HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1075.2023-0080
Educational Sciences, 2023, Volume 68, Issue 4, pp. 21-33
This paper is available online at http://stdb.hnue.edu.vn

A CORPUS-BASED INVESTIGATION INTO LEXICAL COVERAGE AND AWL OCCURRENCE IN THE ENGLISH TEST OF VIETNAM'S NATIONAL HIGH SCHOOL GRADUATION EXAMINATION

Nguyen Thi Nhan[1*], Do Hai Vu[2], Vu Do Hoai Phuong[2], and Do Tran Phuong Anh[2]
[1]*Faculty of English, Hanoi National University of Education*
[2]*Undergraduate student, course 70, Faculty of English, Hanoi National University of Education*

Abstract. This corpus-based study examines all codes of the English papers of Vietnam's National High School Graduation Examination (from 2019 to 2022) regarding lexical coverage as well as academic words appearing in the texts. With the utilization of the Vocabprofile program available on https://www.lextutor.ca/ and two corpora, the Academic Word List (AWL) and BNC-COCA-25K, the results of the investigation revealed that exam-takers were generally required to master the first 3000-word families for 95% comprehension and roughly the first 5000-word families for the rate of 98%, which could be a challenge for them. Another concern is the percentage of AWL items that appeared in these papers was sharply lower than the average in academic contexts. The results not only offer teachers and students insights into the required language input but also emphasize the significance of the number of academic words for comprehension. Additionally, our study indicates that the lexical demands across different examination years appear to be consistent.

Keywords: Lexical coverage, Lexical complexity, Academic word list, English test, Vietnam's National High School Graduation Examination.

1. Introduction

In the context of teaching and learning English as a foreign language, a profound understanding of the lexical threshold holds significant implications, particularly in courses that center around reading. Such comprehension enables teachers and curriculum designers to establish vocabulary objectives and design syllabi that concentrate on lexical development (Laufer & Ravenhorst-Kalovski, 2010) [1; 15-16]. Also, it was confirmed that unsatisfactory comprehension is more commonly correlated to insufficient lexical coverage rather than an abundance of vocabulary (Laufer, 1989) [2; 321]. In addition, an important objective for language learners is to acquire a vast vocabulary in their second language (Nation, 2013) [3; 17] particularly as academic vocabulary forms a critical subset of vocabulary knowledge crucial in text comprehension. Hence, these three concepts of lexical threshold—text coverage, and academic vocabulary-play a pivotal role for educators and learners in the context of English language teaching and learning.

Various studies have examined language input and the use of academic language in educational materials. For example, the language input in Indonesian textbooks was highly challenging for students (Aziez & Aziez, 2018) [4; 76], while the lexical coverage of 895 reading

Received July 11, 2023. Revised August 10, 2023. Accepted September 1, 2023.
Contact Nguyen Thi Nhan, e-mail address: nhannt@hnue.edu.vn

passages in Chinese Senior High School Entrance Tests was 92.82%, lower than the 95% threshold (Jin & Li, 2016) [5; 962]. The Thailand University Admission English Tests contained 4.58% of the total number of words that belonged to the academic word list (AWL) (Cherngchawano & Jaturapitakkul, 2014) [6; 21]. Other studies including an investigation into English articles on medicine showed that 10.07% of words from English articles on medicine were taken from the AWL (Chen and Ge, 2007) [7]. Likewise, an analysis of the Malaysian University English Test (MUET) (Charles, Shabdin & Affendi, 2021) [8] concluded that 15 out of 18 passages required Malaysian students to master no less than 6000-word families so that they would understand 95% of the texts' content. An inspection of academic spoken language revealed that 4.41% of the words were academic (Dang & Webb, 2014) [9].

Yet, the number of studies in Vietnam on the lexical threshold, text coverage, academic vocabulary, and academic lexical items is limited. The investigation into the English Test of Vietnam's National High School Graduation Examination was one that examined these three dimensions (Vu, 2019) [10; 19-38]. In this study, 20 English papers were collected over 16 years starting from 2002. The results showed that Vietnamese students generally needed to know 5000-word families to acquire a 95% understanding of the exams and up to 14000 for the rate of 98%. However, variations between different years were possible. The author also concluded that the AWL words were used less frequently in the English Test of VNHSGE in comparison with other academic examinations and encouraged high school students to learn this academic word list due to its frequency in assessing English proficiency (Vu, 2019) [10; 19-38].

