การสำรวจคำศัพท์ภาษาอังกฤษสำหรับห้องเรียนเฉพาะด้าน: เพื่อปรับใช้สำหรับสาขาพลศึกษาและวิทยาศาสตร์การกีฬา

SPORT NEWS VOCABULARY INVESTIGATION FOR ESP CLASSROOMS: IMPLICATIONS FOR PHYSICAL EDUCATION AND SPORT SCIENCE

ผ่องอำไพ คงเจริญ*

Pong-ampai Kongcharoen*

ภาควิชาภาษาต่างประเทศ คณะมนุษยศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

Department of Foreign Languages, Faculty of Humanities, Kasetsart University.

**Corresponding author, e-mail: pongampai.k@ku.th*

Received: *5 May 2021;* ***Revised:*** *2 September 2021;* ***Accepted:*** *30 September 2021*

บทคัดย่อ

งานวิจัยชิ้นนี้ศึกษาการใช้คำศัพท์ในข่าวกีฬาเพื่อหาคำศัพท์ที่เป็นประโยชน์แก่นิสิตสาขาพลศึกษาและวิทยาศาสตร์การกีฬา โดยใช้การรวบรวมคลังคำศัพท์ข่าวกีฬา คลังคำศัพท์ข่าวกีฬาที่ใช้ในการศึกษาครั้งนี้ประกอบด้วย ข่าวกีฬาที่มาจากหนังสือพิมพ์ภาษาอังกฤษ 12 สำนักพิมพ์ รวมเป็นจำนวนคำทั้งสิ้น 3,571,501 คำ โดยเก็บข้อมูลข่าวกีฬาในรูปแบบอิเล็กทรอนิกส์ทุกวันตั้งแต่วันที่ 1 ตุลาคม 2561 – 30 พฤศจิกายน 2561 สำนักพิมพ์ที่ใช้ประกอบด้วย Sporting News, The Independent Sport, ABC News, BBC Sports, CNN, ESPN, Fox Sport USA, Fox Sport, Guardian Sport, Sky Sport, The Daily Telegraph และ US News หลักเกณฑ์ในการเลือกสำนักพิมพ์ คือ ความสามารถในการเข้าถึงหนังสือพิมพ์ในรูปแบบอิเล็กทรอนิกส์และไม่เสียค่าใช้จ่าย นอกจากนี้หนังสือพิมพ์ที่เลือกเป็นหนังสือพิมพ์ที่มาจากประเทศที่ใช้ภาษาอังกฤษเป็นภาษาที่ 1 จำนวน 3 ประเทศ ได้แก่ สหราชอาณาจักร สหรัฐอเมริกา และออสเตรเลีย ผลที่ได้พบว่า นอกจากคำศัพท์ที่พบโดยทั่วไปและคำศัพท์วิชาการแล้ว ยังมีคำศัพท์อีก 3,043 กลุ่มคำ ที่นิสิตควรรู้เพื่อช่วยให้เข้าใจข่าวกีฬาได้ดี คำศัพท์ 3,043 กลุ่มคำ สามารถแบ่งออกเป็น 3 ประเภท ได้แก่ คำที่มีความหมายทั่วไป แต่มีความถี่ไม่มากพอที่จะจัดอยู่ใน General Service List คำที่มีความหมายทั่วไปและมีความหมายเฉพาะด้านกีฬา และคำที่มีความหมายเฉพาะด้านกีฬาเท่านั้น

คำสำคัญ: คลังคำศัพท์ การศึกษาคำศัพท์ ข่าวกีฬา

Abstract

This paper investigates sports news text in order to ascertain a useful word list for physical education and sport science students. A Sports News Corpus was compiled for this study. The Sports News Corpus consists of 12 newspapers written in English with a total 3,571,501 running words. This sports news was collected every day in their electronic version from 1 October 2018 – 30 November

2018. The 12 newspaper agencies included: Sporting News, The Independent Sport, ABC News, BBC Sports, CNN, ESPN, Fox Sport USA, Fox Sport, Guardian Sport, Sky Sport, The Daily Telegraph, and US News. The newspapers were chosen based on their electronic availability and open access. The selected newspapers originated in the United Kingdom, the United States of America, and Australia. Apart from GSL and AWL, the results reveal that students should be familiar with 3,043 word families in order to comprehend sports news. The 3,043 word families, can be categorized into 3 types which are words with common meanings but do not appear frequently in general contexts, words that appear common but include specialized meanings, and words with specialized meanings.

