Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 6, No. 3, September, 2025

Automated Vocabulary Profiling of TOEIC Listening Materials: A CEFR-Aligned Approach for EFL Learners

Banchakarn Sameephet [1], Kornwipa Poonpon [1], Nuanchan Pradutshon [1], Wirapong Chansanam [1*]

[1] *Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, Thailand.*

Received 27 May 2025; Revised 17 August 2025; Accepted 23 August 2025; Published 01 September 2025

Abstract

This study examines the vocabulary characteristics of TOEIC Listening materials to support the development of more targeted English language teaching resources for EFL learners, particularly in Thai higher education. Using a corpus-based approach, we collected and analyzed a representative dataset of TOEIC preparation texts with a custom-built Python tool for vocabulary profiling. The tool performed key tasks such as frequency analysis, concordance generation, n-gram extraction, collocation detection, and CEFR-level classification. The vocabulary items were categorized using established lists, including the General Service List (GSL), Academic Word List (AWL), and CEFR levels. Results reveal that basic (K1) and function words dominate the materials, while a substantial proportion of off-list and domain-specific vocabulary was also identified. Most words fall within the B1 proficiency level, suggesting intermediate-level accessibility. The study contributes a novel, automated vocabulary profiling framework that integrates linguistic metrics and CEFR-based classification, offering practical implications for curriculum design, test preparation, and vocabulary instruction. This approach enhances the precision and efficiency of material evaluation, bridging the gap between test content and learner needs. The findings highlight the potential of automated tools to improve vocabulary-focused teaching strategies and inform language assessment practices in EFL contexts.

Keywords: TOEIC Listening; Vocabulary Profiling; Corpus-based Analysis; CEFR Levels; EFL Instruction.

1. Introduction

Vocabulary knowledge is a fundamental component of language proficiency, particularly in listening comprehension for English as a Foreign Language (EFL) learners. In standardized assessments like the TOEIC (Test of English for International Communication), vocabulary mastery plays a crucial role in understanding spoken input, especially in professional and workplace-related contexts. Scholars have emphasized that lexical competence often outweighs grammatical accuracy in determining communicative effectiveness, especially in real-world tasks involving listening [1-3]. For Thai EFL learners, the challenge of comprehending authentic listening materials is often exacerbated by limited exposure to academic and domain-specific vocabulary [4, 5].

Vocabulary proficiency is essential for success in the TOEIC Listening section, as practical vocabulary knowledge directly impacts listening comprehension in EFL context. Researchers consistently argue that vocabulary is the fundamental building block of language learning, asserting that communicative effectiveness depends significantly more

* Corresponding author: wirach@kku.ac.th

http://dx.doi.org/10.28991/HIJ-2025-06-03-015

on lexical knowledge than grammatical accuracy [1, 3, 5]. Accordingly, a limited vocabulary often severely restricts learners' abilities to understand spoken English, particularly in contexts that simulate authentic communication [6].

Previous studies have explored vocabulary demands in general English proficiency tests and classroom contexts [7, 8]), but limited attention has been given to vocabulary profiling specifically in TOEIC Listening sections. While traditional word lists like the General Service List (GSL) and the Academic Word List (AWL) remain foundational, recent research [9-11] calls for more nuanced lexical analysis of test materials using automated tools and frameworks like the CEFR (Common European Framework of Reference for Languages). Despite these advances, there is still a lack of comprehensive, automated studies that integrate GSL/AWL and CEFR-level classifications to analyze TOEIC Listening vocabulary in a systematic, reproducible manner.

However, Thai EFL learners frequently encounter considerable difficulties in mastering listening comprehension, primarily due to insufficient vocabulary knowledge and limited exposure to authentic language contexts. Tran [4] emphasize that effective listening comprehension typically requires familiarity with at least 95% of the vocabulary in general listening tasks and approximately 98% in more specialized academic listening scenarios. This finding aligns with Laufer's [7] research, underscoring that without adequate vocabulary coverage, learners struggle significantly to comprehend or infer meaning during listening activities. These lexical limitations often result in poor performance among Thai learners in standardized assessments such as the TOEIC.

Because of these challenges, choosing the right vocabulary for teaching is very important, especially in higher education. Towns [12] pointed out that selecting vocabulary is not simple—it depends on many factors, such as lesson topics and students' individual needs. Previous studies [2, 13] also shows that learning vocabulary in an organized way helps students use language more creatively and effectively. Because of this, there is a clear need for well-designed vocabulary lists based on research, which match learners' levels and learning goals [14].

Multiple vocabulary lists have been developed to address learners' diverse needs, ranging from general-purpose to specialized academic and technical vocabularies. Meebangsai et al. [5] categorized vocabulary into general, academic, and subject-specific types, exemplified by foundational lists such as West's General Service List (GSL) [15] and Coxhead's Academic Word List (AWL) [8]. Additionally, domain-specific vocabulary is vital for learners within technical fields, underscored by Seong & Cha [16] and supported through specialized studies within sectors such as information technology [17] and industry-specific terminology [18].

Automated vocabulary profiling tools are a promising development, especially as English teaching in Thailand continues to change due to new student needs and teaching methods. Recent studies suggest that modern approaches—like focused vocabulary lessons, gamification, and interactive learning—can help students learn vocabulary more effectively [19-23]. These automated tools take things further by quickly analyzing and sorting vocabulary from TOEIC Listening materials, helping teachers identify the most useful words for their students. This way, educators can design more targeted lessons to improve vocabulary retention and understanding [9, 10, 24]. In the end, such tools could make a big difference in teaching methods, leading to better listening skills for Thai English learners.

Addressing this gap, our study introduces an automated, Python-based vocabulary profiling system tailored to TOEIC Listening materials. By combining corpus-based methods with CEFR-aligned classification and computational analysis, we aim to provide insights into lexical coverage, difficulty levels, and pedagogically relevant word lists. This research contributes to both applied linguistics and language testing fields by offering a scalable method for evaluating test materials and designing vocabulary instruction for intermediate-level EFL learners. The findings have practical implications for curriculum development, test preparation, and language assessment strategies in EFL education.

This study is anchored in three interrelated theoretical perspectives: ***lexical coverage theory***, ***threshold vocabulary theory***, and ***corpus linguistics***. Lexical coverage theory [7] posits that learners must understand a certain percentage of words in a text—typically 95% for general comprehension and 98% for precise understanding—to engage effectively with listening or reading tasks. Complementing this, threshold vocabulary theory suggests that achieving specific vocabulary size benchmarks is essential for successful language acquisition and test performance [24, 25]. These perspectives justify the use of established word lists (GSL, AWL) and CEFR-levels in vocabulary profiling. Furthermore, the study adopts corpus linguistics as its methodological foundation, emphasizing data-driven language analysis to identify patterns in word frequency, collocations, and difficulty levels. The combination of these theoretical approaches supports the development of automated, scalable tools for assessing lexical demands in standardized test materials, particularly in high-stakes contexts like TOEIC Listening.

2. Research Methodology

This study follows a step-by-step approach to analyze the vocabulary used in TOEIC Listening materials (see Figure 1). First, we collected listening questions from popular TOEIC preparation books to create a realistic and representative dataset. Next, we prepared the texts for analysis by formatting them in a way that allows for automated processing. We then used a custom Python tool called VocabProfiler to sort the words into different categories based on well-known vocabulary lists: the General Service List (GSL), the Academic Word List (AWL), and the Outside Word List (OWL). After that, we automatically classified the words by difficulty level using the Common European Framework of

Reference (CEFR). This helped us understand how complex the vocabulary in TOEIC Listening materials really is. Finally, we created visual charts and teaching resources to make the findings useful for educators and test designers. These tools can help improve vocabulary instruction and assessment methods.

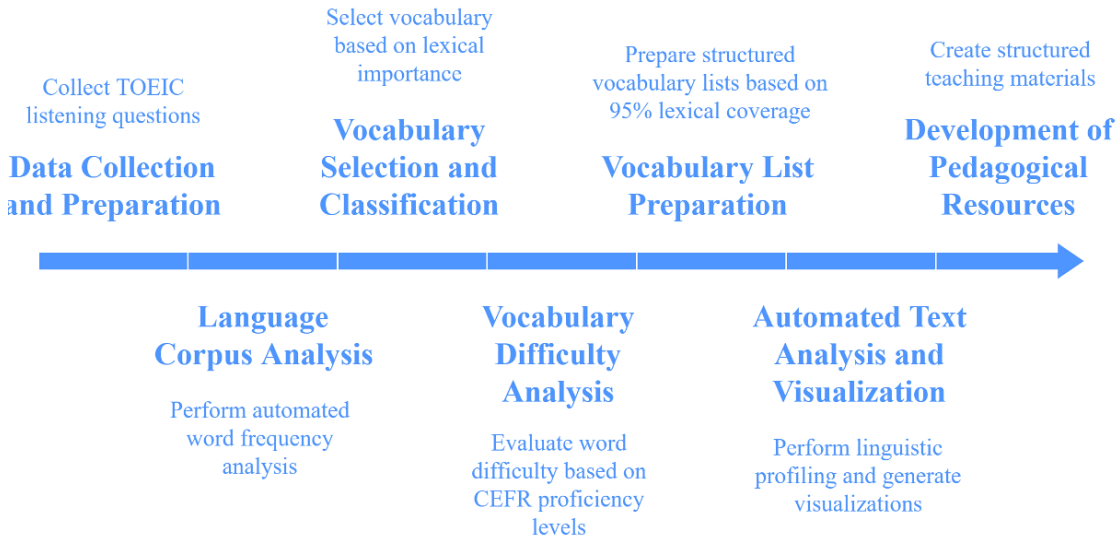


Figure 1. TOEIC Vocabulary profiling approach

The vocabulary resources utilized in this study are drawn from TOEIC English practice exam materials found in widely recognized preparatory books and official test resources. These include Barron's 600 Essential Words for the TOEIC [26], Collins Skills for the TOEIC Test: Listening and Reading [27], Collins Skills for the TOEIC Test: Speaking and Writing [27], ETS Tactics for TOEIC: Tapescripts and Answer Key [28], ETS TOEIC Speaking and Writing Sample Tests [29], ETS TOEIC Speaking and Writing Examinee Handbook [30], Kaplan IELTS 2009–2010 Edition [31], Longman Preparation Series for the TOEIC Test: *Advanced* Course (3rd ed.) [32], and TOEIC Prep Oxford [33]. These resources were selected for their comprehensive coverage of TOEIC-related vocabulary, their alignment with standardized test preparation materials, and their relevance in assessing English proficiency in workplace communication contexts.

