PASAA Journal
Volume 71, July – December 2025, 386–418

PASAA

Establishing English Science Textbook Word Lists (ESTWLs) for Thai EFL Secondary Students: A Corpus-Based Approach

Thana Kruawong[a] and Supakorn Phoocharoensil[b]

[a] Kasetsart University, Bangkok, Thailand

[b] Thammasat University, Bangkok, Thailand

**Corresponding author: supakorn.p@litu.tu.ac.th*

Article information	
Abstract	A number of researchers have worked on creating word lists focused on science to help students in their specific academic fields. However, these lists are usually intended for English for Specific Purposes in university-level EFL programs and may not be suitable for EFL students at other educational levels, like secondary or primary school students. This research aimed to investigate corpus analysis on specialized vocabulary found in English science textbooks at the lower secondary level of Thai schools. The English Science Textbook Corpus (ESTC) comprises 46 textbooks. The corpus, containing 2,076,389 running words, was processed with AntWordProfiler (Anthony 2022) to extract target vocabulary items for learners at this level. Another evaluation through a teacher judgment rating-scale later revealed that 408-word types fulfilled the established criteria. With the completion of the definition review, there were 408-word types that fit according to science teachers with a rating scale approach. After removing function words and expert rating on a scale modified from Chung and Nation (2004),

E-ISSN: 2287-0024

	480-word types were chosen as pedagogically beneficial. These were divided into three lists: the Physical Science List, the Biological Science List, and the Earth & Space Science List. The ESTWLs aim to help lower secondary students learn science and improve their vocabulary in Content and Language Integrated Learning (CLIL) contexts. Science and English teachers may also find the ESTWLs valuable to help them identify core vocabulary for instruction. The full lists will be available at: https://sites.google.com/view/estwl.
Keywords	Specialized Word Lists; Corpus-Based Word Lists; Science Textbooks; Science Secondary School
APA citation:	Kruawong, T., & Phoocharoensil, S. (2025). Establishing English science textbook word lists (ESTWLs) for Thai EFL secondary students: A corpus-based approach. *PASAA Journal*, *71*, 386–418.

1. Introduction

Vocabulary knowledge is a necessary condition for foreign or second proficiency development, and it also is central to academic achievement (Kaya & Charkova, 2014; Rose & McKinley, 2022). In English Medium Instruction (EMI) environments, a considerable number of L2 learners have difficulty with academically-focused content learning as a result of too little vocabulary knowledge (Başbek et al., 2014; Evans & Green, 2007; Chang, 2010). The absence of technical vocabulary has been a frequent reason for learners' difficulty to comprehend academic texts in ESL and EFL contexts.

While general academic vocabulary is widely studied (e.g., Coxhead, 2000; Davies & Gardner, 2013), research on the technical words of science is still scarce to date (Tong et al., 2014). A few researchers recommend focusing on general word lists because it is more generalizable. However, knowledge of content-specific vocabulary was a stronger predictor of science reading comprehension than

general academic vocabulary (Taboada, 2012). Especially EFL learners are going to have technical expression which should be explicitly taught in order for them to comprehend science texts and effectively perform tasks (Schmitt et al., 2011).

EMI has also developed in all sectors of education in Thailand (Phonlabutra, 2007). One such model is the English Program (EP), which teaches at least one or more non-English subjects in English. Nonetheless, students may experience difficulty with science subject matter, technical lexis and language processing (Björklund et al., 2006; Croyle & Chaturongakul, 2015; Yassin et al., 2009). It is important to address science vocabulary at the lower secondary level in order to enhance outcomes of scientific literacy.

Corpora have been widely applied to vocabulary teaching and learning (Nation & Waring, 1997). Pedagogic corpora, which pool together texts from educational domains, provide us with insight into the terminology of actual classrooms (Reppen, 2010). Textbooks in particular provide useful resources for building lexically-driven teaching materials (Willis, 2003). Wordlists from corpora contribute to identifying the priorities of vocabulary and planning for instruction (Dang, 2019).

A few recent studies have looked at lists of specialized vocabulary in domains such as science and engineering (Veenstra & Sato, 2018), zoology (Kruawong & Phoocharoensil, 2020), marine engineering (Durović, 2021), neurology (Li et al., 2022) and computer science (Uba et al., 2023). Language researchers in Thailand have been involved, too, with creating word lists for ESP (Laosrirattanachai & Ruangjaroon, 2021; Rungrueang et al., 2022); However, most of these lists are actually intended for university students and only modest research has been conducted to explore the vocabulary needs at the high school level (Coxhead, 2017).

Secondary school learners, especially in EFL settings, face significant challenges in acquiring technical vocabulary aligned with curriculum content (Humphrey, 2016). Science education becomes increasingly specialized from lower to upper secondary levels, making it necessary to prepare learners for subject-specific vocabulary demands (Coxhead, 2017).

To address this gap, the current study develops a set of English Science Textbook Word Lists (ESTWLs) based on a self-compiled pedagogic corpus. The word lists, Physical Science Word List (PhySciWL), Biological Science Word List (BioSciWL), and Earth & Space Science Word List (E&SSciWL), were derived from science textbooks used in Thai lower secondary English Programs. The goal is to identify frequently occurring technical terms that can support teachers in instruction and assist students in mastering science content through English.

2. Literature Review

2.1 Corpus Linguistics and Vocabulary Teaching

Corpus studies provide essential data for understanding specialist vocabulary. Corpora can generate frequency lists and identify lexical differences across disciplines (Hunston, 2022, leading to the development of subject-specific word lists (Coxhead, 2017). Frequency information helps teachers define vocabulary learning goals aligned with learners' needs (Nation & Waring, 1997).

Nation (2001) classifies word lists into four categories. The first is high-frequency words, such as those in West's (1953) General Service List (GSL), which accounts for over 80% of written text (Nation & Waring, 1997). Despite criticisms of being outdated (Richards, 1976), the GSL remains influential and covers 76% of Coxhead's Academic Word Lists (Watson Todd, 2017). Webb and Nation (2012) therefore recommend that lower-proficiency EFL learners initially focus on these high-frequency items.

The second category is academic vocabulary, comprising words like *analyze* and *phenomena* that are essential in tertiary contexts (Farrell, 1990). Coxhead's (2000) Academic Word List (AWL) 570-word families occurring across arts, science, law, and commerce was built from a 3.5-million-word corpus with a frequency threshold of 100 tokens.

The third category includes technical or discipline-specific terms that distinguish fields (Martin, 1976) and appear rarely outside them (Nation & Anthony, 2016) such as *concave*, *refraction*, and *electroscope* in science. The fourth category is low-frequency words, which occur infrequently and do not fit the previous groups.