Keywords: Corpus Linguistics, Word List, Sports News

Introduction

Students who are beginning university level face many obstacles and adjustments, and many of these difficulties involve learning to use language in new ways [1]. This is equally true in Thailand where English is a required subject in all curricula at university level. However, what can be observed as a university level English instructor for many years is that low-proficiency students tend to have more problems learning a new language than high-proficiency ones. Thus, finding the most effective way to teaching English to low-proficiency students is always challenging.

From the researcher's own experience in teaching students from a physical education department in a Bangkok university for almost 10 years, it can be noticed that physical education students have specific difficulties with vocabulary when reading English. In an attempt to help these students, the Basic Physical Education and Sport Science English word list [2] was created. However, as this word list was created from physical education and sport science research articles, the non-GSL and non-AWL word families are still difficult for this group of students. The Basic Physical Education and Sport Science English word list explores the use of GSL, AWL, and non-GSL and non-AWL English content words in physical education and sport science research articles. However, many of the non-GSL and non-AWL content words are medical and research terms. Therefore, this word list seems to be too difficult for this group of students to make use of. As previously mentioned, sports news is taken into consideration when planning and compiling another useful corpus for this group of students.

Learners of English as a Foreign Language (EFL) usually find that vocabulary limitation is one burden in academic reading [3]. It is clear that vocabulary is important in everyday language use, and learners with a larger vocabulary bank in English usually have a higher level of proficiency [4]. So, focusing on vocabulary learning can help these students in learning new languages. However, which vocabulary should be focused on? This is still a challenging question for all educators and teachers when looking for the appropriate vocabulary for their students.

Students from the physical education department at a Bangkok university have faced difficulty focusing on the appropriate vocabulary when learning English. After teaching these students for 10 years, it can be observed that they tend to feel uneasy when taking English courses. So, elevating their motivation

to learn English is always challenging. Corpora of specialized texts seems to be the answer to offering indications about key lexical, grammatical, or textual issues to deal with in ESP classes [5]. In line with Gavioli [5], creating another specialized corpus for this group of students will be beneficial in vocabulary learning. Identifying the appropriate vocabulary that students should focus on when learning English would shed light on how English reading can be taught to students who have low English proficiency, especially students from the Physical Education Department. This research therefore aims to investigate sports news vocabulary to create a useful English word list for physical education students.

Though Kongcharoen's [2] Basic Physical Education and Sport Science English Word List was created to explore the use of GSL, AWL, and non-GSL and non-AWL content words in physical education and sport science research articles, the Basic Physical Education and Sport Science English Word List contains many technical terms in health science and research methodology. However, first- and second-year students are obliged to take foundation English classes, as the research methodology vocabulary appears to be challenging and unfamiliar to this group of students. An analysis of the two word lists, there are only 97 word families from 3,043 word families which co-occur with the Basic Physical Education and Sport Science English word list. Here are some examples of words which co-occur with Basic Physical Education and Sport Science English Word List; abdominal, accelerate, additionally, aforementioned, align, basketball, breakdown, campus, defense, disability, drain, elevate, endurance, feedback, hamstring, hockey, leisure, mobility, overweight, pace, quadriceps, rugby, ski, soccer, surgeon, trainer, trauma.

Before compiling another specialized corpus, a more interesting and relevant reading source needed to be identified. A survey of the readings they paid most attention to in English showed that sports news was always the most popular. Therefore, a Sports News Corpus needed to be collected to provide these students with the appropriate vocabulary. In the year the corpus was compiled, there was a big sport event which these students were interested in, the 2018 Football Word Cup. The World Cup refers to an international football competition that takes place every four years [6].