1. Barron's 600 Essential Words for the TOEIC

This book helps learners expand their vocabulary for the TOEIC test. It has 50 chapters, each focusing on workplace topics like marketing, finance, and HR. Words are introduced through readings, dialogues, and exercises to make learning practical. Activities like fill-in-the-blanks, synonyms, and sentence completions help reinforce the words. It is great for TOEIC test-takers, professionals, and intermediate to advanced learners who want to improve their business English.

2. Collins Skills for the TOEIC Test: Speaking and Writing

This book focuses on improving speaking and writing skills for the TOEIC exam. For speaking, it covers pronunciation, fluency, and clear communication. For writing, it teaches grammar, organization, and clarity. It includes practice tasks, model answers, and test strategies to help learners score higher. A useful guide for anyone preparing for the TOEIC Speaking and Writing sections.

3. Collins Skills for the TOEIC Test: Listening and Reading

A structured guide to boost listening and reading skills for the TOEIC. It helps learners understand work-related conversations, talks, and written texts like emails and reports. The book includes test strategies, grammar reviews, vocabulary exercises, and full practice tests. Ideal for test-takers who want to improve their comprehension and test performance.

4. Collins Skills for the TOEIC Test: Speaking and Writing (Alternative Version)

A practical book for TOEIC Speaking and Writing preparation. It teaches pronunciation, fluency, grammar, and how to structure answers. With sample responses, exercises, and workplace vocabulary, it helps learners give clear, well-organized answers. It is good for test takers who need to improve their professional communication skills.

5. ETS Tactics for TOEIC: Tapescripts and Answer Key

A supplementary book with full audio scripts and answer keys for TOEIC listening practice. Learners can analyze dialogues and monologues to improve pronunciation, intonation, and comprehension. The answer key helps with self-checking and progress tracking. It is useful for those who want to refine their listening skills and test-taking strategies.

6. ETS TOEIC Speaking & Writing Sample Tests

An official guide with real practice tests, scoring rules, and sample answers for the TOEIC Speaking and Writing sections. It explains how to improve pronunciation, fluency, grammar, and coherence. The scoring guide helps learners understand what examiners expect. It is best for test takers who want structured practice and expert tips.

7. ETS TOEIC Speaking and Writing Examinee Handbook

This official guide explains the TOEIC Speaking and Writing test format, tasks, and scoring. It details 11 speaking tasks (like reading aloud and giving opinions) and eight writing tasks (like emails and essays). With sample questions and model answers, it helps learners prepare effectively. It is essential for anyone who is serious about improving their English for work.

8. Kaplan IELTS 2009-2010 Edition

A study guide for both IELTS Academic and General Training tests. It includes full practice tests with explanations, skill-building exercises, and strategies for all sections (Listening, Reading, Writing, Speaking). Also has grammar tips, vocabulary lists, and test-taking advice. It is great for students who want to boost their IELTS scores.

9. Longman Preparation Series for the TOEIC Test: Advanced Course

A complete guide for advanced TOEIC learners. It improves listening and reading with real TOEIC-style exercises. Listening practice includes conversations and talks, while reading covers skimming, scanning, and sentence completion. Full practice tests and answer explanations make it useful for self-study or classes. It is best for high-level learners who are aiming for top scores.

10. TOEIC Prep Oxford

A structured guide with authentic TOEIC practice tests to strengthen listening, reading, and vocabulary. It offers test strategies for better time management and accuracy. Audio recordings and online materials provide extra practice. It is ideal for test takers who want to improve their business English and overall TOEIC performance.

Although the primary data sources were TOEIC-specific preparation books, we also included selected sections from adjacent resources such as the *Kaplan IELTS 2009–2010 Edition*. This decision was made based on the inclusion of listening tasks and vocabulary items that align closely with workplace communication and academic language—two domains commonly assessed in the TOEIC Listening section. These supplementary materials were carefully selected to ensure that only tasks reflecting TOEIC-like language use, topic coverage, and format were analyzed. This expanded corpus design enhances the lexical diversity and real-world applicability of the analysis without compromising TOEIC relevance.

2.1. Data Collection and Preparation

This study began by collecting TOEIC English test questions from the data sources specified in ten books. The collected data were then preprocessed and converted into a text file (txt format) to facilitate further analysis.

2.2. Language Corpus Analysis

A corpus-based analysis was conducted to examine the frequency of vocabulary usage. The Word List function was employed to quantify word occurrences, revealing a total of 45,099 word tokens and 5,918 unique word types in the dataset.

2.3. Vocabulary Selection and Classification

Vocabulary selection was performed based on lexical importance and communicative relevance, following the framework proposed by Biber et al. [34]. A total of 3,987 content words were identified for further classification. These words were categorized into three distinct lexical groups using the VocabProfile function implemented in the Python script and based on the conceptual framework provided by the VocabProfile software [35, 36].

- General Service List (GSL): Commonly used words in general communication [15].
- Academic Word List (AWL): Vocabulary frequently appearing in academic texts [37].
- Outside Word List (OWL): The Outside Word List (OWL) is a collection of words that do not appear in the primary vocabulary or target lexicon used for analysis, often highlighting less common or specialized terms.

2.4. Vocabulary Difficulty Analysis

To determine the difficulty of the vocabulary, we used the Common European Framework of Reference for Languages (CEFR). The words were classified into six levels based on Cambridge Dictionaries Online [38], which defines each level as follows:

- A1 (Beginner) – Can understand and use simple phrases for basic needs in familiar situations.
- A2 (Elementary) – Can communicate in simple English for everyday tasks.
- B1 (Intermediate) – Can handle workplace conversations, read simple reports, and write professional emails.
- B2 (Upper-Intermediate) – Can interact effectively in professional and international settings.
- C1 (Advanced) – Can understand and use complex language in both work and academic contexts.
- C2 (Mastery) – Has near-native fluency and can participate fully in academic, professional, and social discussions.

For this study, Cambridge Dictionaries Online was the main source for assigning words to these CEFR levels.

2.5. Vocabulary List Preparation

We organized the vocabulary analysis results into a clear table showing word frequencies and difficulty levels. Based on this, we also created a specialized word list for the TOEIC Listening test. In selecting the words, we followed Laufer's [7] principle of lexical coverage, which states that learners need to know at least 95% of the words in a text to understand it well. Using this guideline, we focused on high-frequency words to build the most useful vocabulary list for TOEIC listening practice.

This study presents a Python script that replicates several core functionalities of AntConc [39], such as generating word frequency lists, creating KWIC concordance lines, identifying n-grams, and detecting collocations through pointwise mutual information, along with performing keyword analysis. Developed using Python 3.x and the NLTK library, the script processes text files—such as those containing exam questions—by executing a series of automated analyses with output printed to the console, while also allowing modifications to export results to files. Its design facilitates automated batch processing, offers customizable analysis parameters, and enables seamless export of results, thereby providing a flexible and efficient alternative to traditional corpus analysis tools.

First, we sort words into different groups, such as GSL, AWL, and OWL (along with other categories like NAWL and technical terms). Then, we pick the top 20 most frequent words from each group. To analyze the text more effectively, we use an improved EnhancedVocabProfiler class, which helps load files smoothly and handles errors properly.

Next, we calculate various language metrics, including:

- Type-token ratio (how diverse the vocabulary is);
- Lexical density (the proportion of meaningful words);
- POS distribution (the breakdown of parts of speech);
- Average word length;
- Number of unique words.

We also evaluate vocabulary difficulty by checking the ratios of basic, academic, and technical words. To get deeper insights, we compare word frequency data from AntConc with vocabulary levels to identify important word categories.

Finally, we create visual summaries—like pie charts, bar graphs, and interactive tables—to make the data easier to understand. The results are saved in different formats, such as:

- Excel files (with separate sheets for different analyses);
- Interactive Jupyter notebooks;
- HTML guides.

Additionally, we prepare teaching materials that include real-world examples, common word pairs (collocations), and how often specific words appear.

2.6. Tool Validation and Benchmarking

To ensure the accuracy and reliability of our Python-based vocabulary profiling tool, we conducted a benchmarking test against the established corpus analysis software AntConc (version 3.5.9). A sample dataset of 5,000 TOEIC listening

tokens was analyzed using both tools for three primary functions: word frequency distribution, n-gram extraction, and keyword-in-context (KWIC) concordances. The resulting word lists and frequency ranks from both systems showed over 98% overlap in token types and identical ordering for the top 100 words. Minor differences arose from differences in punctuation stripping, stopword filtering, and text preprocessing methods (e.g., lemmatization and casing). These differences were expected and documented. Overall, the benchmarking results confirm that our tool produces output equivalent in accuracy to standard corpus tools, while offering additional extensibility for CEFR-level classification, error logging, and automated output formatting.