2.2 The construction of science-related word lists

Corpus research has led to the creation of domain-specific science word lists, including the Korean Science Textbook Corpus (K-STeC) by Yun et al. (2018), the Middle School Vocabulary Lists (MSVL) developed by Greene and Coxhead (2015), and locally developed Thai lists including the Science Academic Word List (SAWL) by It-ngam and Phoocharoensil (2019) and the Zoology Word List (ZAWL) by Kruawong and Phoocharoensil (2020).

Although methodologically strong, many lists rely heavily on tertiary-level academic texts, limiting their relevance for younger learners. Using journal articles instead of classroom materials reduces applicability for EFL secondary students. A research gap exists because no word list has been developed specifically for secondary-level science EFL contexts aligned with classroom textbooks. Therefore, pedagogically relevant corpora such as lower secondary science textbooks are needed. This study responds to that gap by constructing and validating word lists derived from the authentic instructional materials used in Thailand's English Program.

2.3 Corpus-Based Word Lists Development

There are methodological decisions that need to be made to produce word lists from corpus such that they can be of use for teaching. Corpora need to be carefully selected: the larger corpora might not capture the language needs of particular learner groups, while the smaller pedagogic corpora are more likely to reflect curricular contexts, but may suffer from a lack of coverage (Nation & Webb 2011; Toriida 2016).

Common criteria for word list generation are frequency and range. However, frequency in itself is not a guarantee of pedagogical utility especially when the words are non-domain-specific (Schmitt & Schmitt, 2014). Learners in settings such as Thai secondary science may benefit from exposure to both HFWs and domain-specific vocabulary words to enhance content learning.

Lexical profiling labels words by groups like GSL, AWL and specialist terms. It can certainly be of use but if left unfiltered, may even serve to mask subtlety or technical meaning unless it is complemented by expert judgement providing pedagogical depth (Chung & Nation, 2004). However, previous research typically does not demonstrate how expert ratings apply to instructional decisions or student uptake.

To meet these challenges, the current study is a mixed-method analysis that combines frequency and range analyses based on corpora with expert validation. Unlike previous studies based on tertiary academic corpora, this study employs lower-secondary science textbooks to increase relevance, transparency and replicability for EFL settings. The aim is to create additional vocabulary lists from English textbooks in science which were already introduced in Thai lower secondary schools. These lists will be helpful for teachers to make syllabus, in classroom teaching and for students who are self-studying. Thus, the research question that drives this study is *What are the specialized words that commonly*

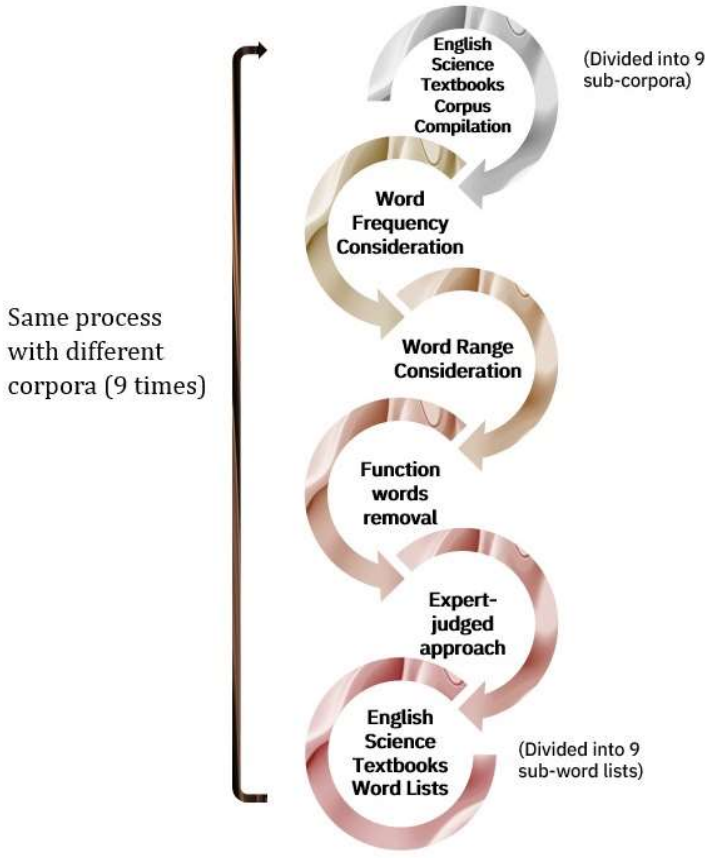
appear in science textbooks for lower secondary school level used in English program in Thailand?

3. Methodology

To identify and analyze the specialized vocabulary that frequently occurs in lower secondary science textbooks used in English Program schools in Thailand., Systematic procedure for designing, and validating English Science Textbook Word Lists is represented in Figure 1 below.

Figure 1

The Process of Creating and Validating English Science Textbook Word Lists



3.1 The English Science Textbook Corpus (ESTC) Compilation

Following Willis's (2003) recommendation that pedagogic corpora be derived from the actual materials students use, the ESTC was built from science

textbooks currently adopted in lower secondary English Programs (Grades 7–9). A total of 54 textbooks from nine publishers were selected to ensure curricular representativeness and reflect authentic learner input. These materials covered the three major science strands in the revised Thai Basic Education Curriculum B.E. 2560 (A.D. 2017): Physical Science, Biological Science, and Earth & Space Science (Office of the Basic Education Commission, 2017). Constructing the corpus from textbooks directly encountered by learners ensured that subsequent vocabulary analysis aligned with classroom practices and the linguistic demands of the national curriculum as proposed as in the table below.

Table 1

*Sub-Divided Corpora Based on the Topics**

ESTC	Sub-divided corpus	Topics
Physical Science Sub-Corpus (PhySciSC),	PS1	The properties of substances, components of substances, and the relationship between the properties and structures of substances, as well as intermolecular forces. The natural principles underlying changes in the states of matter, solubility, and chemical reactions.
	PS2	The essential role of forces in everyday life, the effects resulting from the application of forces on objects, and the characteristics of different types of motion and object properties.
	PS3	The importance of energy, energy transformation and transfer, and the natural principles related to the action of substances, waves, sound, light, and electricity.
Biological Science Sub-Corpus (BioSciSC)	BS1	The diversity of biological systems, the relationships between living and non-living things, and the interactions among different living organisms within ecological systems. The processes of energy transfer, transformation, and flow within ecosystems; the significance of natural resources, environmental issues, and their impacts.
	BS2	The properties of organisms, life the basic unit of life the movement through cells and the interrelationship between structure and function within organ systems in related groups of animals and humans which

ESTC	Sub-divided corpus	Topics
		work together. The relationship between structure and function within organ systems in plants which work together.
	BS3	the processes and extent of inheritance, material genetic, genetic changes in organisms, biodiversity and evolution of organisms that incorporate the use of knowledge for the organisms.
Earth and Space Science (E&SSciSC)	ES1	the components and relationships of the Earth's atmosphere, the changes on the weather and climate including the impacts on the organisms and environment. **Adding from the sub strand*
	ES2	The components and interrelationships of the Earth system, the processes of change inside the Earth and on the Earth's surface, natural disasters, and the processes involved in changes in atmospheric conditions and global climate.
	ES3	The components, characteristics, formation processes, and evolution of the universe, the solar system, planets, and celestial bodies, as well as interactions within the solar system that impact life.