Vietnam's National High School Graduation Examination (VNHSGE) (Kỳ thi THPT Quốc gia) has been held annually since 2015, according to Decision No. 3538/QĐ-BGDĐT [11]. As of 2017, according to the MOET, five papers have been incorporated into the assessment: Mathematics, Literature, Foreign Language, Natural Sciences, and Social Sciences (according to Circular No. 04/2017/TT-BGDĐT) [12]. The English test of VNHSGE plays a pivotal role in assessing students' English proficiency once they have finished grade 12. In reality, the results of the English tests in the past three years did not show positive outcomes (Nguyen & Duong, 2020) [13], which is verifiable via the average scores of this test published by the MOET (4.6 in 2017, 3.91 in 2018, 4.36 in 2019, 4.578 in 2020, 5.84 in 2021, 5.15 in 2022).

Therefore, the authors decided to investigate the English test of VNHSGE from 2019 to 2022, aiming to address the public curiosity about the uniformity among all codes in the same year through three questions:

(1) What is the required number of word groups that students must learn in order to achieve a vocabulary coverage of 95% and 98% for the English paper included in VNHSGE between 2019 and 2022?

(2) To what extent are the academic words (AWL) present in The English Test of VNHSGE from 2019 to 2022?

(3) Are the levels of lexical complexity the same across the test codes?

2. Content

2.1. Theoretical background

2.1.1. Corpus

Definitions of corpus

A corpus is defined as an electronically stored collection of texts, written or spoken, sampled to represent a particular language or language variety (O'Keeffe et al., 2007) [14; 1]. Another definition by Dash and Arulmozi (2018) [15; 4] defines it as a language database statistically sampled for analyzing, describing, and investigating various aspects of language. Both definitions

emphasize the value of corpora in linguistic research but differ in their specific details. The corpus in this study is a collection of texts sampled for the investigation of language used in the context of the national exam.

Types and Tokens

In the book *Statistics in Corpus Linguistics*, a token, also known as a running word, refers to a single occurrence of a word and any sequence of letters or numbers separated by white space. A type is illustrated as a distinctive form of a word present in the corpus (Brezina, 2018) [16; 39]. As word counts, including mostly token and type counts, are integral to most statistical analyses discussed in this study, it is crucial to understand these definitions thoroughly.

BNC-COCA-25K framework

The BNC/COCA word family lists comprise 29 lists of word families, out of which 25 (K-1 to K-25) lists are created based on frequency and range data. The remaining four lists include a list of proper names that continue to expand, a list of marginal words comprising swear words, exclamations, and alphabet letters, a list of transparent compounds, and a list of acronyms (Cobb, 2023) [17]. Multiple studies have utilized the BNC-COCA-25K framework such as the investigation into lexical coverage of reading passages in Chinese Senior High School Entrance Tests (Jin & Li, 2016) [5], the analysis of the Malaysian University English Test (Charles, Shabdin & Affendi, 2021) [8] and the investigation into the English Test of Vietnam's National High School Graduation Examination between 2002 and 2018 (Vu, 2019) [10].

2.1.2. Vocabulary coverage in English as a Foreign Language learning

Text coverage

Lexical coverage, or vocabulary coverage, is the words in the text that are understood by readers (Laufer and Ravenhorst-Kalovski, 2010) [1; 17].

Lexical threshold

First introduced in 1989, the concept of the lexical threshold refers to the minimum vocabulary required for satisfactory reading comprehension (Laufer, 1989) [2; 319]. The 95% lexical coverage can guarantee basic text comprehension (Laufer, 1989) [2; 321] while this number should be raised to 98% (Hu & Nation, 2000) [18; 422]. Even though the 3% seems minor, it involves a significant number of word families. Regardless of the optimal percentage, there is a significant correlation between lexical coverage and reading comprehension (Rahmat & Coxhead, 2021) [19; 806].

Word family

In terms of reading, a word family is formed by a base, and all its inflections, and derivations, which English learners are able to understand when encountering without knowing each of them separately (Laurie Bauer & Paul Nation, 1993) [20; 253]. As an example, with a base form *develop*, its inflections could be *developmental, developmentally, developmentwise;* its derivations could be *undeveloped, undevelopable, semi-developed.*

Word classification

The idea of word categories (high and low) was proposed in 2006 (Nation, 2006) [21; 62]. However, an additional word classification mid-frequency vocabulary was introduced in 2013 (Nation & Anthony, 2013) [22; 9]. The first-word category is high-frequency words, which is encountered most frequently in any text (Dang, 2020) [23; 297]. They possibly are responsible for 80% of text comprehension. Although mid-frequency words occur less frequently than high-frequency words, they are frequent enough to be a sensible learning goal after high-frequency words (Schmitt & Schmitt, 2012) [24; 495]. Finally, the low-frequency group is small in number, scarce in occurrence, and may be used in particular contexts (Nation & Anthony, 2013) [22; 8].