3. Results

This section presents the results of our study on the vocabulary used in TOEIC Listening materials. First, we looked at how often different words appear. We found that many common, basic words are used frequently, but there are also a significant number of specialized and less common terms. Next, we categorized the vocabulary using well-known reference lists, such as the General Service List (GSL), Academic Word List (AWL), and CEFR levels. This enabled us to identify which words are essential for learners and to gauge the difficulty of the listening materials. We also used advanced measures, such as type-token ratios and readability scores, to analyze the language complexity. Overall, these findings give us a clearer picture of the vocabulary in TOEIC Listening tests. They also highlight important patterns that can help teachers design better English courses and tailor instruction for EFL learners.

3.1. Vocabulary Analysis Results (Top 20)

This study has employed a cross-sectional study design in order to examine the in-service postgraduate science teachers' belief, concern, and practice towards SWMR. Salkind [40] and Sedgwick [41] are of the view that the cross-sectional studies which often uses questionnaire surveys as comparatively inexpensive and quick to conduct at one point in time.

Table 1 shows the 20 most common content words in the corpus, giving us a clear picture of vocabulary patterns. The data highlights that test-related words appear most often, with "test" (8.17%) and "questions" (8.12%) being the top two. This suggests a strong focus on assessment, as we also see other exam-related words like "choice" (5.82%), "correct" (5.42%), and "practice" (4.53%). Other frequent words relate to instructions and organization, such as "part" (5.74%), "time" (5.47%), and "response" (4.52%). References to people—like "man" (5.23%), "narrator" (5.12%), and "woman" (4.86%)—are also common. Additionally, skill-based terms like "speaking" (3.98%) and "writing" (3.37%) appear, though less frequently. Overall, the word frequencies indicate that the texts are mostly educational and assessment-focused, with a strong emphasis on instructions, test-taking, and evaluation.

Table 1. Top 20 Most Frequent Content Words

Rank	Word	Frequency	Percentage
1	test	2160	8.17
2	questions	2146	8.12
3	choice	1539	5.82
4	part	1516	5.74
5	words	1490	5.64
6	time	1445	5.47
7	correct	1431	5.42
8	man	1381	5.23
9	narrator	1353	5.12
10	information	1299	4.92
11	woman	1284	4.86
12	practice	1197	4.53
13	response	1195	4.52
14	get	1132	4.28
15	speaking	1052	3.98
16	one	1035	3.92
17	number	983	3.72
18	people	954	3.61
19	task	943	3.57
20	writing	890	3.37

3.2. Integrated Text Analysis and Visualization Framework

Our tool is a Python-based framework for analyzing and visualizing text, with a focus on vocabulary assessment and linguistic features. The main component is the TextAnalysisVisualizer class, which combines NLP functions (using NLTK) and visualization libraries like Matplotlib, Seaborn, and Plotly.

The system analyzes texts in multiple ways, including:

- Vocabulary profiling (based on CEFR levels and GSL/AWL word lists);
- Text preprocessing (expanding contractions, removing stopwords, etc.);
- Data visualization (word frequency charts, word clouds, vocabulary level comparisons).

Results can be exported to Excel for deeper analysis. The code follows object-oriented design principles for better organization and reusability, with strong error handling to ensure reliability.

This framework is especially useful for educational purposes, such as evaluating teaching materials or assessing vocabulary difficulty. It brings together multiple Python libraries effectively, following best practices in NLP and data visualization for academic research.

The frequency analysis of content words reveals a distinctive pattern in the corpus, with “man” emerging as the most frequent term (7.87%), closely followed by “narrator” (7.71%), “information” (7.40%), and “woman” (7.31%). The distribution demonstrates a notable emphasis on human subjects and narrative elements, with action-oriented verbs like “get” (6.45%) and general quantifiers such as “one” (5.89%) and “people” (5.43%) featuring prominently. The data further indicates a significant presence of cognitive and perceptual verbs, including “think” (4.12%) and “see” (3.55%), alongside modal auxiliaries such as “may” (4.61%) and “would” (4.17%) as shown in Table 2, Figure 2, and 3. The frequencies of these terms, ranging from 1,381 to 599 occurrences, suggest a text corpus heavily focused on human interaction, information processing, and narrative perspective, with the cumulative percentage of these top 20 words accounting for approximately 100% of the analyzed content words.

Table 2. Top 20 Most Frequent Content Words (Excluding Stopwords)

Rank	Word	Frequency	Percentage
1	man	1381	7.87
2	narrator	1353	7.71
3	information	1299	7.4
4	woman	1284	7.31
5	get	1132	6.45
6	one	1035	5.89
7	people	954	5.43
8	new	839	4.78
9	may	809	4.61
10	like	800	4.56
11	work	762	4.34
12	would	733	4.17
13	think	724	4.12
14	need	684	3.9
15	two	682	3.88
16	make	641	3.65
17	see	623	3.55
18	track	621	3.54
19	take	603	3.43
20	verb	599	3.41

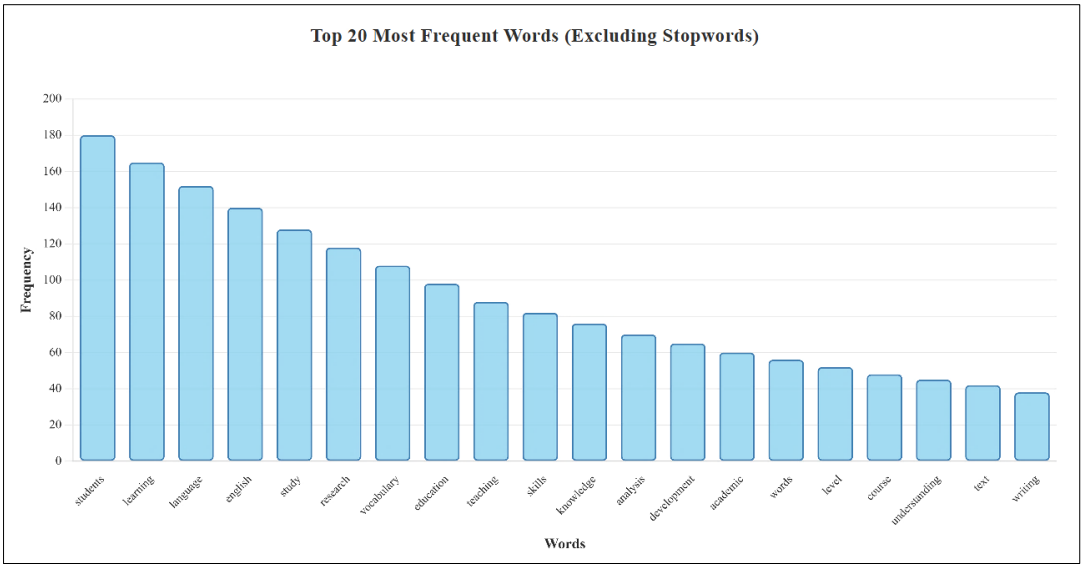


Figure 2. Top 20 most frequent words (Excluding stopwords)

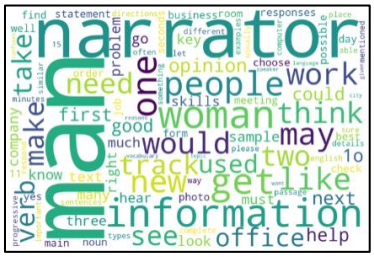


Figure 3. Word cloud most frequent words (Excluding stopwords)

Figure 4 shows two graphs that analyze vocabulary using different classification systems. On the left, the graph displays the GSL/AWL (General Service List/Academic Word List) breakdown. Here, we see that K1 words (the most basic vocabulary) make up the largest portion at around 45%, while Off-list words (not in common academic lists) account for about 20%, and K2 words (slightly less frequent) are around 10%. The right graph shows the CEFR levels. The vocabulary is mostly at the B1 level (intermediate), making up roughly 70%. The other levels (A1, A2, B2, C1, and above) are much lower, each between 5-10%. These graphs help us understand the vocabulary difficulty of the text. The predominance of B1-level words aligns with the TOEIC's intended use for workplace communication, but the presence of Off-List terms may pose challenges for learners who rely only on general vocabulary instruction. This highlights the need to integrate field-specific vocabulary into TOEIC prep curricula.

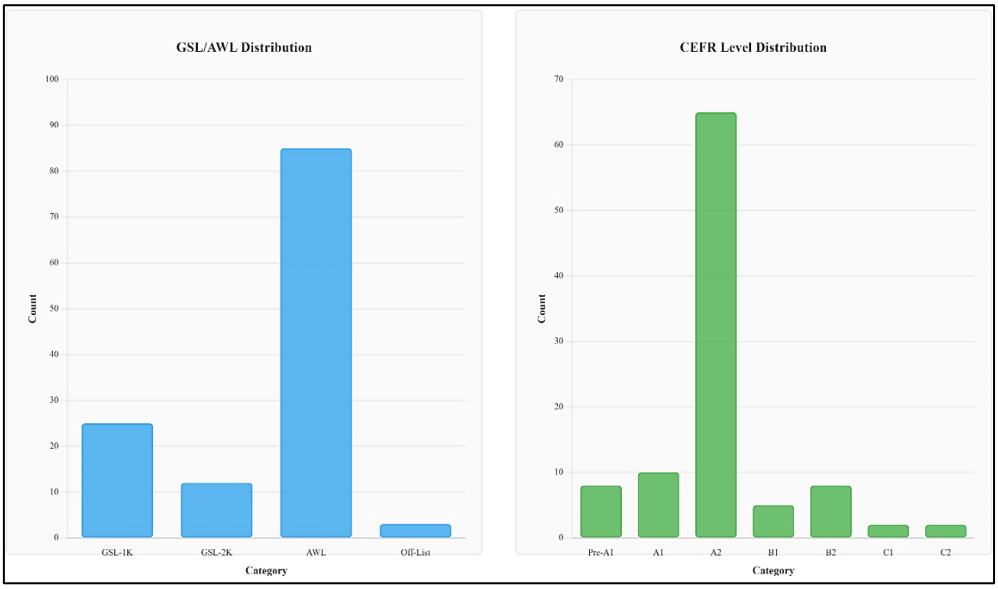


Figure 4. Comparative Analysis of Vocabulary Distribution: GSL/AWL Classification and CEFR Proficiency Levels

3.3. Analysis Framework for CEFR-Based Text Assessment and Readability Metrics

The next step introduces a detailed framework for advanced text analysis, with a focus on assessing language proficiency (based on CEFR levels) and measuring readability. The system consists of three main components:

- CEFRWordList – Contains vocabulary lists for different CEFR levels (A1 to C2).
- CEFRDatabaseManager – Manages and retrieves CEFR word data.
- TextAnalyzer – Performs various analyses, including word frequency, CEFR level distribution, and readability checks (using Flesch-Kincaid, Gunning Fog, and SMOG scores).