**Adapted from Thailand's revised curriculum of B.E. 2560 (A.D., 2017) Learning sub strands*

The textbooks were converted into plain-text files (.txt) through direct extraction or manual conversion, depending on the original format. Non-linguistic elements, such as images, graphs, charts, and reference lists, were removed to maintain consistency and ensure that only analyzable text was included. The files were then processed using AntFileConverter (Anthony, 2022) to prepare them for corpus analysis software.

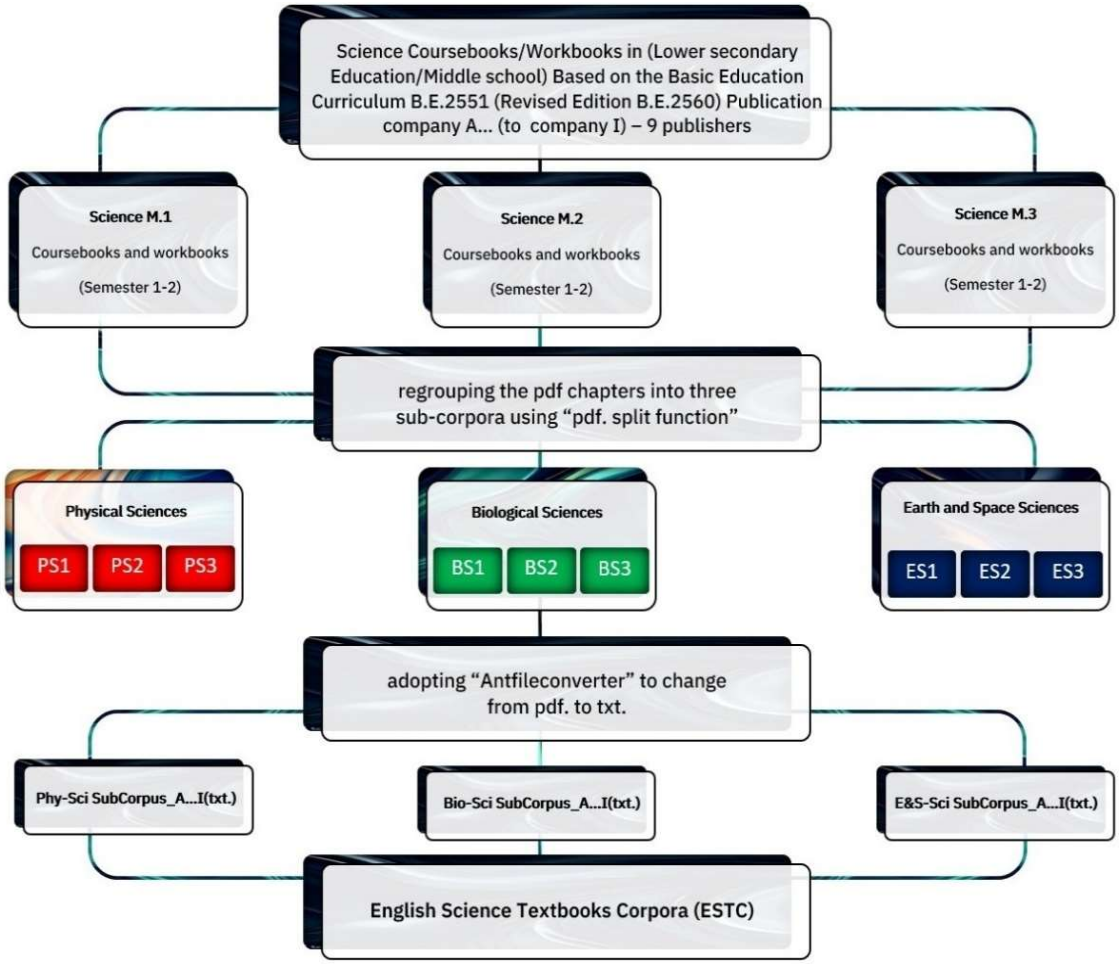
The finalized ESTC comprised three sub-corpora which reflects the curriculum strands comprising: 1. Physical Science Sub-Corpus (PhySciSC), 2. Biological Science Sub-Corpus (BioSciSC), and Earth & Space Science Sub-Corpus (E&SSciSC). Nine topic-specific sub-corpus were grouped based on the sequencing of textbook content. The division on topic-level was inspired to enhance pedagogical usability by letting teachers map the resulting word lists to

textbook units or lesson plans. Figure 2 illustrates the corpus compilation process, and Table 2 summarizes the overall structure and size of the ESTC.

E-ISSN: 2287-0024

Figure 2

The Process of Compiling English Science Textbook Corpora (ESTC)



To clarify the composition of each learning strand, Table 2 summarizes the sources and token totals that make up the ESTC.

Table 2

Sources and Tokens of English Science Textbook Corpus (ESTC)

Learning Strands	English Science Textbook Corpora	Token	Total
1	Physical Science Sub-Corpus (PhySciSC)	PS1 = 425,459 PS2 = 171,915 PS3 = 296,538	893,912

Learning Strands	English Science Textbook Corpora	Token	Total
2	Biological Science Sub-Corpus (BioSciSC)	BS1 = 130,302 BS2 = 434,508 BS3 = 103,363	695,173
3	Earth and Space Science (E&SSciSC)	ES1 = 92,661 ES2 = 285,318 ES3 = 109,325	487,304
			2,076,389

The Physical Science Sub-Corpus The size of (PhySciSC) is 893,912 words. BioSciSC contains 695,173 running words. In addition, the Earth and Space Science Sub-Corpus (E&SSciSC) consists of 487,304 running words. The ESTC consists of 2,076,389 running words. The words were all copied and pasted into Notepad software and saved as. TXT files.

3.2 Word Frequency Consideration

The first stage involved screening words by frequency following Coxhead (2000). Items were excluded if their relative frequency did not meet a minimum threshold determined by the size of each sub-corpus. For instance, in the Physical Science sub-corpus containing 425,459 tokens, the cut-off was set at 13 occurrences. This procedure ensures that only words appearing often enough to warrant instructional attention remain on the list (Lindquist, 2018). Examples of how these frequency thresholds were calculated for the Physical Science Sub-Word List (PhySciWL) are shown below.