Table 1. Word categories (Adopted from Nation & Anthony, 2013)

Level	1000-word family lists	Example
High frequency	1000-3000	*air, jump, world…*
Mid frequency	4000-9000	*endemic, dove, circus…*
Low frequency	> 10000	*vinify, cava, haole…*

2.2. Methodology

2.2.1. Subjects of the study

A corpus of 30791 tokens was built from the downloadable copies of the chosen English Test of VNHSGE from 2019 to 2022. In the VNHSGE, there were 24 exam codes, developed from 4 original ones in this high-stake examination (Nguyen, T., 2017) [25], (Le, T., 2017) [26]. For analysis purposes, the corpus was divided into 20 sub-corpora (16 for the original codes and 4 for separate years).

2.2.2. Data collection procedure

The authors developed a corpus from the digital versions of the English Test of VNHSGE spanning from 2019 to 2022. To process the text, the authors adopted the text processing procedure outlined in 2021 (Rahmat & Coxhead, 2021) [19; 807-808], which involved the following steps:

1. File Conversion: The materials were stored as PDF documents and later transformed into text files with a ".txt" extension. Any elements that could not be converted automatically were transcribed manually.

2. Quality Check: Throughout the conversion process, the authors meticulously checked each page for any technical glitches or spelling errors. When such an error was detected, it was corrected with close attention to the original text's format. Additionally, the authors removed any unrecognized text or symbols, including special characters or abbreviations that were not explicitly defined in the papers.

3. Handling Hyphenated Words: With regards to hyphenated multi-word terms, the decision to replace the hyphen with a space to separate them into single words or remove the hyphen altogether and group the multi-item words into the off-list category for further analysis was based on the overlap in meaning between the individual words and the multi-word item (Webb & Nation, 2008) [27; 15]. Some examples of hyphenated words to be separated included *nitrogen-rich, man-made, thousand-mile-wide, e-readers, name-makers, global-scale, etc*.

4. Proper Nouns Handling: Proper nouns were eliminated from the corpus by the Vocabprofile program due to their excessive number and the controversy about whether they are lexical items. The program identifies mid-sentence capitalization as a marker of proper nouns, but its accuracy is limited to about 90% due to reliance on punctuation (Cobb, 2023) [28].

5. Compounding Nouns Handling: Compounding nouns are useful in English, but the sheer number of medium and low-frequency compound words makes creating a comprehensive list difficult. Traditionally, the solution has been to exclude less common words, but this can lead to reduced accuracy. To address this, Lextutor's profilers split compound words into individual parts and assign them to a K-list. This approach reflects how learners perceive and comprehend text. However, there are limitations, including dependence on the BNC-COCA framework and proper nouns only appearing in the 1k list. (Cobb, 2023) [29].

6. Off-list Words: It should be noted that words that do not belong to the BNC-COCA-25K and AWL corpus such as *dishwasher, shopkeeper, or cyberbullying* were listed in the off-list by the Vocabprofile tool. In addition to this, Latin- or French-borrowed words such as *résumé, café, or cafetière* are out of the list.

7. Pilot Analysis: The input texts were uploaded to the Vocabprofilers program for pilot analysis to detect any technical or spelling errors. Following confirmation of the corpora, they were deemed fit for analysis.

2.2.3. Data analysis procedure

To analyze the corpora, the authors utilized the BNC-COCA 1-25K and CLASSIC (GSL/AWL) program available on the website https://www.lextutor.ca/. The programs treated each word in the document as one token, and each distinct word as one type when the materials were uploaded for analysis (Webb & Nation, 2008) [27; 10].

The approach for examining the lexical input present in the corpora was derived from two studies (Rahmat & Coxhead, 2021; Vu, 2019) [19; 808, 11; 20]. The process is outlined below:

To address the research question on the number of word families with which candidates should be acquainted to achieve 95% and 98% lexical coverage of the papers, the outcome table was derived from Vocabprofilers after data had been analyzed by the BNC-COCA 1-25K program. This table comprises the threshold for each of the 1,000-word lists in the datasets. Additionally, the four sub-corpora were analyzed independently to facilitate a more comprehensive comparison of the lexical coverage in each paper.

To answer the second question on the appearance of academic words in these English papers, the CLASSIC (GSL/AWL) program was used for analysis. Initially, each test paper each year was examined separately. After that, 4 sub-corpora each year were merged into a larger sub-corpus which was fed into the Vocabulary Profile program for analyses.

To tackle the question of lexical demands' consistency, the first and second question results were used to form the final answer to the third one.