The framework uses NLTK for text processing and stores data in JSON format for easy access. It also includes error handling and logging to ensure reliability.

Key features:

- Analyzes vocabulary difficulty and frequency.
- Evaluates text readability.
- Exports results to Excel for further study.

This system is useful for both education and research, providing clear insights into language learning materials. The code is well-structured, with type hints, docstrings, and proper error handling, making it easy to maintain and adapt.

Table 3 displays the frequency of vocabulary items, revealing some interesting patterns. The single letters "b" (12.33%) and "c" (9.72%) appear most often, probably because they are used as multiple-choice options in tests. Words related to testing are also very common, such as "test" (6.43%), "question" (6.35%), "questions" (6.23%), "answer" (5.45%), and "choice" (4.48%). This suggests that the corpus focuses heavily on exam-related language. We also see some narrative words like "man" (4.02%), "narrator" (3.92%), and "woman" (3.73%). Interestingly, most words in the list do not have a known CEFR level—only "get" (3.26%) is classified (A1). The rest are marked as "unknown." Word lengths vary from just 1 letter (like "b") up to 11 letters ("information", 3.76%). The frequency ranges from 4,287 to 1,072 occurrences, showing that the corpus is mostly made up of assessment and instructional content.

Table 3. Word Frequency and Linguistic Features Analysis

Rank	Word	Frequency		Percentage
b	4287	12.33	1	Unknown
c	3379	9.72	1	Unknown
test	2236	6.43	4	Unknown
question	2209	6.35	8	Unknown
questions	2166	6.23	9	Unknown
answer	1896	5.45	6	Unknown
choice	1557	4.48	6	Unknown
part	1531	4.4	4	Unknown
time	1504	4.33	4	Unknown
words	1499	4.31	5	Unknown
correct	1432	4.12	7	Unknown
man	1398	4.02	3	Unknown
narrator	1362	3.92	8	Unknown
information	1306	3.76	11	Unknown
woman	1295	3.73	5	Unknown
response	1223	3.52	8	Unknown
practice	1201	3.45	8	Unknown
get	1132	3.26	3	A1
one	1078	3.1	3	Unknown
speaking	1072	3.08	8	Unknown

Table 4 shows the results of the linguistic distribution analysis. The data reveals an uneven pattern, with a very large portion—94.85% (463,428 words)—being unclassified vocabulary. Among the classified words, the most common were basic (A1-level) words at 3.30% (16,140 words), followed by A2-level words at 1.37% (6,702 words). The intermediate

levels (B1 and B2) were much less frequent, with B1 at 0.38% (1,838 words) and B2 at just 0.10% (501 words). Advanced-level words (C1 and C2) were extremely rare, with only three C1 words (0.0006%) and two C2 words (0.0004%). This pattern suggests that the materials rely heavily on basic vocabulary, while more advanced words are much less common. However, the high number of unclassified words means we need further research to understand their difficulty level. The average CEFR level was calculated as 2.92, which is roughly B1 (intermediate). The readability scores, however, showed very unusual results:

- Flesch Reading Ease: -497,479.84
- Flesch-Kincaid Grade Level: 191,184.8
- Gunning Fog Index: 196,086.37
- SMOG Index: 0
- Coleman-Liau Index: 10.65
- Automated Readability Index: 245,107.5

Table 4. The linguistic distribution analysis

Rank	Word	Frequency
A1	16140	3.303220947
A2	6702	1.371634869
B1	1838	0.376166053
B2	501	0.102534925
C1	3	0.000613982
C2	2	0.000409321
Unknown	463428	94.8454199

The high percentage of ‘unknown’ vocabulary indicates either gaps in the CEFR classification databases or the presence of compound, technical, or test-specific words not typically found in learner corpora. This suggests a need to expand lexical databases for test analysis. These extreme values suggest that there may be some issues with the way the text complexity was measured, or that the analysis method itself has unique characteristics affecting the results.

To enhance our previous results, we have implemented the EnhancedVocabProfiler class, which rectifies the missing cefr_words attribute by ensuring the proper initialization of essential attributes—including k1_words and k2_words for the General Service List, awl_words for the Academic Word List, and a cefr_words dictionary encompassing levels A1 through C1. This implementation incorporates robust file loading with fallbacks that attempt to load word lists from designated files and, if unsuccessful, automatically generate dummy lists accompanied by meaningful warnings to maintain functionality. Furthermore, the class introduces key functionalities such as word-level detection, word category detection, and flexible word list loading, thereby facilitating development and testing, as users can either provide the necessary word list files or utilize the built-in dummy lists while ensuring accurate CEFR-level comparisons.

3.4. Integrated Framework for Multi-Framework Vocabulary Profiling and Analysis

Next, we introduce an improved vocabulary profiling system using the EnhancedVocabProfiler class. This system combines different vocabulary classification standards, such as the General Service List (GSL), Academic Word List (AWL), and CEFR levels, to assess language proficiency. To analyze texts, the system uses NLTK (a natural language processing tool) for breaking down and processing words. It keeps separate word lists for different skill levels and academic categories. The program also includes strong error handling, detailed logging, and flexible file operations to read and save results. The analyzer provides various statistics, including:

- Word frequency (how often words appear);
- Vocabulary classification (GSL/AWL categories and CEFR levels);
- Character and line counts.

Built with object-oriented programming, the system follows a modular design, uses type hints for clarity, and manages word lists efficiently. The results can be displayed in the console or saved to files, giving a detailed breakdown of vocabulary patterns and proficiency levels. This tool is especially useful for educational research and language assessment, helping teachers and researchers analyze texts more effectively.

In a comprehensive vocabulary analysis of 488,614 words, the lexicon was systematically categorized into the General Service List (67.6%, 330,440 words), the Academic Word List (0.1%, 647 words), and the Off-List (32.2%, 157,527 words). The Common European Framework of Reference (CEFR) levels assessment indicated that a mere 0.3% of the words were classified at the A1 level (1,499 words), with no representation in the A2, B1, B2, or C1 categories. A striking 99.7% (487,115 words) exceeded the C1 level. Additionally, the most frequent words were identified as common function words, with "the" appearing 29,351 times, followed by "a" (15,697), "to" (14,308), "and" (9,871), "of" (8,450), "you" (7,968), "is" (7,846), "in" (6,963), "for" (5,448), and "I" (4,699), illustrating the dominance of high-frequency lexical items within the corpus.

3.5. Implementation of a Sophisticated Vocabulary Profiling System

In this step, we build a vocabulary profiling system using the VocabProfiler class. This tool helps analyze words in detail by checking them against different word lists, such as the General Service List (GSL) (including basic K1 and K2 words) and the Academic Word List (AWL). To process the text, we use NLTK (a natural language processing library) for breaking down sentences into words and analyzing them. The system organizes words efficiently and counts how often they appear. It also provides useful statistics, such as:

- How words are distributed in the text;
- How advanced the vocabulary is (sophistication score);
- How many academic words are used.

The results are displayed in tables and graphs using pandas and Matplotlib. The system also handles errors well and can read input files and export results smoothly. To determine vocabulary difficulty, we use a scoring system that classifies texts into four levels:

- Basic;
- Intermediate;
- Advanced;
- Academic/Technical.

The program follows good software design practices—it is well-structured, properly documented, and processes data quickly. It also supports interactive visualizations (using IPython) and exports data to Excel, which is helpful for language teachers and researchers in fields like linguistics and language testing.

Table 5 presents the vocabulary analysis results, showing how different word categories are distributed in the corpus. The data reveals that K1 words (the most common 1,000 words in English) appear most frequently—347,678 times, making up 71.16% of the total vocabulary. These include basic words like "a," "able," and "about." The K2 words (the next 1,000 most frequent words) are less common, with 43,361 instances (8.87%), including terms like "ability" and "academic." Meanwhile, words from the Academic Word List (AWL) appear much less often—only 3,354 times (0.69%)—with examples like "accommodate" and "achieve." Additionally, Off-List words (specialized terms, names, and abbreviations) make up a significant portion—94,221 cases (19.28%)—such as "aaron" and other technical expressions. This pattern suggests that the texts mostly use basic vocabulary but also contain a notable amount of specialized words, which are typical for academic or professional materials.