PhySciWL01

$$100 \times \frac{425{,}459}{3{,}500{,}000} \approx 13$$

PhySciWL02

$$100 \times \frac{171{,}915}{3{,}500{,}000} \approx 5$$

PhySciWL03

$$100 \times \frac{296{,}538}{3{,}500{,}000} \approx 8$$

Table 3

The Cut-Off Point for the Frequency Count on Each Sub Corpus

English Science Textbook Word Lists (ESTWLs)	Sub topic-word list	Corpus Size	Cut-off point for the frequency count
Physical Science	PhySciWL01	42,5459	13
Sub-Word List	PhySciWL02	17,1915	5
(PhySciWL)	PhySciWL03	296,538	8
Biological Science	BioSciWL01	13,0302	4
Sub-Word List	BioSciWL02	434,508	12
(BioSciWL)	BioSciWL03	103,363	4
Earth and Space	E&SSciWL01	92,661	3
Science Sub-Word	E&SSciWL02	285,318	8
List (E&SSciWL)	E&SSciWL03	109,325	3

In short, for the PhySciWL Data, words found at least 5 times satisfied the Word Frequency requirement. For BioSciWL, words had to occur at least four times, and for E&SSciWL a minimum of three occurrences was met the Word Frequency request. The frequencies were calculated by the AntWordProfiler in this study.

3.3 Word Range Consideration

To ensure lexical consistency across materials, a word was required to appear in at least 50% of the publishers that is, in five out of nine textbook sets. This range criterion enhances the generalizability and stability of the final word lists by minimizing the influence of publisher-specific terminology (Coxhead, 2000).

3.4 Function Word Removal and Lexical Profiling

High-frequency grammatical items were removed using Nation's (2018) function word list. Because function words (e.g., articles, prepositions) are typically

mastered early in EFL learning and provide limited disciplinary meaning, excluding them ensures that the lists focus on semantically meaningful science vocabulary (Ward, 2009).

Words were also profiled to see if they were from GSL (West, 1953) and AWL (Coxhead, 2000) using AntWordProfiler (Anthony, 2022). Profiling is crucial in differentiating general, academic and discipline specific terms and to help teachers to know which items need focused instructionally. Research suggest that university students need to deal with the study of GSL vocabulary before they encounter academic words and that AWL is neglected in content classes where it is often thought students are already acquainted with its items (Farrell, 1990). Since both lists were relevant, they were included them in the current study.

3.5 Expert-Judged Appraoch

For pedagogical validity, words that were selected through corpus-based filtering were rated by 15 science teachers using a 4-point scale which was modified from Chung and Nation (2004). Participants were purposefully identified based on three criteria: (1) having three or more years of teaching lower secondary science subject in English Programs, (2) with professional qualification at undergraduate level in either Natural Sciences or Science Education to ensure reliable and informed judgements, and (3) who have two years' working experience using an EP textbook. Participants were recruited via professional teacher networks, with efforts to recruit teachers across a range of school types and region included.

4. Results

This study employed a mixed-methods design that integrated quantitative corpus analysis with qualitative expert evaluation to ensure both methodological rigor and pedagogical relevance. In the first phase, corpora were built, frequency and range analyses run and lexical profiling with AntWordProfiler to objectively discover lexicon frequent and widely spread in science textbooks. These findings

were reinforced during the qualitative phase by judging from experienced science teachers about relevance in context and their discipline as to the identified words, as well as qualitatively comparing the ESTWLs with MSVL for practical utility. Utilizing both corpus data and practitioner intuition, the research increased the rigour, validity and classroom utility of the ensuing word lists.

Before processing the data, the tokens were converted to types, with each corpus containing 70,060 types in PhySciSC, 70,002 types in BioSciSC, and 50,437 types in E&SSciSC. These were then examined using Lexical Frequency, Range, Lexical Profiling, and the expert-judge method. Table 4 shows the number of words in the English Science Textbook Word Lists.

Table 4

Number of Words in the English Science Textbook Word Lists After Processing

Research Procedure	**Physical Science Word Lists (PhySciWL)**			**Biological Science Word Lists (BioSciWL)**			**Earth&Space Science Word Lists (E&SSciWL)**		
	01	**02**	**03**	**01**	**02**	**03**	**01**	**02**	**03**
Tokens to types	30,998	14,589	24,473	17,432	34,852	17,718	15,163	20,371	14,903
Word Frequency and Range	218	240	205	219	219	217	214	184	231
Function word Removals	156	180	147	162	167	163	155	124	151
Expert-judged approach	120	125	120	62	60	76	75	47	75
Types	97	111	106	56	53	66	64	44	63
Total		314			175			171	
Types in total	**480-type word** *There are some words that can be found in more than one sub list.								

When analyzing the frequency criteria using the AntWordProfiler, there were approximately 180-240 words that passed both frequency and range in each sub-corpus. Most of such words are function words, for example, and, as, by, is, of, on, the, this, etc. To refine these lists further and ensure the remaining words were domain-specific, function words were removed using a reference function word list (FWL). Tables 5–7 show the updated breakdown of each topic area.

Table 5

Word Filtering Summary for PhySciWL01 (Physical Science)

Filtering Step	Word Count
Words passing frequency and range filters	218
Words removed based on function word list	62
Words forwarded to experts	156

Table 6

Word Filtering Summary for BioSciWL01 (Biological Science)

Filtering Step	Word Count
Words passing frequency and range filters	253
Words removed based on function word list	73
Words forwarded to experts	180

Table 7

Word Filtering Summary for E&SSciWL01 (Earth & Space Science)

Filtering Step	Word Count
Words passing frequency and range filters	197
Words removed based on function word list	58
Words forwarded to experts	139

As previously emphasized, Nation (2009) recommends that students with lower levels of English begin by focusing on high-frequency general and academic vocabulary, such as words from West's (1953) General Service List (GSL) and

Coxhead's (2000) AWL. Although these lists were not the direct reference lists for criterion of filtering in this study, their pedagogical implications are that it is crucial to determine EFL learners' essential vocabulary systematically. The present study optimizes this approach by focusing on technical terms commonly found in lower secondary science instruction materials and helping Thai EFL students in English Programs. The GSL and AWL continue to serve as foundational vocabulary resources widely adopted in textbooks and language curricula (Safari, 2020).

As the expert-judged approach was used in the study, each word list was then given to one of five professors in each sub-area. According to the teachers' input, 314 words in the PhySciWL, 175 words in the BioSciWL, and 171 words in the E&SSciWL were related to the physical science, biological science, and earth and space science topics, respectively. These were the final words, which were organized into word-types and separated into nine sub-lists. The ESTWLs Series has been completed and can be used for learning and teaching in lower secondary school classroom settings.