2.3. Results and discussions

2.3.1. Lexical demands of the English Test of VNHSGE

2.3.1.1. 2019 Examination

Table 2. The lexical levels in the 2019 papers

Freq.Level	Code 1	Code 2	Code 3	Code 4
K-1	76.6	81	78.4	78.2
K-2	88.7	90	88.8	88.5
K-3	**96.6**	**95.9**	94.5	**95.1**
K-4	**98.7**	97.8	**96.8**	96.8
K-5	99.1	**98.4**	**98.6**	**98.4**
K-6	99.5	98.7	99.1	99
K-7	99.7	99	99.2	99.1
K-8	99.8	99.1	99.3	99.7
K-9	99.9		99.4	
K-13				99.8
K-15			99.5	
K-17		99.2		99.9
K-25			99.6	

Off-list: affordably, amiability, babysitter, cafés, cyber, cyberbullying, dropouts, equivalency, landfills, shopkeepers, wrongdoings.

Table 2 shows the data on lexical demands among four codes of the English Test of VNHSGE in 2019. It is evident that the lexical difficulty levels of all 2019 sets of questions slightly varied between the codes. Specifically, Codes 1, 2, and 4 required test-takers to reach the K-3 frequency level to acquire 95% comprehension of contents, but it was K-4 for this point in code 3. Meanwhile, code 1's 98% threshold stood at K-4, but K-5 for the other codes. Additionally, several unfamiliar words ranging from K-13 up to K-25 including *broadsheets, desalinate, desalinated, desalination* (K-13); *bur* (K-15); *broads, ceilidh* (K-17); *dhow, Hogmanay* (K-19); *cyberspace* (K-25) were used in codes 2, 3 and 4. The words *dhow and Hogmanay* are considered to be technical terms that are seen in a limited number of situations and seemingly challenging to guess the meaning based on the surrounding context. Similarly, at level K-25, test-takers needed to know the combining form *Cyber*, which is listed in the B2 level based on CEFR (Oxford Learner's Dictionaries) [30]; consequently, a number of them might not have understood this word because it was not a member of the expected lexical list after having finished high school in Vietnam (Circular No. 32/2018/TT-BGDĐT) [31; 26]. Then, it can be deduced that with a lexical load of 3000-word families, Vietnamese students had hit 95% of these papers, and needed roughly 5000 of the most frequent words to understand nearly the entire text in 2019. As illustrated, the level of lexical use in 4 codes in 2019 was slightly different.

2.3.1.2. 2020 Examination

Table 3. The lexical levels in the 2020 papers

Freq.Level	Code 1	Code 2	Code 3	Code 4
K-1	81.2	81.4	80	81.5
K-2	91	91	90.9	90.4
K-3	**95.9**	**96.5**	**95.7**	**96.3**
K-4	97.7	97.6	**98.1**	**98.2**
K-5	**98.3**	**98.6**	98.6	98.9
K-6	98.9	99.2	99.1	99.3
K-7	99.1	99.3	99.4	
K-8	99.2		99.5	99.4
K-9		99.4		99.5
K-10	99.3			
K-12		99.5		99.6

off-list: eco, gridder, gridders, heartbeats, hydro, kWh, outsourced, résumé, unadventurous, urbanites.

Table 3 illustrates the lexical demands of 4 question papers for the English Test of VNHSGE in 2020. Overall, students who did these papers experienced the nearly-the-same level of vocabulary. To be more precise, the third 1000-word family was the threshold for 95% comprehension of all 4 codes. Apart from this similarity, the differences between them are mainly laid in the 98% threshold, which required students to grasp the K-4 word group in codes 3 and 4; K-5 in 1 and 2. It can be summarized that word frequency varied from K-2 to K-12. K-9 word list included *convection, crossword, extrovert, tertiary,* and *unwind*; K-10 *inflammable* and *sufficiency*; K-11 *broadband*; K-12 *elixir, overrated, and pandemic*; K-13 *offside*, K-17 *hydropower*; K-21 *fashionista*. Among these words, many were tougher than the expected vocabulary level - B1 (according to Circular No. 32/2018/TT-BGDĐT) [31; 26], namely *hydropower* (C2), *offside* (C2), *broadband* (C1), *pandemic* (B2) (Oxford Learner's Dictionaries.

Besides, it was hard to infer the meanings of such compound nouns as *offside* and *broadband* whether splitting them apart or depending on the contexts. As a result, full comprehension of the 4 codes in 2020 might have been demanding for Vietnam's high school students.

