicated that the English examination for government officials employed a significant number of high-frequency words. The government official examination comprises multiple sections, including conversation, vocabulary, cloze test, and reading comprehension, where these words may commonly appear in both questions and answer choices. Moreover, the utilization of these terms was justifiable and pertinent to the aims of the governmental official examinations, which seeks to evaluate students' comprehensive English ability. The 270 word families associated with the AWL suggest that governmental official examinations predominantly utilize numerous specialized terms or vocabulary relevant to undergraduates. Table 3 showed the overlapping words in comparison with other word lists.

Table 3

The Overlapping Words in GOE Word List with Other Word Lists

Word List	GOE Word Families	Percentage
The 1st 1,000 GSL	895	83.38
The 2nd 1,000 GSL	531	6.22
AWL	270	2.41
Not in lists	1389	7.99
Total	3085	100

Table 3 indicates that the coverage of both the 1st 1,000 and 2nd 1,000 GSL in the Government Official Examination Word List (GOEWL) was 89.60 percent, however the Academic Word List in OWL is covered at merely 2.41 percent. The research indicates that both the Government Official Examination Word List

(GOEWL) and the General Service List (GSL) were valuable resources for students preparing for their national examinations.

Discussion

The final results of the Government Official Examination Word List (GOEWL) study indicated both parallels and differences when compared to prior research findings in this corpus-based analysis. Cherngchawano and Jaturapitakkul (2014) reported that the General Service List (GSL) covered around 85 percent of the vocabulary utilized in the Thai National Examination, including O-NET, A-NET, GAT, and BGAT, but the Academic Word List (AWL) comprised merely 4.5 percent of the vocabulary questions in these assessments. The statistical findings from the study by Cherngchawano and Jaturapitakkul closely resembled the data from the Government Official Examination Word List (GOEWL). As much as 89.60 percent of the vocabulary of GOEWL is derived from the first and second 1,000 words of the GSL. Merely 2.41 percent of the GOEWL lexicon was comprised of AWL terms. This data revealed two main points. The initial point was to verify that high-frequency words significantly contribute to the government official examination in the English topic. The vocabulary intended for undergraduates was infrequently utilized in the same assessments, as the objective of the Academic Word List (AWL) is to establish an English foundation for university students (Coxhead, 2000). The second point from this data was to endorse both GSL and GOEWL as suitable word lists for the national tests in Thailand, as GSL covered up to 89.60 percent of the test vocabulary, and GOEWL was developed to align with the examination objectives. The results indicated the distinctions between GOEWL and other prior word lists. Certain word lists contain a higher number of AWL words compared to their

representation in GOEWL. The Academic Word List (AWL) shown a greater representation in the journal of literature and language teaching compiled by Chanasattru and Tangkiengsirisin (2017). SSWL comprised 394 frequency headwords. Specifically, 127 of the 394 items were classified as belonging to the AWL, or 32.23 percent. Unlike academic wordlists designed for discipline-specific vocabulary (e.g., Khany & Kalantari, 2021), GOEWL captures high-frequency words crucial for governmental exam success. The lower AWL representation suggests that government exams assess general English proficiency rather than specialized academic knowledge.

The data suggests that these three wordlists contained a higher proportion of terms categorized as AWL than the Government Official Examination Word List. The Academic Word List vocabulary included only 4.30 percent of the official government lexicon. The disparity in word count stemmed from the specific intended purpose. The Government Official Examination Word List was created for undergraduates to prepare for governmental examinations. Thus, the words in GOEWL denote authentic high-frequency vocabulary that may be regularly utilized across different sectors of the age-group examination context. Nonetheless, the alternative wordlists aimed to provide essential vocabulary understanding within each specified context as evidenced by Khany and Kalantari (2021) which identified 658 accounting academic word families constituted 10.16% of the whole corpus. Subsequent investigation revealed that of the identified high-frequency word families, only 354 corresponded with those enumerated in the Academic Word List (AWL). Thus, the differing amounts of AWL words in each list are beneficial for incorporation into classroom instruction or material development for students preparing for government examinations, enabling them to enhance their English

vocabulary repertoire for higher education or any governmental English examinations.