Evaluation is a critical stage in word list development, typically involving corpus comparisons and lexical coverage analysis to determine usefulness (Dang & Webb, 2016). Because assessing a list only against the corpus from which it was derived limits validity (Coxhead, 2000), many studies test their lists on external corpora (Brezina & Gablasova, 2015). As Coxhead (2017) notes, new lists should be compared with existing ones and supported with qualitative data, so the ESTWLs were also structurally compared with sample items from the MSVL (Greene & Coxhead, 2015). For this evaluation, three experienced science teachers one native English speaker, one bilingual Thai-English speaker, and one Thai speaker each with over five years of teaching experience, assessed the appropriateness of ESTWLs and MSVL items for middle school science instruction. Results are presented below.

Table 8

The Comparison of Percentage to Choose the Word in the List

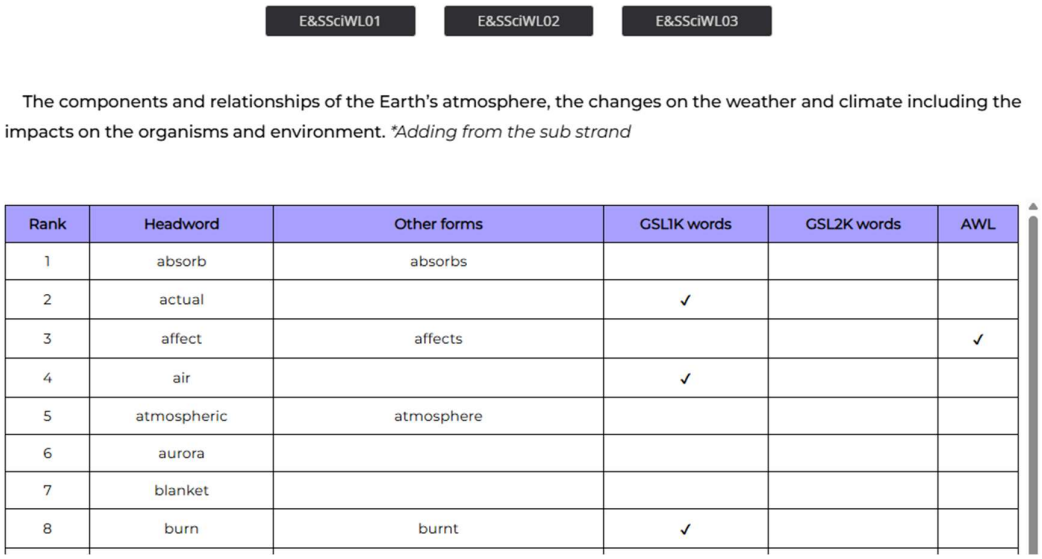
	% of choosing words in English Science Textbook Word Lists	% of choosing words in Middle School Vocabulary Lists – Science (Greene & Coxhead, 2015)
Expert 1	59.16%	57.70%
Expert 2	64.50%	63.60%
Expert 3	73.04%	70.20%

Based on the evaluation results in Table 7, all three experts agreed that the ESTWLs are more appropriate for Thai EFL lower secondary learners than the MSVL. A likely reason is that many ESTWLs items overlap with the high-frequency vocabulary in the General Service List (GSL), which is especially valuable for EFL learners because it enhances learning efficiency and broad comprehension (Nation, 2001).

The final result, **the 480 Words-Types of the ESTWLs** were published for academic and pedagogical use on a dedicated website (https://sites.google.com/view/estwl). The site provides three core lists, Physical Science, Biological Science, and Earth & Space Science, each divided into nine sub-lists, along with rank, headword, other forms. GSL and AWL items are clearly marked to support EFL learners and exam preparation. An example from one of the online tables is shown below.

Figure 3

Vocabulary Items in the Earth & Space Science Word List (E&SSciWL01) from https://sites.google.com/view/estwl



E&SSciWL01 E&SSciWL02 E&SSciWL03

The components and relationships of the Earth's atmosphere, the changes on the weather and climate including the impacts on the organisms and environment. **Adding from the sub strand*

Rank	Headword	Other forms	GSL1K words	GSL2K words	AWL
1	absorb	absorbs			
2	actual		✓		
3	affect	affects			✓
4	air		✓		
5	atmospheric	atmosphere			
6	aurora				
7	blanket				
8	burn	burnt	✓		

The table presents a ranked selection of vocabulary items identified from the sub-topic, *Earth & Space Science Word List.* The table provides several columns to guide interpretation: **Rank** indicates the word's frequency or importance within the corpus of science sub-topics; **Headword** is the main form used for counting and analysis; **Other Forms** lists additional morphological or inflected versions (e.g., absorbs, affects, atmosphere) that are intended to support recognition of word families; **GSL1K Words** identifies whether it appears in the first 1,000 most frequent words as well as GSL2K Words (i.e., second 1,000 words), marking mid frequency vocabulary that would be useful for learners moving into academic English; and **AWL** marks these items as belonging to Coxhead's Academic Word List, marking it as essential for academic reading and writing across disciplines.

5. Discussion

This study contributes to discipline-specific vocabulary research in several important ways. First, while earlier work such as Coxhead's (2000) Academic Word List and It-ngam and Phoocharoensil's (2019) Science Academic Word List

established the value of identifying high-frequency academic and technical vocabulary, these studies primarily focused on tertiary-level materials. The present research diverges by concentrating specifically on lower secondary science textbooks, directly addressing Coxhead's (2017) call for more vocabulary studies at the school level, where learners' linguistic and cognitive needs differ significantly from those of university students.

Second, this study contributes towards the field by using textbooks (instead of Academic Journal Articles) as the corpus, with a view to assuming that lexical items are truly reflected in authentic texts exposure for adolescents in classrooms. The method is a response to Humphrey's (2016) claim that word research in classrooms tends to overlook the practical needs of younger students. Through anchoring data analysis, the ESTWLs is more representative of the linguistic circumstances of Thai EFL students when using English as a medium for learning science.

Third, the study enhances methodological soundness by combining corpus-based frequency data with judgments of experts, thereby supporting hybrid approaches to epistemology as recommended by Chung and Nation (2004). These twin processes serve to increase the content validity of words by not only making sure these are statistically robust, but actually pedagogically sound within science teaching.

Fourth, while the ESTWLs conceptually overlap with Greene and Coxhead's (2015) MSVL, it differs by design explicitly for EFL learners in a bilingual or EMI context instead of a native-English-speaking environment. This is important, as vocabulary requirements and comprehension difficulties are far from identical in EMI environments (Evans & Green, 2007). ESTWLs thus addresses a shortage of an appropriate list, targeted to the Thai EFL students who study science in English.