2.3.1.3. 2021 Examination

Table 4. The lexical levels in the 2021 papers

Freq.Level	Code 1	Code 2	Code 3	Code 4
K-1	85.8	85.6	82.9	86.2
K-2	94.8	94.7	94.1	**95.4**
K-3	**98.8**	**98.7**	**98.3**	**98.3**
K-4	99.4	99.3	99.2	99.6
K-5	99.9	99.8	99.8	99.8
K-6	100	99.9	99.9	
K-9		100		
K-10			100	99.9

off-list: words in this sub-corpus are all in BNC-COCA-25K

Table 4 compares the frequency levels (K-1 from K-25) of 4 codes in the 2021 English Test of VNHSGE. Notably, no lexical items are grouped into the off-list. It is noticeable that 3000-word families equaled 98% coverage of all codes and 2000-word families for approximately 95%. More importantly, the 100% threshold varied in these papers. It was K-6 for total comprehension in code 1, but K-9 in code 2, and K-10 in code 3. Differently, the outcome from the VP tool stopped at 99.9% for the highest understanding, which meant that a small number of words were in K-26 and likely beyond. Moreover, there were no low-frequency words in code 1; these lexical items were used limitedly in the others (*sociable*, K-9, and *spokes*, K-10). Because of these relatively modest lexical demands, students likely found these papers manageable.

2.3.1.4. 2022 Examination

Table 5. The lexical levels in the 2022 papers

Freq.Level	Code 1	Code 2	Code 3	Code 4
K-1	85.2	87.4	86	83.8
K-2	93.4	94.1	94	93.3
K-3	**97.8**	**97.9**	**97.8**	**97.8**
K-4	**98.7**	**99**	**98.8**	**98.6**
K-5	99.4	99.7	99.6	99.5
K-6		99.8	99.7	99.6
K-7	99.5	99.9	99.9	99.7
K-8	99.7	100	100	99.8

off-list: café, de.

Table 5 describes the various lexical coverage and text comprehension of the 2022 English Test of VNHSGE. Knowing 3000-word families guaranteed 95% to nearly 98% understanding of the whole document, and 4000-word families for 98% to 99%. Students just needed mid-frequency words to gain the 100% threshold for code 2 and code 3; however, it required a number of low-frequency words to reach 99.9% of language coverage in codes 1 and 4, which was K-11

and K-10 respectively. This year, just 2 low-frequency words appeared in the exam, specifically, *outperform* (K-10) *and talkative* (K-11). If students had been unfamiliar with the word *outperform*, they could have easily broken the word down into two smaller parts: the prefix *out-* meaning greater, better (Oxford Learner's Dictionaries | Find Definitions, Translations, and Grammar Explanations at Oxford Learner's Dictionaries, n.d.) [30] and the base *perform* to guess the meaning.

2.3.1.5. Discussions

Lexical demands and text comprehension are key to language processing. The former refers to a text's vocabulary and grammar, while the latter refers to a reader's ability to understand and interpret it. An analysis of the BNC-COCA-25K corpus showed that the input language coverage of test papers changed slightly each year from 2019 to 2022 and that the language input was less challenging than from 2002 to 2018.

For a 95% comprehension rate, 2000-word families (2021) and 3000-word families (2019, 2020, 2022) were sufficient, compared to an average of 5000-word families from 2003 to 2018. The study found that lexical demands in 2019, 2020, and 2022 were as difficult as in 2005 but easier than in 2002, where 12,000-word families were required for 95% understanding.

Exam-takers needed to know around 4000-word families to reach the 98% threshold from 2019 to 2022, compared to over 14000 in previous years. Low-frequency words were added to the papers from 2002 to 2018, with a higher density than from 2019 to 2022. The 2021 papers had the simplest lexical input in four years, possibly due to school closures during the pandemic. Test-takers needed to master high-frequency words for a 95% comprehension rate, mid-frequency words for 98%, and low-frequency words for a thorough understanding.

Vietnamese students typically have a vocabulary load of 2500 words upon finishing secondary education (Circular No. 32/2018/TT-BGDĐT) [31; 26], which can be a challenge for higher scores on the English exam. The research recommends high school students, especially 12^{th}-graders, need to learn more than 2500 expected lexical items, particularly in the K-5 list, to achieve the 98% threshold, and receive appropriate vocabulary instruction and practice to improve their results.

2.3.2. The coverage of AWL in the English Test of VNHSGE

The idea of the AWL was developed by Averil Coxhead, a lecturer in Applied Linguistics at the Victoria University of Wellington in New Zealand. Coxhead analyzed a large corpus of academic texts from various disciplines and identified the most frequently occurring words. The list is divided into 10 sublists, with each sublist containing 60 words that increase in difficulty and complexity (Coxhead, 2000) [32; 228].