Literature Review

Text Coverage

Reading for basic comprehension normally refers to the comprehension of a subset of individual ideas which relate to the thematic content or the main idea of the text [7]. In reading comprehension, vocabulary always plays an important role. There are many ways to decide how many words a learner of English as a second or foreign language needs to know in order to comprehensibly read a text without external support [8]. The examination of the relationship between text coverage and reading comprehension for non-native speakers of English with a fiction text revealed that with a text coverage of 80%, no one perceived adequate comprehension. With a text coverage of 90%, a small group gained adequate comprehension. With a text coverage of 95%, a few more gained adequate comprehension. At 100% coverage, the majority gained adequate comprehension. From the statements above, 95% seems to be a promising amount of text coverage in order to gain adequate comprehension [9].

Laufer [10] claims that 95% coverage is needed for reasonable comprehension of a text. Congruent with Laufer [10], Webb and Rodgers [11] suggest that 95% coverage is necessary to have

adequate understanding of television programs. Schmitt et al. [12] have worked on text coverage and reading comprehension and have suggested that 98% is more promising. However, the same paper points out that 95% text coverage seems to work as well if the comprehension rate is set at 60%, whereas 98% coverage is most suggested for academic reading and academic courses. Besides text coverage, background knowledge is another important key for reading comprehension. Hence, having a specialized corpus with similar content to students' background knowledge should be beneficial and should assist in students 'reading comprehension.

Besides Webb and Rodgers [11], there are many scholars who have worked on the vocabulary coverage of various texts [13, 10, 14-15], but few scholars have worked on the vocabulary coverage of sports news. Therefore, the aim of this study is to determine the lexical coverage of sports news that students need to know in order to comprehensibly read sports news. Coverage here refers to " the percentage of known words in discourse and is a valuable measure because it may indicate the vocabulary size necessary to understand a text as well as to incidentally learn words in the text" [11].

Having a larger vocabulary size will provide the learner with a more extensive database to guess the meaning of unknown words or behavior of newly learned words [7]. Many scholars have worked on how much text coverage is required to comprehend any texts when reading. For example, Nation's [8] research has set out to see how sizeable receptive vocabulary is needed for typical language use such as reading a novel, reading a newspaper, watching a movie, and taking part in a conversation. " The English language contains approximately 114,000 word families, excluding proper names" [16]. Educated monolingual native speakers know approximately 20,000 word families [16-17]. However, "acquiring 20,000 word families is a very challenging goal for second or foreign language learners" [14].

A particular use of word lists estimates the coverage of a loss over a text or corpus, and the purpose of this estimate is to find out how many words are required to understand a text [18]. Many scholars examine word coverage in some particular texts. For instance, Hirsh and Nation [13] found that the 2,000 most common word families would provide 90% coverage of teenage novels' corpus. However, Laufer [10] suggests that 95% coverage is needed for comprehension. Zeeland and Schmitt [15] also found that 95% coverage is sufficient for listening comprehension. Furthermore, Hu and Nation [9] examined the relationship between text coverage and reading comprehension for non-native speakers of English with a fiction text. They discovered that with a text coverage of 80%, no one gained adequate comprehension. With a text coverage of 90%, a small minority gained adequate comprehension. With a text coverage of 95%, a few more gained adequate comprehensions, but they were still a small minority. At 100% coverage, most gained adequate comprehension. However, 100% coverage seems to be impossible when 95% coverage is more reasonable to deal with. Thus, this study is conducted to create a useful word list from sports news after reaching 95% coverage from Sports News corpus.

Why are Vocabulary and Specialized Corpora Important?

Biber [1] states that vocabulary used in university contexts is an integral part of the development of effective teaching materials and approaches. Vocabulary has also played an essential role in putting corpus-based research into the academic language [19]. There are many available corpora to be used at

hand such as British National Corpus (BNC) [20], Corpus of Contemporary American English (COCA) [21], The Corpus of Historical American English (COHA) [22], etc. However, corpora of specialized texts seem to be beneficial material in providing indications of key lexical, grammatical, or textual issues to deal with in ESP classes [5]. This is because general English corpora cannot provide information for specialized language, while a smaller, specialized corpora can do the trick [5]. General corpora are more suitable for studying the structure and the use of language. In contrast, specialized corpora which focus on specific genres are better when using for exploring language in specific setting [23]. In order to focus on a more specific range of topics and text types, specialized corpora are the most suitable tool for a specific language [5]. Therefore, a Sports News Corpus was compiled in order to provide a useful lexical profile and create a focused word list for students who are in the field of physical education and sport science.