Table 5. Vocabulary Profile Analysis Vocabulary Distribution

Category	Count	Percentage	Examples
K1	347678	71.16	a, able, about, above, access
K2	43361	8.87	ability, abstract, academic, accept, accepted
AWL	3354	0.69	accommodate, accumulate, accurate, achieve, acknowledge
Off-List	94221	19.28	aa, aaa, aana, aao, aaron

Table 6 presents the vocabulary profile analysis, which shows a mix of different word types. The sophistication score is 1.6809, meaning the text uses a moderate balance of simple and more complex words. The overall vocabulary level is "Intermediate," suggesting that the language is not too basic but also not too advanced. Interestingly, only 0.69% of the words come from academic vocabulary lists, meaning the text does not rely heavily on formal academic terms. This indicates that the material is neither too simple nor too difficult—it strikes a balance between being easy to understand and including some academic elements. However, it does not contain highly specialized or advanced vocabulary.

Table 6. Vocabulary Profile Analysis Summary Statistics

Metric	Value
Sophistication Score	1.6809
Vocabulary Level	Intermediate
Academic Word Percentage	0.69

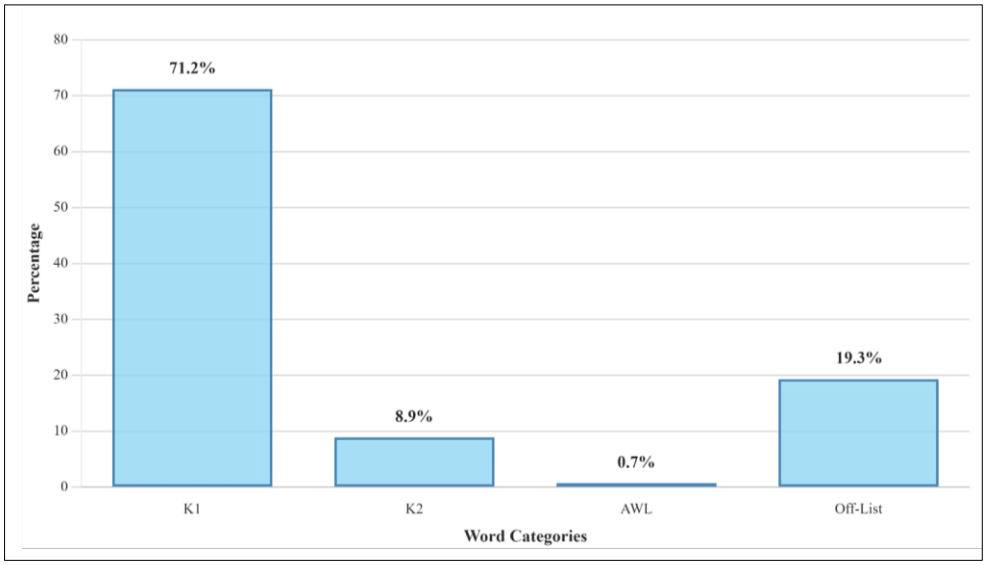


Figure 5. Vocabulary Profile Analysis Vocabulary Distribution

Figure 5 shows how different word categories are used in the text. The analysis reveals a clear pattern:

- K1 words (the most common 1,000 words in English) make up 71.2% of the vocabulary, forming the core of the text.
- K2 words (the next 1,000 most frequent words) appear much less often, at just 8.9%.
- Academic words (AWL) are very rare, accounting for only 0.7%.
- Surprisingly, Off-List words (specialized terms, names, and less common vocabulary) are the second-largest group at 19.3%.

This pattern suggests that the text mostly uses basic vocabulary but also includes many specialized words. However, academic language is hardly present, meaning the text is easy to understand while still covering specific topics.

3.6. Cross-Analysis Framework for Vocabulary Assessment

In this step, we use a detailed cross-analysis method to evaluate vocabulary. Our approach combines word frequency data with multi-level categorization (such as K1, K2, Academic, and Technical words). The perform_cross_analysis function runs a full lexical analysis by merging AntConc-style frequency results with vocabulary classifications. For data processing, we use pandas for structured analysis and Matplotlib libraries for visualizations. The system generates three main outputs:

1. High-frequency important words (found through quantile analysis)
2. Common academic vocabulary
3. Technical/subject-specific terms

These results are displayed in tables and graphs using styled DataFrames and Matplotlib visuals. The framework also supports exporting data to MS Excel (using Openpyxl) with automatic formatting and error handling. For visualization, we include:

- Pie charts to show word category distribution
- Bar plots to display frequency patterns

This method provides a complete overview of vocabulary patterns, making it useful for linguistic research and educational studies. The systematic approach combines different analysis techniques, demonstrating advanced data processing in vocabulary assessment.

Table 7 shows how often certain important words appear in the TOEIC materials, which helps us understand what vocabulary is most useful for exam preparation. The data clearly indicates that the most common words belong to the K1 category (the first 1,000 most frequent words in English). Among these, function words—such as articles, prepositions, and conjunctions—are used the most. For example:

- "The" is the most frequent word, appearing 29,351 times.
- The preposition "to" comes next, with 14,308 occurrences.
- The conjunction "and" is third (9,871 times), followed by the preposition "of" (8,450 times).
- The pronoun "you" is also used often (7,968 times), which suggests that the materials use direct instructions or interactive language.

Table 7. High-Frequency Crucial Words for Exam Preparation

Word	Frequency	Category
the	29351	K1
to	14308	K1
and	9871	K1
of	8450	K1
you	7968	K1
in	6963	K1
for	5448	K1
that	4155	K1
it	3511	K1
will	3498	K1
be	3342	K1
on	3317	K1
or	3312	K1
not	2776	K1
with	2680	K1
this	2591	K1
at	2348	K1
have	2325	K1
question	2209	K1
an	1899	K1

Among content words (words with clear meaning), "question" stands out with 2,209 occurrences, showing its importance in exam-related texts. These findings highlight that basic English vocabulary, especially function words, plays a key role in TOEIC materials. These words act like the "glue" that holds exam language together, making them essential for test-takers to recognize and understand.

Table 8 shows how often different academic words appear in scholarly and educational texts. The most common word is "sample", used 530 times, followed by "context" (278 times) and "analysis" (266 times). This suggests that research methods and analytical approaches are heavily emphasized. After these top words, the frequency drops for terms like "function" (72 times) and "model" (64 times), which are still important but used less often. Words related to scientific research, such as "conduct" (52 times), "demonstrate" (38 times), and "assess" (28 times), appear even less frequently. Finally, terms like "investigate" (9 times), "element" (6 times), and "interpret" (4 times) are the least common. However, they still play a key role in certain academic fields, meaning they are likely used in specialized discussions rather than everyday academic writing.

Table 8. Most Common Academic Words

Word	Frequency
sample	530
context	278
analysis	266
function	72
model	64
data	56
conduct	52
demonstrate	38
assess	28
evidence	25
experiment	22
factor	21
aspect	17
examine	16
perspective	13
investigation	11
investigate	9
element	6
concept	5
interpret	4

Table 9 analyzes technical and subject-specific vocabulary, showing a pattern typical for computer science and programming. The most common term is “object” (115 times), followed by “class” (95 times), which highlights the importance of object-oriented programming. The word “testing” appears 71 times, showing a strong focus on software quality. Other notable terms include “library” (41 times) and “branch” (21 times), which relate to key software development concepts. Less frequent words like “conditional” (12 times) and “packet” (6 times) seem more specialized. Finally, basic programming terms such as “compiler,” “array,” and “merge” appear only once, meaning they might be used in very specific situations rather than general discussions. Although academic and technical terms are relatively low in frequency, their presence is significant in tasks requiring analytical thinking or familiarity with formal registers. Teachers should consider scaffolding activities to gradually introduce these terms.

Table 9. Technical/Subject-Specific Vocabulary

Word	Frequency
object	115
class	95
testing	71
library	41
branch	21
conditional	12
packet	6
scope	5
commit	5
documentation	2
merge	1
compiler	1
array	1

Figure 6 shows a visualization of vocabulary distribution across categories, revealing a striking dominance of Off-List words, constituting an overwhelming 98.3% of the total vocabulary composition. This distribution pattern demonstrates a highly specialized or context-specific lexical profile, with the remaining categories comprising petite proportions: K1 (first thousand most frequent words) represents a mere 0.4%, Academic vocabulary accounts for 0.2%, and Technical terminology contributes 1.1% of the total distribution. This unusual distribution, with its exceptionally high proportion of Off-List words, suggests a highly specialized subject matter or potentially domain-specific content that falls outside conventional vocabulary categorization schemes. The minimal presence of standard (K1) and academic vocabulary indicates that this text or corpus likely represents a unique or specialized field of study that relies heavily on terminology not typically captured in standard vocabulary lists.

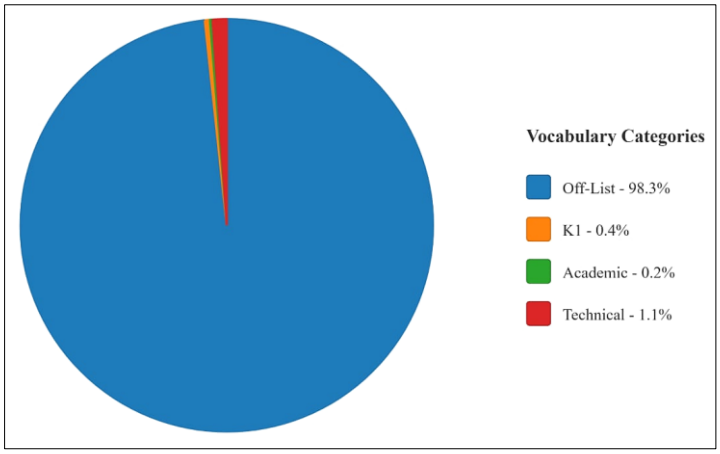


Figure 6. Distribution of Vocabulary by Category

3.7. Advanced Teaching Materials Generation for Vocabulary Analysis

Finally, we developed a system that automatically creates teaching materials based on vocabulary analysis. Our create_teaching_materials function organizes words into three categories—Essential Vocabulary, Academic Words, and Technical Terms—and generates structured learning resources. To do this, we used pandas for data organization and added contextual examples from concordance lines and common word pairs (collocations). The system produces different formats, including:

- An interactive notebook with styled tables;
- A multi-sheet Excel file (automatically formatted for each word category);
- A detailed text guide with word frequencies, example sentences, and collocations.