2.3.2.1. 2019 Examination

Table 6. The presence of AWL in the 2019 English Test of VNHSGE

Code	Types (%)	Token (%)
1	80 (12.20)	111 (8.1)
2	58 (8.96)	89 (5.3)
3	83 (12.75)	109 (8.0)
4	69 (10.52)	91 (6.5)

It is clear that there was a minor discrepancy in AWL occurrence in types between code 2 of 2019 and the others. Regarding unique words, except for code 2 where total AWL items covered 8.96%, the others contained more than 10%. As a result, the percentage of AWL members in these papers was not seemingly analogous.

2.3.2.2. 2020 Examination

Table 7. The presence of AWL in the 2020 English Test of VNHSGE

Code	Types (%)	Token (%)
1	74 (11.58)	105 (7.2)
2	59 (9.28)	105 (7.5)
3	74 (10.82)	119 (7.6)
4	71 (10.97)	106 (7.8)

It can be inferred from Table 7 that the differences in using AWL in these papers appeared insignificant since the AWL proportion in tokens ranged from 7.2% to 7.8%. To be precise, there were approximately 70 different lexical items belonging to AWL in the first, third, and fourth codes, but this data was reduced by some 10 words in the second one.

2.3.2.3. 2021 Examination

Table 8. The presence of AWL in the 2021 English Test of VNHSGE

Code	Types (%)	Token (%)
1	32 (6.79)	55 (4.5)
2	32 (6.84)	55 (4.6)
3	32 (6.64)	61 (5.3)
4	27 (5.73)	43 (3.6)

The number of AWL words in all question sets this year was nearly the same picture, where 32 different items from this list were put into these papers, but 27 in code 4. In other words, 32 unique AWL words accounted for around 6.5% of the entire text in codes 1, 2, and 3 while 27 took up 5.73% in code 4.

2.3.2.4. 2022 Examination

Table 9. The presence of AWL in the 2022 English Test of VNHSGE

Code	Types (%)	Token (%)
1	44 (7.87)	75 (5.9)
2	39 (7.26)	61 (5.3)
3	42 (7.64)	63 (5.4)
4	48 (8.38)	77 (6.1)

The AWL type in papers 1, 2, and 3 was some 7.5% of the total text types whereas it was slightly higher in code 4 (8.38%). These data proved that 4 question sets in 2022 were designed comparably in terms of academic words.

2.3.2.5. Discussions

Table 10. The presence of AWL in the English Test of VNHSGE from 2019 to 2022

Year	Types (%)	Tokens (%)
2019	244 (13.95)	390 (8.1)
2020	189 (12.66)	417 (8.1)
2021	51 (6.59)	213 (4.8)
2022	69 (8.14)	274 (6.4)

Table 10 demonstrates the use of AWL in the National English Exam papers in a 4-year period from 2019. Overall, it can be deduced that the difference in the use of AWL in all codes each year was trivial aside from the year 2019; therefore, the exam design guaranteed equality towards all sets of questions. Also, the percentage of AWL items decreased sharply from 2019 to 2021 but increased slightly in 2022. Specifically, there were 244 unique words in AWL presented in the 2019 paper, but this size dwindled to 189 in 2020; to 69 in 2022; dramatically to 51 in 2021. Even though the 2019 and 2020 papers contained the highest proportion of AWL's words, which were both 8.1% of cumulative coverage, this contribution was lower than the average AWL presence, roughly 10% of cumulative coverage (Coxhead, 2000) [32; 226].

This could suggest a shift in the academic lexicon used in these English papers within the given period, especially the enormous downfall between 2020 and 2021. Another inference would be that a valid reason driving this decline in 2021 possibly coincided with the driver of lexical uncomplicatedness mentioned earlier. Moreover, the difficulty levels of these papers may have been slashed to some extent over four years since the number of academic items packaged in texts strongly correlated with a deeper comprehension of passages (Lawrence, J. F. et al., 2022) [33; 681].

2.3.3. Implications

The findings revealed that Vietnamese high school students potentially found it challenging to reach the 95% and 98% threshold in comprehending English papers because a vast population of them did not master the targeted lexical size issued by the MOET, and they seemingly could not completely address English papers because of lexical barrier (Vu & Nguyen, 2019) [35; 32]. Hence, there would be several practical implications after the study was conducted.

2.3.3.1. English exam design process

Firstly, the amount of academic content included in a text is closely linked to the level of difficulty required to comprehend passages more deeply within that (Lawrence, J. F. et al., 2022) [33; 681]. In essence, the MOET should ensure the same level of lexical demands and academic items between different codes each year in order to breed equality for all exam-takers.