Coxhead [18] states that foreign language learners need a large vocabulary to deal with their studies in an academic or professional environment. Students need to know the vocabulary of their field well enough to function as professionals. Moreover, Adamson [24] states that vocabulary knowledge is often taken into account to measure students' language learning progress. Therefore, setting vocabulary goals for language courses, guiding learners in their independent study, and informing course designers in deciding on course materials, selecting text, and developing learning activities is always a critical element for creating specialized corpora [25]. In China, vocabulary teaching and learning are seen as the most important elements of language teaching [14]. In Thailand, we might need to explore further, in what aspects, vocabulary teaching can be important.

According to Yu and Renandya's [14] research, words are not created equal: some are more frequently used, and therefore more useful. As a result, exploring the more useful words in specific fields needs to be conducted. Yu and Renandya [14] also add that the words which appear more frequently and have large proportion in daily use should receive a priority in second language classrooms. Hence, West's [26] famous GSL word families are included in this study as the essential words students should know. Understanding what aspects of word knowledge are important for learners and what materials design elements can help learning are also points for learning specialized vocabulary [18]. So, in this study, a sports news corpus is used as a specialized corpus to determine the specialized words that students can focus on. Coxhead [18] also adds that specialized vocabulary is important because knowledge of the vocabulary of a field is tightly related to content knowledge of the discipline. Flowerdew [27] also states that "specialized smaller corpora offer more advantages than general corpora from a methodological perspective because they provide more contextual information". Personally, I agree with this point; specialized corpora can provide ways to explore specialized vocabulary.

There have been quite a variety of news corpora such as CC-News-En: A Large English News Corpus [28], A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception [29], Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention [30], LOTUS-BN: A Thai Broadcast News Corpus [31], etc. For sports news, Al-Khawaldeh et al. [32] explored Arabic sports journalistic texts to find discourse markers. However, an English sports news word list for language classrooms has not been conducted. In the hope of creating another useful

word list for physical education students, a sports news corpus is needed as a specialized corpus for specific group of students.

High Frequency Words

Corpora can assist in creating important vocabulary lists which can be used as a guide for course designing and materials preparing [33]. Well-known word lists include West' s [26] General Service List (GSL) and Coxhead' s [34] Academic Word List (AWL). However, Gardner and Davie' s Academic Vocabulary List (AVL) [35] created in 2013 is more reliable as a core academic word list because the AVL consists of more tokens than AWL with a corpus of 120 million words, which means it can be seen as a more reliable word list for academics. However, AVL is not as popular as AWL among language teachers. As AWL is often the most utilized, it makes AWL worth exploring so AWL coverage is also investigated in this study.

Besides using GSL and AWL, Moon [36] suggests that all corpora show which words are used in their constituent texts and how frequently they appear. The widespread available corpora and the ease of automated word counts seem to offer every teacher the possibility of creating vocabulary lists tailored to their students' needs [33]. Studies like these are based on word lists that identify the most critical words in various fields [1].

Any text will contain high-frequency items [37], and these items make up the majority of running words in texts. Besides that, high-frequency words can take on specialized meanings in particular contexts [18]. The most frequent words in any corpus are grammatical or function words, but working down frequency lists can reveal key items in the genre, enabling teachers to identify and teach basic items in their classes [38]. There are many reasons why frequency words are vital for students to learn; one reason is that the more frequent a word is, the more critical it is to learn [33]. Priority is given to describing the most typical uses of the commonest words on the assumption that if something has happened often enough in the past, then it is likely to happen often in the future [38]. Hence, frequency lists can be valid documents for lexicographers and language syllabus and materials designers to create a more practical material for their students [39].