Lastly, the results contribute to and go a step further in previous efforts to connect corpus linguistics with classroom instruction (Dang & Webb, 2016; Nation, 2016). Through the combination of analysis, alignment with the curriculum, and teacher expertise this study offers a replicable and pedagogically sound framework for constructing vocabulary lists at secondary level in EFL settings. The ESTWLs therefore functions not only as a class-room tool, but also as a methodological model for future studies in this area of applied linguistics.

6. Limitations of the Study

The ESTWLs, however practical and validated they may be, have a number of limitations that should be acknowledged. The corpus was derived from Thai English Program textbooks in compliance with the Thailand national Basic Education Curriculum, so the addition of materials from other EFL settings might enhance robustness. Additionally, the analysis was limited to single-word types; research should investigate multiword chunks and lexical bundles that are fundamental for scientific literacy. Also, some conceptually significant terms *(e.g. mesosphere, eclipse and photorespiration)* which were calculated to be below the frequency/range cutoff values, were not included in this list. It is therefore recommended that language and science teachers work together to complement the ESTWLs with low-frequency, content-specific vocabulary, ensuring that these words are incorporated in lists linked to curriculum glossaries or terminology banks.

7. Pedagogical Implications

The ESTWLs serve as a useful resource for lower secondary science classrooms, focusing on the high - frequency and conceptually significant vocabulary to use for English - medium instruction. The lists can be used by teachers to establish learning outcomes, monitor learners' lexical growth, adjust readings to levels of proficiency, gauge text difficulty and support program development (Davies & Gardner 2013) as well to develop additional resources to build on the scientific language. The ESTWLs help students to develop their

vocabulary and prepare for further study of science by providing learning in each topic across the three strands, presented in nine curriculum topics that focus on various objectives. The lists also constitute useful resources for curriculum developers. What is more, science teachers can exploit the ESTWLs to focus on top target words, while language teachers can create vocabulary-oriented activities supporting science learning. As high frequency word knowledge facilitates understanding in subject specific texts (Ball, Kelly & Clegg, 2015) the ESTWLs represents a strong inter-connection of language and science education which in turn fosters interdisciplinary teaching and enhanced learner achievement.

8. Recommendation for Further Research

The results of this study emphasize the significance of basing vocabulary teaching on curriculum-based corpora; therefore, future investigations should further investigate and validate content area word lists also in other domains and at other educational levels. It is suggested that researchers investigate the functions of the ESTWLs in authentic classroom environments, relating to students' reading comprehension, content knowledge and their long-term vocabulary retention. Other research can also consider the incorporation of corpus-based vocabulary tools within CLIL courses and investigate how often digital or AI-supported tools could be useful for personalized vocabulary learning. It is suggested that a community of practice of language teachers, science teachers, and curricular developers cooperatively work to integrate word lists for effective use in teaching, assessing and creating materials so as to create better EFL students access to science content and more equitable participation in English-medium science education.

9. Conclusion

This study developed the ESTWLs for Thai EFL lower secondary students using a systematically compiled corpus of 54 science textbooks totaling 2,076,389 words. After analyzing frequency and range with AntWordProfiler, removing function words, and conducting expert evaluation, 480 key vocabulary items were

identified across three strands, Physical Science, Biological Science, and Earth & Space Science each further divided by topic. The resulting lists support learners' vocabulary development in English-medium science classes and serve as a practical resource for subject and language teachers in CLIL settings.

10. About the Authors

Thana Kruawong, Ph.D., is a lecturer in the Division of English Language Teaching Faculty of Education, Kasetsart University, Bangkok, Thailand. He was a Ph.D. Candidate in English Language Teaching at Language Institute of Thammasat University, Bangkok, Thailand. He has research interests in Corpus Linguistics and Content and Language Integrated Learning approach.

Supakorn Phoocharoensil, Ph.D. is an Associate Professor of English, works at the Language Institute, Thammasat University. He is the Editor-in-Chief of the LEARN Journal. His research interests include Second Language Acquisition, Corpus Linguistics and the English collocations and Formulaic language.

11. References

Anthony, L. (2022). *AntWordProfiler* (Version 2.0.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Başıbek, N., Dolmacı, M., Cengiz, B. C., Bür, B., Dilek, Y., & Kara, B. (2014). Lecturers' perceptions of English medium instruction at engineering departments of higher education: A study on partial English medium instruction at some state universities in Turkey. *Procedia – Social and Behavioral Sciences*, 116, 1819–1825. https://doi.org/10.1016/j.sbspro.2014.01.477

Ball, P., Kelly, K., & Clegg, J. (2015). *Putting CLIL into practice*. Oxford University Press.

Björklund, S., Mård-Miettinen, K., Bergström, M., & Södergård, M. (Eds.). (2006). *Exploring dual-focused education: Integrating language and content for*

individual and societal needs. Centre for Immersion and Multilingualism, University of Vaasa.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22. https://doi.org/10.1093/applin/amt018

Chang, Y. Y. (2010). English-medium instruction for subject courses in tertiary education: Reactions from Taiwanese undergraduate students. *Taiwan International ESP Journal*, 2(1), 53–82. https://doi.org/10.6706/TIESPJ.2010.2.1.3

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*(2), 251–263. https://doi.org/10.1016/j.system.2003.11.008

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2017). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. Routledge.

Croyle, J., & Chaturongakul, P. (2015). *An investigation into student and teacher difficulties with English language science instruction at a late partial immersion English program in Northeastern Thailand* [Master's thesis, Thammasat University]. Thammasat University Digital Collections. https://digital.library.tu.ac.th/tu_dc/digital/api/DownloadDigitalFile/download/105316

Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). Routledge.

Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL – International Journal of Applied Linguistics*, *167*(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan

Davies, M., & Gardner, D. (2013). *A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists*. Routledge.

Durović, Z. (2021). Corpus linguistics methods for building ESP word lists, glossaries, and dictionaries: The example of a marine engineering word list. *Lexikos*, 31(1), 259–282. https://doi.org/10.5788/31-1-1647

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes*, *6*(1), 3–17. https://doi.org/10.1016/j.jeap.2006.11.005

Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary* (ED332551). ERIC. https://files.eric.ed.gov/fulltext/ED332551.pdf

Greene, J. W., & Coxhead, A. (2015). *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*. Brookes.

Humphrey, S. (2016). Academic literacies in the middle years: A framework for enhancing teacher knowledge and student achievement. In J. Arnold (Ed.), *Academic literacies for middle years* (pp. 43–59). Routledge.

Hunston, S. (2022). *Corpora in applied linguistics* (2nd ed.). Cambridge University Press.