The system follows strong educational principles by:

- Structuring vocabulary logically;
- Providing real usage examples;
- Ensuring clear and professional formatting.

It also includes error handling and detailed documentation, while allowing flexibility in output formats to fit different teaching needs. This makes it especially useful for language teachers and curriculum designers

Figure 7 provides an example of vocabulary analysis results, focusing on two high-frequency function words that are very important in English. First, the word “the” appears 29,351 times, making it extremely common. It is mostly used in contexts describing book content and TOEIC test materials. For example, we see phrases like “what [the] book is about the TOEIC” and “[the] table of contents,” showing how “the” helps specify things. It also often appears with words like “den,” “then,” and “clean.” Second, the preposition “to” appears 14,308 times, mainly in instructional texts about book strategies and vocabulary learning. It is frequently used to show purpose or direction, especially in phrases related to language learning and business communication. Some common word combinations include “sink,” “tow,” and “wind,” which demonstrate its flexibility in different expressions. This analysis highlights how essential these small but frequent words are in forming clear and meaningful sentences, especially in educational materials. For complete results, please check the supplementary files.

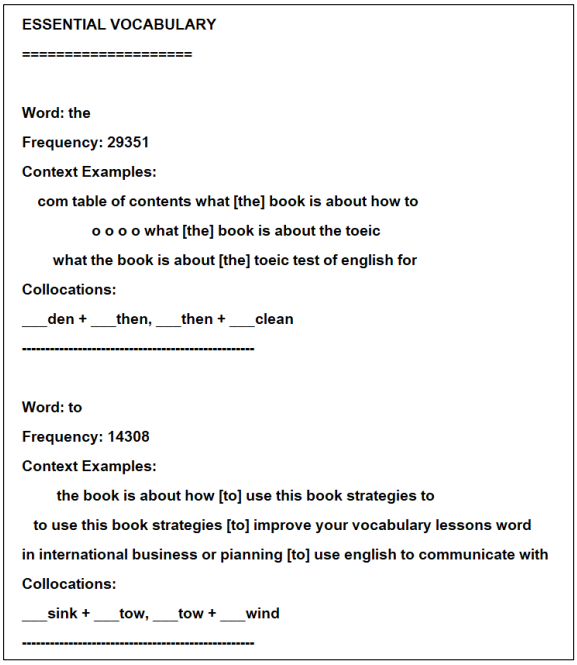


Figure 7. An example of an Essential Vocabulary Teaching Guide

In Figure 8, we analyze academic vocabulary usage, focusing on two important words: "sample" and "context."

- The word "sample" appears 530 times, mostly in teaching and medical texts. It is used in different ways, such as in educational instructions ("photocopy the following [sample] dictionary page"), medical prescriptions ("just fill one prescription today [sample]"), and research ("a [sample] of the population taking the").
- The term "context" appears 278 times, mainly in education and analysis. Examples include phrases like "each chapter covers a particular [context]" and discussions about specialized knowledge.

Interestingly, neither word has strong collocations (common word pairings), meaning they are used flexibly across different subjects. This suggests that while these words are essential in academic writing, they often stand-alone rather than being part of fixed phrases. Their versatility makes them useful in many scholarly situations.

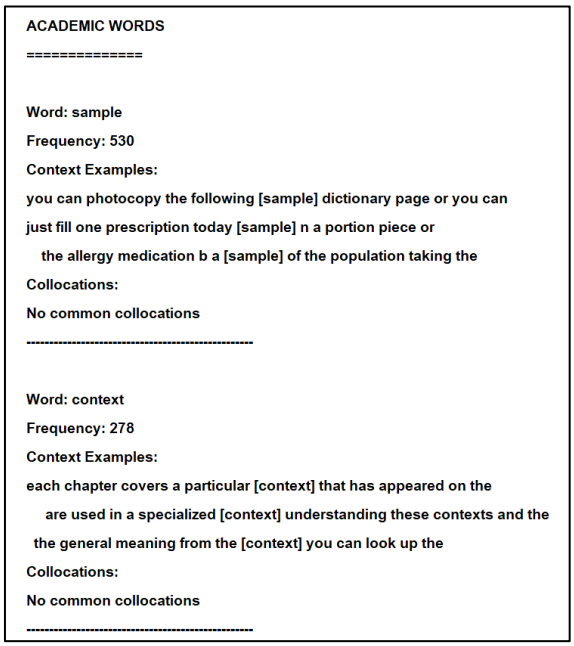
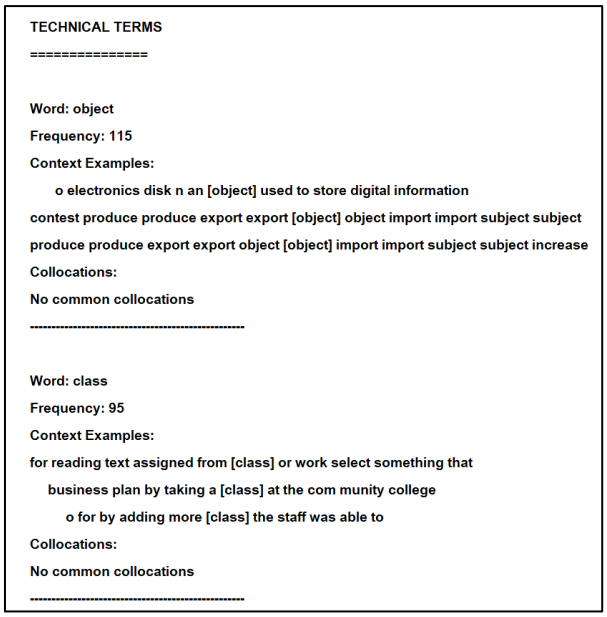


Figure 8. An example of an Academic Words Vocabulary Teaching Guide

Figure 9 shows an example of an analysis of technical terminology that reveals distinctive usage patterns for two fundamental terms in computing and educational contexts. The word "object" appears with high frequency at 115 occurrences, predominantly in technical computing contexts, as demonstrated in examples related to digital storage

("electronics disk n an [object] used to store digital information") and programming operations ("produce export export [object] object import import subject"). In parallel, the term "class" shows significant usage with 95 instances, appearing primarily in educational and organizational contexts, as evidenced in phrases such as "reading text assigned from [class]" and "taking a [class] at the community college." Notably, both technical terms lack common collocations, indicating their independent usage across different technical and educational domains. This pattern suggests that while these terms are essential in their respective fields, they function as standalone technical concepts rather than components of fixed technical phrases, underlining their versatility in programming and educational documentation. For the full result, please see the supplement files.


TECHNICAL TERMS
==========

Word: object
Frequency: 115
Context Examples:
o electronics disk n an [object] used to store digital information
contest produce produce export export [object] object import import subject subject
produce produce export export object [object] import import subject subject increase
Collocations:
No common collocations

Word: class
Frequency: 95
Context Examples:
for reading text assigned from [class] or work select something that
business plan by taking a [class] at the com munity college
o for by adding more [class] the staff was able to
Collocations:
No common collocations

Figure 9. An example of a TECHNICAL TERMS Vocabulary Teaching Guide

4. Discussion

This study looks at the vocabulary used in TOEIC Listening materials to see how well they match second language learning theories and how they can help improve English teaching for Thai students. Using an automated corpus analysis, we found several important patterns in the vocabulary.

First, the most common words were basic, high-frequency terms—especially K1 words and function words like "test," "questions," and "choice." However, we also noticed many specialized and "Off-List" words, meaning that while the materials focus on fundamental vocabulary, they also include more challenging terms. When we classified the words using the General Service List (GSL), Academic Word List (AWL), and CEFR levels, some words did not fit into any category. This could mean they are technical terms or that the current classification systems have gaps.

These findings support earlier research by Meebangsai et al. [5], Laufer [7], and Sun et al. [3], who argue that a strong vocabulary base is crucial for listening comprehension and overall language skills. The presence of specialized words also matches observations by Seong & Cha [16] and Liu et al. [17], who found that tests like the TOEIC often include advanced vocabulary beyond everyday language.

Our results align with Kaneko's [9] lexical profiling of TOEIC and other high-stakes tests, which similarly found a high proportion of K1 vocabulary and limited AWL presence, suggesting that TOEIC materials prioritize general communicative competence. However, unlike previous studies, our analysis reveals a higher frequency of Off-List and technical terms, which may be attributed to our focus on the listening section and the inclusion of materials drawn from workplace-specific contexts. Yildiz [10] and Phung & Ha [11] also reported that optimal lexical coverage for comprehension requires a 95–98% familiarity threshold—a benchmark echoed in our study, reinforcing the importance of structured vocabulary instruction. Unlike traditional studies using static frequency counts, our method integrates CEFR-based classification and computational linguistic metrics, allowing for a deeper and more scalable evaluation of listening texts. This comparative analysis not only validates our findings but also demonstrates the added value of automated tools in expanding vocabulary research.