A specialized approach to making word lists would begin with no existing word list to represent existing knowledge. West's [26] GSL and Coxhead's [34] AWL are examples of important academic word lists. 2000 most frequent word families in GSL were created by West in 1953 with a 5-million-word corpus. In compiling the General Service List [26], "West and his colleagues considered a number of criteria other than frequency: less frequent works were included if they could be used to convey a range of important concepts; words were excluded if a synonym was available; words needed to be stylistically neutral; and intensive emotional words were not included" [33].

Aside from GSL, AWL was created by Coxhead in 2000 with a corpus of 3.5 million words. In order to create AWL, Coxhead first screened GSL out of the list. Then she looked at frequency and range by including any words which counted more than 100 times from the whole corpus and 10 times from each sub-corpus.

A clear advantage of this approach is that it fits well with particular groups of learners regarding their language learning needs [18]. Hyland [38] suggests that the advantage of using frequency word lists in EAP is the construction of vocabulary lists such as the Academic Word List [34]. The top-rated word lists are West's [26] GSL which contains 2,000 common words, and Coxhead's [34] AWL, which consists of 570-word families. Besides these two popular word lists, there are many more word lists for specialized contexts, such as Tourism, Hotel, and Airline Business Word List [40], Zoology Academic Word List [41], The Word List of Hospitality Service [42], Science Academic Word List [43], Environmental Academic Word List [44], Nursing Academic Word List [25], Medical Academic Word List [45], and many more. However, few scholars explore sports news or create word lists from sports news. Even though the Basic Physical Education and Sport Science English Word List [2] was created for physical education students, that word list was based on research articles. In an attempt to find the more relevant words for Physical Education students, the sports news corpus was compiled as a source of information that students could employ often. Coxhead [18] also suggests that research needs to ensure that corpora are looked at more closely, and word lists are used to guide decisions on which words to look at first because of their frequency and range.

Objectives

This research aims to investigate sports news vocabulary to create a useful English word list for physical education students. After reaching 95% coverage of sports news, what are non-GSL and non-AWL word families from a sports news corpus that students need to pay attention to?

Methods

Corpus Compilation

As Coxhead [18] claims, a specialized approach to making word lists needs to be conducted in order to provide word lists representing existing knowledge. In the hope of finding the appropriate vocabulary to focus on for students from the Physical Education Department, a Sports News Corpus was created. The Sports News Corpus consisted of 12 newspapers written in English with a total of 3,571,501 running words. This sports news was collected every day in its electronic version from 1 October 2018 – 30 November 2018. The 12 newspaper agencies included: *Sporting News, The Independent Sport, ABC News, BBC Sports, CNN, ESPN, Fox Sport USA, Fox Sport, Guardian Sport, Sky Sport, The Daily Telegraph, and US News*. The newspapers were selected based on the availability in an electronic version and free access. Another criterion in selecting the newspapers was they had to originally come from 3 inner circle countries which were the United Kingdom, the United States of America, and Australia. They can be divided as follow:

United States: CNN, ESPN, Fox Sport USA, US News.

United Kingdom: BBC Sport, Guardian Sport, The Independent Sport, Sky Sport.

Australia: ABC News, Fox Sport, The Daily Telegraph, Sporting News.

In an attempt to create a variety of language used for the word list, the selected sports news needed to be from 3 different countries which are the United State of America, the United Kingdom, and Australia so that students could encounter a greater variety of words.

Table 1 The total number of running words from each newspaper.

Newspapers	Total running words	Number of lines	Number of types	Number of tokens
ABC News	216288	16945	14511	219568
BBC Sport	282,144	20,196	14,943	281,638
CNN	263,317	9,869	12,466	267,583
ESPN	340,169	9,897	16,214	343,915
Fox Sport USA	158,619	6,417	12,067	160,629
Fox Sport	400,428	39,583	19,043	404,384
Guardian Sport	360,133	7,551	22,587	372,103
Sky Sport	242,020	11,071	12,577	243,900
Sporting News	355,028	34,312	16,784	352,892
The Daily Telegraph	374,587	32,258	18,801	377,087
The Independent Sport	322,416	17,899	16,975	330,881
US News	216,615	7,499	11,902	216,921

Reference: AntWordProfiler software.