It-ngam, T., & Phoocharoensil, S. (2019). The development of science academic word list. *Indonesian Journal of Applied Linguistics*, *8*(3), 657–667. https://doi.org/10.17509/ijal.v8i3.15269

Kaya, J., & Charkova, K. (2014). The most and least frequent vocabulary learning strategies of high school English language learners. *International Journal of English Language Education*, *2*(2), 122–141. https://doi.org/10.5296/ijele.v2i2.6019

Kruawong, T., & Phoocharoensil, S. (2020). Developing an English zoology academic word list: A corpus-based study. *Thoughts*, *2*, 63–78. https://doi.org/10.58837/CHULA.THTS.2020.2.3

Laosrirattanachai, P., & Ruangjaroon, S. (2021). Corpus-based creation of tourism, hotel, and airline business word lists. *LEARN Journal: Language Education and Acquisition Research Network*, *14*(1), 50–86. https://so04.tci-thaijo.org/index.php/LEARN/article/view/248677/168988

Li, M., Lee, Y., & Liu, D. (2022). Development of a neurology word list: A corpus-based study. *Language and Cognitive Science*, *6*(1). https://doi.org/10.35534/lcs.0601003

Lindquist, H. (2018). *Corpus linguistics and the description of English*. Edinburgh University Press.

Martin, A. V. (1976). Teaching academic vocabulary to foreign graduate students. *TESOL Quarterly*, *10*(1), 91–97. https://doi.org/10.2307/3585942

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2009). *Teaching ESL/EFL reading and writing* [eBook edition]. Routledge.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle.

Nation, P. (2018). Reading a whole book to learn vocabulary. *ITL – International Journal of Applied Linguistics*, 169(1), 30–43. https://doi.org/10.1075/itl.00005.nat

Nation, P., & Anthony, L. (2016). Measuring vocabulary size. In R. Routledge (Ed.), *Handbook of research in second language teaching and learning* (pp. 355–368). Routledge.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 6–19). Cambridge University Press.

Office of the Basic Education Commission. (2017). *Indicators and core science learning content: Revised edition 2017 under the Basic Education Core Curriculum B.E. 2551 (A.D. 2008)*. Ministry of Education, Thailand. https://drive.google.com/file/d/1_ALwE9xuCL3Fjet3Xl4gYjBj8p_1zLaA/view

Phonlabutra, K. (2007). *Learning in an English content-based program in a junior-high school in Thailand: A case study* [Doctoral dissertation, The

University of Arizona]. UA Campus Repository. http://hdl.handle.net/10150/194337

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89. https://doi.org/10.2307/3585941

Rose, H., & McKinley, J. (2022). May I see your credentials, please? Displays of pedagogical expertise by language teaching researchers. *The Modern Language Journal*, *106*(3), 528–546. https://doi.org/10.1111/modl.12794

Rungrueang, T., Boonprasert, P., Poempongsajaroen, S., & Laosrirattanachai, P. (2022). Corpus-based approach to generate a word list for food service. *THAITESOL Journal*, *35*(1), 57–76. https://so05.tci-thaijo.org/index.php/thaitesoljournal/article/view/258591/174027

Safari, M. (2020). Vocabulary lists for EAP and conversation students. *International Journal of Foreign Language Teaching and Research*, *8*(29), 131–163. https://sanad.iau.ir/fa/Journal/jfl/DownloadFile/990805

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

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Taboada, A. (2012). Relationships of general vocabulary, science vocabulary, and student questioning with science comprehension in students with varying levels of English proficiency. *Instructional Science*, *40*(1), 901–923. https://doi.org/10.1007/s11251-011-9196-z

Tong, F., Irby, B. J., Lara-Alecio, R., & Koch, J. (2014). Integrating Literacy and Science for English Language Learners: From Learning-to-Read to Reading-to-Learn. *The Journal of Educational Research*, *107*(5), 410–426. https://doi.org/10.1080/00220671.2013.833072

Toriida, M. C. (2016). Steps for creating a specialized corpus and developing an annotated frequency-based vocabulary list. *TESL Canada Journal*, *34*(1), 87–105. https://teslcanadajournal.ca/index.php/tesl/article/view/1257

Uba, S. Y., Irudayasamy, J., & Hankins, C. A. (2023). A new computer science academic word list. *World Journal of English Language*, *13*(5). https://doi.org/10.5430/wjel.v13n5p319

Veenstra, J., & Sato, Y. (2018). Creating an institution-specific science and engineering academic word list for university students. *Journal of Asia TEFL*, *15*(1), 148–166. https://doi.org/10.18823/asiatefl.2018.15.1.10.148

Ward, J. (2009). EAP reading and lexis for Thai engineering undergraduates. *Journal of English for Academic Purposes*, *8*(4), 294–301. https://doi.org/10.1016/j.jeap.2009.10.002

Watson Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, *45*, 31–39. https://doi.org/10.1016/j.esp.2016.08.003

Webb, S., & Nation, P. (2012). Teaching vocabulary. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green & Co.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge University Press.

Yassin, S. M., Marsh, D., Tek, O. E., & Ying, L. Y. (2009). Learners' perceptions towards the teaching of science through English in Malaysia: A quantitative analysis. *International CLIL Research Journal*, *1*(2), 54–69. https://jyx.jyu.fi/bitstreams/14d6ab9b-2a06-40cc-80a4-5c3a19c67f97/download

Yun, E., Kim, J., Nam, K., Song, H., Ok, C., Choi, J., & Park, Y. (2018). Building Korean science textbook corpus (K-STeC) for research of scientific language in education. *Journal of the Korean Association for Science Education*, *38*(4), 575–585. https://doi.org/10.14697/JKASE.2018.38.4.575

12. Appendix

The 480 Words -Types of the English Science Textbook Word List (ESTWLs)

The ESTWLs series are separated into 3 sub-word lists and 9 sub-topic lists. The Full ESTWLs is available at *https://sites.google.com/view/estwl/* To help readers interpret the table accurately, the following conventions are used.

- Words in *italic* indicate that additional morphological or derived forms appear in the full list on the website.
- Words that are underlined show cross-list occurrences, meaning they appear in more than one sub-list due to their relevance across multiple science topics.
- Words marked with an asterisk (*) are included in Coxhead's Academic Word List (AWL), highlighting their importance not only in science education but also in broader academic English contexts.