Secondly, this examination is designed for those who have finished or graduated from secondary education in general, and some other special test-takers designated by the Minister of Education and Training (15/2020/TT-BGDĐT) [34; 7] but the language input in the given English papers was almost higher than expected. As a result, lexical selection needs to belong to the expected word list (32/2018/TT-BGDĐT) [31; 26].

2.3.3.2 EFL teaching and learning process

First, it benefits EFL teachers if they can grasp high-frequency and mid-frequency word lists, which typically equal 95% and 98% comprehension of the English papers respectively. Once mastered, they will be able to make samples of English papers with the same lexical demands as the sample English exam questions published annually by MOET since 2015. An authentic resource for these lists can be found on the website https://www.lextutor.ca/

Second, with AWL items accounting for a certain number of words in these English papers, a lack of frequent academic words possibly leads to a barrier in understanding passages. For this reason, learning academic vocabulary should be one of the priorities during preparation for this high-stake examination.

3. Conclusion

On the one hand, the study examined the vocabulary requirements of the English Test of VNHSGE from 2019-2022, finding that high-frequency words are necessary for general comprehension, while mid-frequency words provide up to 98% coverage. Less common words are also important for full comprehension. Students needed to know about 3000-word families

for a 95% threshold in 2019, 2020, and 2022, but the same number of words achieved the 98% threshold in 2021. Test-takers in 2019, 2020, and 2022 needed to understand some 5000-word families for 98% comprehension. However, the lexical choices were similar each year, and Vietnamese students' vocabulary upon finishing secondary education may not meet the lexical demands in the papers.

On the other hand, AWL usage varied each year, with the highest number of unique AWL words in 2019 and the lowest in 2021. AWL's presence decreased from 2019 to 2021 but increased slightly in 2022. Despite this, the AWL proportion in each code within the same year was analogous to some extent; the huge gap in AWL population between question sets was found in 2019 and 2020, and for the years 2021 and 2022, these differences seemed insignificant. This signals a potential change in the academic language used in English papers over this period.

With regard to lexical consistency, test designers guaranteed the same or nearly the same quality of lexical demands between all codes in the given period, resulting in equality for test-takers.

Hopefully, these conclusions can fill the public questions on the lexical input levels in the English test of VNHSGE from 2019 to 2022.

Recommendations for further studies

The study entirely concerns the dimensions of language coverage and AWL presence in the English Test of VNHSGE between 2019 and 2022. Other elements including the grammatical diversity, the readability level of the materials, and different aspects of effective evaluations (such as practicality, validity, reliability, and authenticity), etc. are not covered in the scope of the study. Therefore, this might be a paucity of this paper, which is a potential for other authors to exploit in future studies in this field.

REFERENCES

[1] Laufer, B., & Ravenhorst-Kalovski, G. C., 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* (Vol. 22) [1].

[2] Laufer, B., 1989. What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines*: Vol. (pp. 316-323) fix. Multilingual Matters.

[3] Nation, I. S. P., 2013. *Learning vocabulary in another language* (2nd ed.). Cambridge, England: Cambridge University Press.

[4] Aziez, F., & Aziez, F., 2018. The Vocabulary Input of Indonesia's English Textbooks and National Examination Texts for Junior and Senior High Schools. *TESOL International Journal*, *13*(3), 66–77.

[5] Jin, T., Li, Y., & Li, B., 2016, December. Vocabulary Coverage of Reading Tests: Gaps Between Teaching and Testing. *TESOL Quarterly*, *50*(4), 955–964. https://doi.org/10.1002/tesq.324.

[6] Cherngchawano, W., & Jaturapitakkul, N., 2014. Lexical Profiles of Thailand University Admission Tests. *PASAA*, *48*, 1–27. https://eric.ed.gov/?id=EJ1077888.

[7] Chen, Q., & Ge, G. C., 2007, January. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*(4), 502–514. https://doi.org/10.1016/j.esp.2007.04.003.

[8] Charles, B. M., & Affendi, S. A., 2021. Word Frequency Level and Lexical Coverage in the Reading Comprehension Texts of the Malaysian University English Test. *PASAA*, *61*, 33–60.

[9] Dang, T. N. Y., & Webb, S., 2014, January. The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76. https://doi.org/10.1016/j.esp.2013.08.001.

[10] Vu., 2019, December. A corpus-based lexical analysis of Vietnam's high-stakes English exams. https://www.researchgate.net/publication/334657600_A_corpus-based_lexical_analysis_of_Vietnam's_high-stakes_English_exams

[11] Vietnam MOET., 2018. Decision No. 3538/QĐ-BGDĐT: Approving the plan to develop and improve the quality of higher education in the period of 2018-2025. Hanoi, Vietnam.