Data Process and Analysis

All the download samples were collected and converted into text forms (i.e. txt.) and AntWordProfiler was used to analyze the coverage of the texts. AntWordProfiler software was developed by Laurence Anthony and can be downloaded for free at https://www.laurenceanthony.net/software /antwordprofiler

To analyze the data, the frequency and distribution of word tokens and types in the corpus were determined using AntWordProfiler which is a freeware tool for profiling the vocabulary level and complexity of texts. AntWordProfiler shows the coverage of the words in Sports News Corpus by installing GSL and AWL as the based lists. After running the program, the words that placed in ‘out of based list’ were examined.

The 1st 1000 GSL, 2nd 1000 GSL, and AWL were used to show the coverage of the running words in the Sports News Corpus. At first, Coxhead’s [34] method was adopted. Since there were 3,571,501 running words in the Sports News Corpus, which was quite similar to Coxhead’s corpus in creating the AWL, Coxhead’s [34] frequency and dispersion were adopted. Any words which appeared more than 100 times in the whole Sports News corpus and appeared at least 10 times in each sub-corpus were included. However, from this frequency and dispersion, the coverage failed to reach 95%. Thus,

a more precise method was used in order to reach 95% coverage. The running words in each sub-corpus are not exactly the same amount, so $\frac{\text{the running words from each outlet x 100}}{\text{All running words}}$ was applied. For example, ABC News, $\frac{216{,}288 \text{ x } 100}{3{,}571{,}501}$ = 6.055, any words which occurred more than 6 times were included. After running the formula with each newspaper, the number of occurrences were as follows:

1. ABC News with total running words: 216,288 = 6.055 (Approximately 6 times).
2. BBC Sport with total running words: 282,144 = 7.899 (Approximately 8 times).
3. CNN with total running words: 263,317 = 7.372 (Approximately 7 times).
4. ESPN with total running words: 340,169 = 9.524 (Approximately 10 times).
5. Fox Sport USA with total running words: 158,619 = 4.441 (Approximately 4 times).
6. Fox Sport with total running words: 400,428 = 11.211 (Approximately 11 times).
7. Guardian Sport with total running words: 360,133 = 10.083 (Approximately 10 times).
8. Sky Sport with total running words: 242,020 = 6.776 (Approximately 7 times).
9. Sporting News with total running words: 355,028 = 9.940 (Approximately 10 times).
10. The Daily Telegraph with total running words: 374,587 = 10.488 (Approximately 10 times).
11. The Independent Sport with total running words: 322,416 = 9.027 (Approximately 9 times).
12. US News with total running words: 216,615 = 6.065 (Approximately 6 times).

Any words which appeared more than 100 times in the whole corpus and appeared as assigned number in each sub-corpus have been included. Again, this method failed to reach 95% coverage. With this method, the data could only reach 84.69% coverage in the Sports News Corpus. Hence, the criterion was adjusted again to include any words which had a frequency over 50 times for the whole corpus and appeared at least 6 times in each sub-corpus. Still, the coverage percentage did not reach 95%. This time, the text coverage only reached 90.78%. After several failed attempts, the criterion was continually adjusted until the coverage finally reached 95% with any words which appeared more than 9 times in the whole corpus. The text coverage from this criterion was 95.45%.

After reaching 95% text coverage in sports news corpus, the non-GSL and non-AWL words were carefully investigated. At this stage, the proper nouns and acronyms were eliminated. From 9,235 words, there were 5,148 proper nouns and acronyms. The total number of words after eliminating proper nouns and acronyms was 4,224 words which were later put into families.

Besides the non-GSL and non-AWL words, the AWL word families were considered in this study. AWL coverage in the Sports News corpus appears only 3.36%. It may be inferred that the AWL word families have not played a very important role in the Sports News Corpus.

After running all processes, the results are as follows:

Table 2 Results after reaching 95% coverage.