The 1st Sub - Physical Science Word Lists (PhySciWL) Separated into Three Sub-Topics

PhySciWL-1 (Topic: Matter/ Chemical Reaction) 97-Word Type

according	add	alcohol	apart	apparatus	appear
arrange	beaker	*become*	carbon	cause	*characteristic*
close	*collide*	common	compare	*complete*	composition
conclusion	consist*	contain	cook	copper	*Crystal*
depend	diesel	*difference*	dioxide	*dissolved*	due
even	*example*	*exist*	explain	*forces*	freely
further	*gas*	helium	ice	industrial	information
iron	kerosene	kinetic	lighter	liquid	make
material	matter	means	*mix*	model	*movement*
observation	oil	oxygen	paper	particle	pattern
physical*	*place*	position	*pour*	*produce*	*property*
recorded	rubber	sand	*scientist*	separate	shape
show	solid	solution	space	speed	*states*
steam	step	*stir*	stone	strong	sulphate
table	take	temperature	test	theory*	tiny
total	tube	*use*	vibrate	volume*	water
weak					

PhySciWL-2 (Topic: Force/Motion) 111-Word Type

able	acceleration	actual	area*	arrow	ask
better	calculate	consider	decide	*define**	density
description	determine	diagram	difference	*direction*	displacement*
distance	draw	*east*	empirical*	encounter*	*end*
energy*	*event*	evidence*	example	explain	field
figure	final	*fly*	focus*	force	*give*
go	*great*	group	*head*	*heat*	height
hold	hot	initial*	*instrument*	Kevin	know
later	left	length	let	line	list
locate*	magnetic	magnitude	mass	*measure*	metre
momentum	motion	*movement*	need	north	object
only	path	physical*	*point*	position	proportional*
quantity	reach	represent	rest	right	rule
scalar	*scale*	science	*show*	size	smart
solution	south	specific*	speed	*start*	station
stop	straight	study	*swim*	table	temperature
then	therefore	time	torque	total	*travel*
understand	unit	*use*	vector	velocity	volume*
water	weight	west			

PhySciWL-3 (Topic: Energy/Waves) 106-Word Type

able	adjacent*	amplitude	apparatus	axis	bar
*benefit**	bottom	*broadcast*	*carry*	circular	common
complete	*component**	compression	conclusion*	create*	crest
data*	describe	*diagram*	direction	displacement*	distance
draw	electromagnetic	energy*	equivalent*	*example*	explain
fall	*fix*	floor	focus*	formation	*form*
frequency	graph	hertz	highest	horizontal	important
include	know	light	longitudinal	lowest	*make*
material	maximum*	*measure*	medium*	model	motion
movement	observation	often	parallel*	part	*particle*
path	period*	perpendicular	phase*	*place*	plane
point	*position*	procedure*	*produce*	radio	rarefaction
represent	same	science	short	Show*	simple
sound	source*	spectrum	spread	spring	step

straight	stretched	surface	system	table	televisions
term	time	*touch*	toward	transverse	*travel*
trough	type	*use*	various	vertical	*vibration*
wall	water	*wave*	wavelength		

The 2nd Sub - Biological Science Word Lists (BioSciWL) Separated into Three Sub-Topics

BioSciWL-1 (Topic: Ecosystem/Biodiversity) 56-Word Type

abiotic	*add*	*air*	animal	arrange	bacteria
biotic	breed	cause	chain	community*	complete
consumer*	*decompose*	*difference*	earth	*ecology*	*ecosystem*
energy*	environment*	*example*	fertile	flowers	food
forest	frog	*grassland*	*habitats*	insect	*interaction**
living	marine	microorganism	microscope	*mineral*	natural
nectar	nutrients	offspring	photosynthesis	*plant**	*pollen*
population	pond	*producer*	relationship	*rocky*	*scientific*
shelter	snail	soil	species	stream	transfer*
water	web				

BioSciWL-2 (Topic: Cells/Organism structure and Functions) 53-Word Type

apparatus	base	blood	body	bone	brain
cell	cheek	coarse	compare	compound*	cork
cytoplasm	describe	diameter	diaphragm	diffusion	division
energy*	excretion	eye	*form*	*function**	growth
human	knob	lens	magnify	marrow	material
membrane	methylene	*microscopes*	*mitochondrion*	multicellular	nucleus
observe	*organism*	osmosis	permeable	plant	procedure*
process*	*produce*	respiration	*shape*	slide	solution
stage	*structure**	*substance*	tube	unit	

BioSciWL-3 (Topic: Inheritance/ Heredity) 66-Word Type

allele	baby	biodiversity	blood	*body*	cause
cell	centromere	characteristic	chromatin	chromosome	coiled
cytoplasm	diploid	*disease*	diversity*	*division*	DNA
dominant*	*ecosystem*	effect	environment*	factor*	father

female	fertilization	form	gain	generation*	*gene*
genetics	genotype	heredity	*human*	individual*	information
inheritance	*instruction**	living	male	meiosis	membrane
Mendel	mitosis	model	modify*	mother	nucleus
offspring	order	*organism*	*parent*	*pea*	people
phenotype	physical*	ratio*	recessive	relationship	resemble
scientific	species	sperm	*structure**	trait	zygote

The 3rd - Earth & Space Science Word Lists (E&SSciWL) Separated into Three Sub-Topics

E&SSciWL-1 (Topic- Atmosphere/ Weather/ Climate) 64-Word Type

absorb	actual	affect*	air	atmospheric	aurora
blanket	burn	cause	climate	cloud	cold
compare	composition	contain	cyclone	depend	direction
earth	entire	environment*	exosphere	extend	force
forecast	formation	future	*gas*	gravity	ground
harmful	humidity	*impact**	layer*	light	merge
mesosphere	*meteor*	*occur**	ozone	period*	precipitation
prediction*	pressure	rays	*refer*	satellites	*start*
state	stratosphere	sun	*temperature*	thermosphere	thunderstorm
time	tropical	troposphere	ultraviolet	vapor	*vary**
visible*	*weather*	wet	wind		

E&SSciWL-2 (Topic- Earth/ Geoscience) 44-Word Type

affect*	air	clay	climate	combination	*component**
composition	contain	*decompose*	describe	*factor**	fall
formation	found	growth	humus	identify*	irrigation
marble	material	matter	mineral	mixture	*movement*
organic	organism	*particle*	*pebble*	planet	plant
pressure	process*	*property*	*rock*	sand	shape
silt	*soil*	space	*structure**	texture	*vary**
water	*weather*				

E&SSciWL-3 (Topic- Universe/ Space Science) 63-Word Type

angle	appearance	asteroid	*attracting*	autumn	axis
belt	circle	comet	constant*	create*	daytime
December	*distance*	dwarf	earth	effect*	equation*
equator	explain	exploration	focus*	force	gravitation
gravity	hemisphere	Kuiper	mass	moon	moonrise
moonset	motion	Neptune	newton	nighttime	north
notice	object	occurrence*	*orbit*	phase*	phenomena*
planet	Pluto	proportional*	revolution*	ring	rotate
season	south	space	spring	*star*	*state*
summer	sun	sunlight	system	technology*	tidal
tilt	wane	winter			