Recommendations

To enhance validity, more tests could be included in the data collection. A longitudinal study may be conducted to compare examinations on a regular basis because Thailand has had standard official government exams for more than 20 years. On the basis of existing corpora, numerous scholars have attempted to assemble and produce new word lists. For instance, a New General Service List (NGSL) of essential vocabulary for second language learners was created by Browne et al. (2014). The NGSL is a significant revision of Michael West's 1953 GSL and contains the most significant high frequency words in the English language for second language learners (Browne, 2014). Gardner and Davies (2014) presented the New Academic Vocabulary List (NAVL), which is the other list of high frequency words. The 120 million-word academic sub-corpus of the 425 million-word Corpus of Contemporary American English (COCA) serves as the basis for the NAVL. They asserted that the NAVL employed a greater quantity of texts and broader coverage than the AWL.

It recommended that other researchers utilize the NGSL and the NAVL (2014), both potential world lists, as a foundation for their future study projects in order to examine the government tests in Thailand. The future research should compare GOEWL vocabulary with CEFR-based wordlists to evaluate whether government exams align with international proficiency standards. Comparing university students' vocabulary knowledge against GOEWL could also reveal

whether academic instruction sufficiently prepares students for national-level assessments.

References

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2022). What can corpus software do? In *The Routledge handbook of Corpus linguistics* (pp. 103-125). Routledge.

Apichat, B., & Fatimah, N. (2022). Students' difficulties in learning English speaking: A case study in a Muslim high school in the South of Thailand. *Teaching English as a Foreign Language Journal*, *1*(1), 13-22.

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction, 3*(1), 1-10. https://doi.org/10.7820/vli.v03.1.browne

Browne, C., Culligan, B., & Phillips, J. (2013). *The new academic word list*. http://www.newgeneralservicelist.org/nawl-new-academic-word-list/

Chanasattru, S., & Tangkiengsirisin, S. (2017). The word list distribution in social science research articles. *Arab World English Journal, 8*(4), 412-429. https://dx.doi.org/10.24093/awej/vol8no4.28

Cherngchawano, W., & Jaturapitakkul, N. (2014). Lexical profiles of Thailand university admission tests. *PASAA*, *48*(1), 1-28.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–236. https://doi.org/10.2307/3587951

Coxhead, A. (2018). Vocabulary and English for specific purposes research. Routledge. https://doi.org/10.4324/9781315146478

Crawford, W., & Csomay, E. (2015). *Doing corpus linguistics*. Routledge.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. http://applij.oxfordjournals.org/content/early/2013/08/02/applin.amt015

Khany, R., & Kalantari, B. (2021). Accounting academic word list (AAWL): A corpus-based study. *Journal of Foreign Language Teaching and Translation Studies*, *6*(1), 35-58.

Kongcharoen, P. A. (2023). Investigation of vocabulary for ESP classrooms from academic journals in physical education and sport Science. *Language*, *16*(2), 311-332.

Nation, P., & Webb, S. (2011). Researching and analyzing vocabulary. Heinle Cengage Learning.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University press.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *Journal of English for Specific Purposes, 28*(3), 170 – 182.

West, M. (1953). *A general service list of English words*. Green and Co.

Worarattapong, A., & Phoocharoensil, S. (2023). Who blames: Police or protesters?: A Corpus-based Study of ideological bias in anti-government protest news. *JHUSOC, 21*(2), 151-172. [in Thai]

Wudthayagorn, J. (2022). An exploration of the English exit examination policy in Thai public universities. *Language Assessment Quarterly*, *19*(2), 107-123.

Xodabande, I., & Xodabande, N. (2020). Academic vocabulary in psychology research articles: A corpus-based study. *MEXTESOL Journal*, *44*(3), 1-21.

Zhang, X., & Sukying, A. (2021). Receptive and productive knowledge of lexical collocations in Thai university learners of English. *European Journal of English Language Teaching*, *6*(6), 266-285.

Author

Dr.Darunee Yotimart

General Education Program, Faculty of Sports Communication, Thailand National Sports University 239 Ongkarn 2 Road, Muang Chaiyaphum District, Chaiyaphum Province 36000

Tel: 083-932-4965 E-mail: ajarnlidatnsu@gmail.com