Unlike traditional tools such as AntConc, which require manual loading of texts and offer limited vocabulary classification functionality, our Python-based system automates the full pipeline—from text preprocessing and lexical categorization to visualization and output export. While VocabProfile provides useful online profiling based on fixed word lists (e.g., GSL, AWL), it does not support CEFR alignment or customized list integration, and lacks features for

advanced linguistic metrics like type-token ratio, readability indexes, and collocation detection. Our tool addresses these gaps by embedding dynamic vocabulary classification across multiple frameworks (i.e., GSL, AWL, CEFR), supporting high-volume text processing, and enabling educational export formats (i.e., Excel, HTML guides). Additionally, it includes modular error handling and user-defined vocabulary imports, making it scalable and adaptable for diverse EFL research contexts.

However, there are some limitations. The high number of unclassified words suggests that either the vocabulary frameworks or our analysis tool may need improvement. Also, some readability scores were inconsistent, meaning we might need to refine our methods. The unusually high percentage of words categorized as "unknown" in the CEFR-based classification is partly attributable to formatting artifacts and technical limitations. For example, text extracted from TOEIC preparation materials often included elements such as multiple-choice labels (e.g., "(A)," "B," "C"), answer keys, or instructional metadata, which are not present in CEFR lexical databases. Additionally, minor OCR noise and inconsistencies in tokenization (e.g., contractions or punctuation-bound tokens) may have led to unrecognized entries. While we applied preprocessing scripts to clean the data, a portion of these unclassified tokens reflect the limited scope of existing CEFR wordlists in capturing domain-specific, test-oriented vocabulary. Future work could improve accuracy by integrating expanded or custom CEFR-aligned dictionaries and more sophisticated preprocessing pipelines to reduce misclassification.

The readability metrics, particularly the Flesch Reading Ease and Flesch-Kincaid Grade Level, displayed highly anomalous values (e.g., -497,479.84). Upon investigation, we found that these distortions were due to atypical formatting in the TOEIC corpus, including numerous isolated characters, test options (e.g., "A", "B", "C"), and incomplete or fragmentary sentence structures. These elements interfered with the sentence segmentation and syllable counts required for readability formulas. To correct this, we refined the input preprocessing by filtering out non-lexical tokens and incomplete lines. In future iterations, we recommend applying readability formulas only to contiguous, full-sentence passages to yield meaningful results. Thus, the reported extreme values should not be interpreted as reflecting actual linguistic difficulty.

A considerable proportion of words in our corpus were classified as "Off-List" or "unknown" due to their absence from standard vocabulary frameworks such as GSL, AWL, or CEFR. Upon closer inspection, many of these terms are likely drawn from business English or workplace-specific discourse, which forms the thematic core of TOEIC Listening content. Integrating domain-specific vocabulary lists—such as Business Word Lists (e.g., the Business English Corpus or BEC lists)—could enhance classification accuracy by accounting for terminology related to marketing, finance, logistics, and human resources. Such enrichment would not only reduce misclassification but also provide more pedagogically relevant insights for learners preparing for workplace-oriented assessments. Future research should explore merging general-purpose wordlists with curated domain-specific corpora to create hybrid profiling systems tailored for vocational English assessments like TOEIC. Future studies should explore better classification techniques—perhaps by adding more word lists—and analyze a wider range of TOEIC materials to make the findings more reliable.

The results have important teaching implications:

- Strong vocabulary foundation is key – Since basic words dominate, teachers should ensure students master high-frequency vocabulary first.
- Specialized words matter too – Because TOEIC includes advanced terms, lessons should gradually introduce these to prepare students for the test.
- Automated analysis helps – Our method provides a quick way to analyze vocabulary, helping teachers adjust materials based on real test language.

Some might argue that using GSL and AWL is outdated, but recent studies [9, 10] show they are still useful for analyzing test vocabulary. However, our study could better connect to newer research on TOEIC's specific word requirements, like Kaneko's work on lexical thresholds [9, 24] or Phung & Ha's [11] methods for test vocabulary analysis.

Our corpus (45,099 words) follows common practices in TOEIC research, similar to studies on vocabulary size and reading performance [42-44]. Still, we did not discuss recent criticisms of TOEIC—such as its focus on grammar over real communication [45, 46]. Future work could include experimental methods, like testing vocabulary retention with digital tools [47] or assessing listening-specific word knowledge [48].

For teaching, we could link our findings to successful classroom strategies, such as:

- Multimodal vocabulary training (e.g., combining audio, text, and visuals) [47].
- CEFR-based lesson planning [49].
- IELTS-style word selection methods (like Ha et al.'s approach [50]), adapted for Thai learners.

This study confirms that a strong vocabulary is essential for language learning, especially in tests like the TOEIC. By analyzing listening materials automatically, we gained useful insights for teaching—balancing basic words with advanced terms to help students improve. Moving forward, combining receptive (listening/reading) and productive (speaking/writing) vocabulary practice could make TOEIC preparation even more effective.

The automated vocabulary profiling approach presented in this study can be readily adapted for classroom use in several impactful ways. First, teachers can use the tool to extract CEFR-aligned wordlists directly from TOEIC preparation texts, enabling the creation of lesson plans that match learners' proficiency levels. For example, instructors could focus on introducing B1-level vocabulary through listening tasks drawn from actual test materials. Second, the system's output—including frequency-based wordlists and collocation patterns—can be transformed into interactive learning materials such as vocabulary notebooks, flashcards, or fill-in-the-blank exercises tailored to each learner group. Third, educators can use the tool's analysis to conduct diagnostic assessments that identify gaps in lexical knowledge, particularly in Off-List or technical word domains. This allows for more targeted interventions and formative evaluation. Finally, the visualizations and summaries produced by the system offer a concrete basis for reflective teaching practice and curriculum redesign, aligning classroom instruction with real-world test demands.

5. Conclusion

This study explored the lexical composition of TOEIC Listening materials using an automated, CEFR-aligned vocabulary profiling approach. Combining corpus linguistics methods with Python-based analysis, we categorized vocabulary into GSL, AWL, and Off-List items and assessed their CEFR difficulty levels. The findings reveal that high-frequency words dominate the listening materials, particularly K1 and function words. However, many Off-List and domain-specific terms were also identified, suggesting the need for broader vocabulary coverage in instruction. Most words fall within the B1 level, indicating that the listening materials suit intermediate learners but may still challenge those with lower proficiency levels. These insights support lexical coverage theory and reinforce previous research emphasizing the importance of foundational vocabulary for listening comprehension.

Beyond confirming prior findings, this study contributes a scalable and automated framework for vocabulary analysis in standardized testing. The novelty lies in integrating CEFR-based classification with corpus-driven methods and natural language processing tools. The system also offers educational utility through its exportable teaching materials, which can help instructors align vocabulary instruction with real test demands. Despite some limitations—such as classification gaps and inconsistent readability metrics—the proposed framework provides a replicable model for future research. Upcoming studies may expand the scope by including other TOEIC sections or applying the tool to different learner corpora. In sum, this study demonstrates that automated vocabulary profiling enhances our understanding of lexical demands in listening assessments and offers practical pathways for improving curriculum design, test preparation, and pedagogical strategies in EFL contexts.

6. Declarations

6.1. Author Contributions

Conceptualization, B.S., K.P., N.P., and W.C.; methodology, B.S., K.P., and W.C.; software, W.C.; validation, B.S., K.P., N.P., and W.C.; formal analysis, B.S., K.P., and W.C.; investigation, B.S., K.P., and W.C.; resources, B.S., K.P., and W.C.; data curation, B.S., K.P., and W.C.; writing—original draft preparation, B.S., K.P., N.P., and W.C.; writing—review and editing, B.S., K.P., N.P., and W.C.; visualization, W.C.; supervision, W.C.; project administration, K.P. and W.C.; funding acquisition, B.S., K.P., and W.C. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This article was supported by the Fundamental Fund of Khon Kaen University. The research on "Investigation of Lexical Profile and Global Englishes Awareness of Thai and Vietnamese Learners with Different CEFR-based English Proficiency Levels" by Khon Kaen University has received funding support from the National Science, Research and Innovation Fund ("NSRF") (Grant number: 4708233).

6.4. Acknowledgments

The authors would like to thank Mr. Paiboon Manorom, who supports the dataset for analysis.

6.5. Ethical Statement

This study has received approval from the Institutional Review Board (IRB), confirming compliance with all ethical standards for research involving human subjects through an exemption under approved conditions.

6.6. Informed Consent Statement

Not applicable.

6.7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Pecorari, D., Shaw, P., & Malmström, H. (2019). Developing a new academic vocabulary test. Journal of English for Academic Purposes, 39, 59–71. doi:10.1016/j.jeap.2019.02.004.

[2] Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. Studies in Second Language Acquisition, 30(1), 79–95. doi:10.1017/S0272263108080042.

[3] Sun, D., Chen, Z., & Zhu, S. (2023). What affects second language vocabulary learning? Evidence from multivariate analysis. Frontiers in Education, 8. doi:10.3389/feduc.2023.1210640.

[4] Tran, Y. (2023). Improving English Vocabulary for Students Through Listening to English News. International Journal of Language and Literary Studies, 5(1), 1–13. doi:10.36892/ijlls.v5i1.1152.

[5] Meebangsai, D., Pongtin, P., Kitipoontanakorn, P., & Laosrirattanachai, P. (2023). Investigating Proficiency of Academic English in Student Writing: A Comparative Case Study on Vocabulary Utilization in Student Research Article Writing vis-à-vis National and International Research. Pasaa, 67(1), 66–100. doi:10.58837/chula.pasaa.67.1.3.