[12] Vietnam MOET., 2017. Circular No. 04/2017/TT-BGDĐT: Regulations on quality assurance and accreditation of higher education institutions and programs. Hanoi, Vietnam.

[13] Nguyen, & Duong., 2020, August 27. Vietnamese students perform worst in English in national high school exam. *Vnexpress International*. Retrieved February 17, 2023, from https://e.vnexpress.net/news/news/vietnamese-students-perform-worst-in-english-in-national-high-school-exam-4153084.html

[14] O'Keeffe, A., McCarthy, M., & Carter, R., 2007, May 3. *From Corpus to Classroom: Language Use and Language Teaching*.

[15] Dash, N. S., & Arulmozi, S., 2018, February 12. *History, Features, and Typology of Language Corpora*.

[16] Brezina, V., 2018, September 14. *Statistics in Corpus Linguistics: A Practical Guide*. https://doi.org/10.1017/9781316410899

[17] (Cobb, T. *What is COCA? Why does Lextutor need it?* Accessed 13 Mar 2023 at https://www.lextutor.ca/vp/comp/)

[18] Hu, M. H., & Nation, P., 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

[19] Rahmat, Y. N., & Coxhead, A., 2021, January 31. Investigating vocabulary coverage and load in an Indonesian EFL textbook series. *Indonesian Journal of Applied Linguistics*, *10*(3). https://doi.org/10.17509/ijal.v10i3.31768

[20] Bauer, L., & Nation, P., 1993. Word Families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

[21] Nation, I. S. P., 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

[22] Nation, I. S. P., & Anthony, L., 2013. Mid-frequency readers. *Journal of Extensive Reading*, *1*, 5–16.

[23] Dang., 2020. Corpus-Based Word Lists in Second Language Vocabulary Research, Learning, and Teaching. In *The Routledge Handbook of Vocabulary Studies* (Vol. 19, pp. 288–303). Routledge.

[24] Schmitt, N., & Schmitt, D., 2012, February 7. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/s0261444812000018

[25] Tue Nguyen., 2017, June 25. *Bộ GD-ĐT trả lời về sự "vênh nhau" của các mã đề thi*. Thanh Nien. Retrieved March 17, 2023, from https://thanhnien.vn/bo-gd-dt-tra-loi-ve-su-venh-nhau-cua-cac-ma-de-thi-185674827.htm

[26] Le Thu., 2017, June 24. *Bộ GD&ĐT giải thích "độ vênh" giữa các mã đề thi*. Dan Tri . Retrieved March 17, 2023, from http://dantri.com.vn/giao-duc-huong-nghiep/bo-gddt-giai-thich-do-venh-giua-cac-ma-de-thi-20170624174546004.htm

[27] Nation, I. S. P, & Webb, S., 2008. Evaluating the vocabulary load of written text.

[28] Cobb, T. *Why do proper nouns need handling?* Accessed 13 Mar 2023 at https://www.lextutor.ca/vp/comp/

[29] Cobb, *T. Why do compound nouns need special treatment?* Accessed 13 Mar 2023 at https://www.lextutor.ca/vp/comp/

[30] *Oxford Learner's Dictionaries | Find definitions, translations, and grammar explanations at Oxford Learner's Dictionaries*. (n.d.). Oxford Learner's Dictionaries | Find Definitions, Translations, and Grammar Explanations at Oxford Learner's Dictionaries. Accessed 13 Mar 2023 at https://www.oxfordlearnersdictionaries.com/

[31] Vietnam MOET., 2018. Circular No. 32/2018/TT-BGDĐT: Regulations on the implementation of the national framework for qualifications for higher education in Vietnam. Hanoi, Vietnam.

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

[33] Lawrence, J. F., Knoph, R., McIlraith, A., Kulesz, P. A., & Franc, D. J., 2022, April. Reading Comprehension and Academic Vocabulary: Exploring Relations of Item Features and Reading Proficiency. *Reading Research Quarterly*, *57*(2), 669–690. https://doi.org/10.1002/rrq.434

[34] Vietnam MOET., 2020. Circular No. 15/2020/TT-BGDĐT: Regulations on the organization and operation of National High School Examination. Hanoi, Vietnam: Ministry of Education and Training

[35] Vu, D. V., & Nguyen, C. N., 2019, December. An assessment of vocabulary knowledge of Vietnamese EFL learners. In *The 20th English in Southeast Asia Conference*, Date: 2019/12/06-2019/12/07, Location: National Institute of Education, Nanyang Technological University, Singapore