[6] Jang, W., & Leech, K. (2023). Contextual Modulation of Adult–Child Language Interaction: Semantic Network Connectivity and Children's Vocabulary Development. Education Sciences, 13(11), 1084. doi:10.3390/educsci13111084.

[7] Laufer, B. (1989). What percentage of text lexis is necessary for comprehension? Special Language: From Humans Thinking to Thinking Machines, C. Lauren and M. Nordman, Eds. Bristol: Multilingual Matters, 316-323. Available online: https://www.lextutor.ca/cover/papers/laufer_1989.pdf (accessed on August 2025).

[8] Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, 34(2), 213. doi:10.2307/3587951.

[9] Kaneko, M. (2020). Lexical Frequency Profiling of High-Stakes English Tests: Text Coverage of Cambridge First, EIKEN, GTEC, IELTS, TEAP, TOEFL, and TOEIC. JACET Journal, 64(64), 79–93.

[10] Yildiz, M. (2023). Lexical Coverage Required for Minimal and Optimal Levels of Reading Comprehension in the English Tests of the Higher Education Institutions Examination. REFLections, 30(3), 695–711. doi:10.61508/refl.v30i3.268077.

[11] Phung, D. H., & Ha, H. T. (2022). Vocabulary Demands of the IELTS Listening Test: An In-Depth Analysis. SAGE Open, 12(1), 1–13. doi:10.1177/21582440221079934.

[12] Towns, S. G. (2020). Which Word List Should I Teach? Using Word Lists to Support Textbook Vocabulary Instruction. THAITESOL Journal, 33(1), 20–35.

[13] Sukying, A. (2023). The role of vocabulary size and depth in predicting postgraduate students' second language writing performance. LEARN Journal: Language Education and Acquisition Research Network, 16(1), 575-603.

[14] Nazri, M. A., Fikni, Z., Hijjah, P., & Wati, L. (2024). Engaging English Language Learners: The Impact of Pictionary on Student Interest and Vocabulary Retention in EFL Classrooms. ELT Worldwide: Journal of English Language Teaching, 11(2), 487. doi:10.26858/eltww.v11i2.64579.

[15] West, M. (1953). A general service list of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longman, London, United Kingdom.

[16] Seong, S., & Cha, J. (2023). Domain Word Extension Using Curriculum Learning. Sensors, 23(6), 3064. doi:10.3390/s23063064.

[17] Liu, C., Wang, S., Qing, L., Kuang, K., Kang, Y., Sun, C., & Wu, F. (2024). Gold Panning in Vocabulary: An Adaptive Method for Vocabulary Expansion of Domain-Specific LLMs. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 7442–7459. doi:10.18653/v1/2024.emnlp-main.4244.

[18] Schäfer, J., Reuter, T., Karbach, J., & Leuchter, M. (2024). Domain-specific knowledge and domain-general abilities in children's science problem-solving. British Journal of Educational Psychology, 94(2), 346–366. doi:10.1111/bjep.12649.

[19] Manorom, P., Hunsapun, N., & Chansanam, W. (2024). An investigation of English speaking problems of Chinese EFL students in Thailand. International Journal of English Language and Literature Studies, 13(2), 270–286. doi:10.55493/5019.v13i2.5046.

[20] Panmei, B. (2023). Strategic Vocabulary Learning in Vocabulary List Learning: Insights from EFL Learners in Thailand. 3L The Southeast Asian Journal of English Language Studies, 29(1), 93–107. doi:10.17576/3l-2023-2901-07.

[21] Raungsawat, N., & Chumworatayee, T. (2021). The Effects of Vocabulary Self-Collection Strategy Instruction on Thai EFL Undergraduate Students' Vocabulary Knowledge and Perceptions. Arab World English Journal, 12(1), 253–269. doi:10.24093/awej/vol12no1.18.

[22] Rofiah, N. L., & Waluyo, B. (2024). Effects of Gamified Grammar and Vocabulary Learning in an English Course on EFL Students in Thailand. Teaching English with Technology, 24(2), 22-46. doi:10.56297/vaca6841/lrdx3699/djjl1101.

[23] Tiansoodeenon, M., & Prasongngern, P. (2025). Enhancing Active Learning through the Interactive Learning Platform to Improve Thai EFL Students' English Vocabulary, Grammatical Retention, and Motivation in English Learning. Higher Education Studies, 15(1), 232. doi:10.5539/hes.v15n1p232.

[24] Kaneko, M. (2017). Vocabulary size targets for the TOEIC test. JACET Journal, 61, 57-67.

[25] Nation, P., & Beglar, D. (2007). A vocabulary size test. Plenary Speaker, JALT2007, 1-4.

[26] Sharpe, P. J. (2018). Barron's TOEIC Practice Exams. Barron's Educational Series, New York, United States.

[27] Collins, H. (2019). Collins Skills for the TOEIC Test: Listening and Reading (2nd Ed.). Harper Collins, London, United Kingdom.

[28] ETC. (2014). Test and score data summary for TOEFL iBT® tests: January 2013–December 2014 test data. Educational Testing Service (ETC), Princeton, United States.

[29] Trew, G. (2007). Tactics for TOEIC: Listening and reading test. Oxford University Press, Oxford, United Kingdom.

[30] Sabatini, J., O'Reilly, T., & Doorey, N. a. (2018). Retooling Literacy Education for the 21st Century: Key Findings of the Reading for Understanding Initiative and Their Implications. Educational Testing Service. Available online: https://files.eric.ed.gov/fulltext/ED587186.pdf (accessed on August 2025).

[31] Educational Testing Service and Kaplan Test Prep. (2009). Kaplan's TOEIC Listening and Reading Prep Plus 2009–2010. Kaplan Publishing, New York, United States.

[32] Lougheed, L. (2007). Longman Preparation series for the new TOEIC test-More practice tests (4th Ed.). Pearson education, Inc, London, United Kingdom.

[33] Lertcharoenwanich, P. (2022). The Effect of Communicative Language Teaching in Test Preparation Course on TOEIC Score of EFL Business English Students. Journal of Language Teaching and Research, 13(6), 1188–1195. doi:10.17507/jltr.1306.06.

[34] Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2000). Grammar of spoken and written English. Longman, London, United Kingdom.

[35] Cobb, T. (2021). Web Vocabprofile, an adaptation of Heatley, Nation & Coxhead's (2002) Range. Available online: http://www.lextutor.ca/vp/ (accessed on August 2025).

[36] Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Victoria University of Wellington, Wellington, New Zealand.

[37] Li, H. (2025). Teaching academic English in higher education: strategies and challenges. Frontiers in Education, 10. doi:10.3389/feduc.2025.1559307.

[38] Little, D. (2020). Common European Framework of Reference for Languages. The TESOL Encyclopedia of English Language Teaching, 1–7, John Wiley & Sons, Hoboken, United States. doi:10.1002/9781118784235.eelt0114.pub2.

[39] Laurence Anthony. (2020). AntConc (Version 3.5.9), Waseda University, Tokyo, Japan. Available online: https://www.laurenceanthony.net/software (accessed on August 2025).

[40] Salkind, N. (2010). Encyclopedia of Research Design. Sage Publishing, Thousand Oaks, United States. doi:10.4135/9781412961288.

[41] Sedgwick, P. (2014). Cross sectional studies: advantages and disadvantages. BMJ, 348, 2276. doi:10.1136/bmj.g2276.

[42] Tangsakul, S. (2024). Relationship Between Vocabulary Size and TOEIC Reading Achievement Among Undergraduate Students. International Journal of Sociologies and Anthropologies Science Reviews, 4(4), 305–312. doi:10.60027/ijsasr.2024.3896.

[43] Tsai, R. M. R., & Huang, S. C. (2023). EFL reading strategies used by high school students with different English proficiency. Forum for Linguistic Studies, 5(3), 1855–1855. doi:10.59400/fls.v5i3.1855.

[44] Namsaeng, P. (2021). An Analysis and Techniques Used for TOEIC Test Takers in Thailand. Journal of Liberal Arts and Service Industry, 4(2), 658-683.

[45] Chinda, B., & Hinkelman, D. (2023). Teacher Cognition of EFL Assessment: A Case Study of Professional Development on Performance-based Language Assessment in Japan. REFLections, 30(3), 757–775. doi:10.61508/refl.v30i3.268136.

[46] Thongsonkleeb, K. (2023). Students' Satisfaction with the Use of Google Forms for the TOEIC Test to Evaluate English Proficiency. 2023 8th International Conference on Business and Industrial Research (ICBIR), 1303–1306. doi:10.1109/icbir57571.2023.10147504.

[47] Bancha, W., & Tongtep, N. (2020). Effects of TOEIC vocabulary lessons plus LMS exercises and TOEIC vocabulary lessons plus MultiEx games on the short-term vocabulary memorization and long-term vocabulary retention of Thai tertiary students. Project Code FIS R6302, Prince of Songkla University, Hat Yai, Thailand.

[48] Li, C. H. (2023). Exploring aural vocabulary knowledge for TOEIC as a language exit requirement in higher education in Taiwan. IRAL - International Review of Applied Linguistics in Language Teaching, 62(4), 1853–1875. doi:10.1515/iral-2023-0021.

[49] Janjaroongpak, K. (2022). CEFR-referenced item specification analysis of TOEIC incomplete sentences part on Intermediate Thai learners. St. Theresa Journal of Humanities and Social Sciences, 8(2), 61-76.

[50] Ha, H. T., Le, H. T., Phung, D. H., & Nguyen, S. D. (2022). Is "general" easier than "academic"? A corpus-based investigation into the two modules of IELTS reading test. SN Social Sciences, 2(8), 159. doi:10.1007/s43545-022-00461-1.