LEVEL	FILE	TOKEN	TOKEN%	CUMTOKEN%	TYPE	TYPE%	CUMTYPE%	GROUP	GROUP%	CUMGROUP%
1	1_gsl_1st_1000.txt	2509560	70.27	70.27	3694	6.48	6.48	998	1.97	1.97
2	2_gsl_2nd_1000.txt	188226	5.27	75.54	3030	5.31	11.79	964	1.90	3.87
3	3_awl_570.txt	119865	3.36	78.9	2225	3.90	15.69	563	1.11	4.98
4	out of based list_freq 9.txt	591129	16.55	95.45	9372	16.43	32.12	9372	18.52	23.5
0	-	162721	4.56	100.01	38719	67.88	100	38719	76.50	100
TOTAL:		3571501			57040			50616		

Reference: AntWordProfiler software.

When testing for text coverage, the proper nouns and acronyms were not screened out since this group of words contain quite high coverage with 5,148 words out of 9,372 words. However, students do not need to focus on these proper nouns as they can be changed depending on the situation in the news. As well as the proper nouns, acronyms represent the name of the organizations, places, etc and can be easily inferred.

Besides 2,000 GSL and AWL words, proper nouns, and acronyms, 4,224 words were extracted from this corpus and are put into word families. When putting all the words into word families, 3,043 word families were put in the list for students to make use of.

Results

The text coverage from the first attempt is shown below, i.e. any words which appear more than 100 times in the whole corpus and appear with assigned times in each sub-corpus.

Table 3 Text coverage from the first attempt.

LEVEL	FILE	TOKEN	TOKEN%	CUMTOKEN%	TYPE	TYPE%	CUMTYPE%	GROUP	GROUP%	CUMGROUP%
1	1_gsl_1st_1000.txt	2509560	70.27	70.27	3694	6.48	6.48	998	1.97	1.97
2	2_gsl_2nd_1000.txt	188226	5.27	75.54	3030	5.31	11.79	964	1.90	3.87
3	3_awl_570.txt	119865	3.36	78.9	2225	3.90	15.69	563	1.11	4.98
4	out of based list_freq 100.txt	206777	5.79	84.69	491	0.86	16.55	491	0.97	5.95
0	-	547073	15.32	100.01	47600	83.45	100	47600	94.04	99.99
TOTAL:		3571501			57040			50616		

Reference: AntWordProfiler software.

Although the first attempt failed in reaching 95% coverage, the words included in this stage are directly related to sports or sports news. Words such as *midfielder, golf, teammate, tennis, basketball, striker, athletes, goalkeeper, tackle, medal,* etc. appear at a high frequency here. However, the text coverage does not reach the comprehensible rate.

The text coverage from the second attempt is shown below, i.e. any words which appear more than 50 times in the whole corpus and at least 6 times in each sub-corpus.

Table 4 Text coverage from the second attempt.

LEVEL	FILE	TOKEN	TOKEN%	CUMTOKEN%	TYPE	TYPE%	CUMTYPE%	GROUP	GROUP%	CUMGROUP%
1	1_gsl_1st_1000.txt	2489999	69.72	69.72	3692	6.47	6.47	998	1.97	1.97
2	2_gsl_2nd_1000.txt	188226	5.27	74.99	3030	5.31	11.78	964	1.90	3.87
3	3_awl_570.txt	119865	3.36	78.35	2225	3.90	15.68	563	1.11	4.98
4	out of based list_freq 50.txt	443882	12.43	90.78	2093	3.67	19.35	2093	4.13	9.11
0	-	329529	9.23	100.01	46000	80.65	100	46000	90.88	99.99
TOTAL:		3571501			57040			50618		

Reference: AntWordProfiler software.

In this stage, the specifically related-to-the-field words appear. For example, *league, cricket, stadium, rugby, championship, pitch, penalty, tournament, innings, quarterback,* etc appear at a high frequency at this stage. Besides, names of athletes and places also appear with high frequency at this stage.

The text coverage from the last attempt is shown below, i.e., any words which appear more than 9 times in the whole corpus.

Table 5 Text coverage from the last attempt.

LEVEL	FILE	TOKEN	TOKEN%	CUMTOKEN%	TYPE	TYPE%	CUMTYPE%	GROUP	GROUP%	CUMGROUP%
1	1_gsl_1st_1000.txt	2509560	70.27	70.27	3694	6.48	6.48	998	1.97	1.97
2	2_gsl_2nd_1000.txt	188226	5.27	75.54	3030	5.31	11.79	964	1.90	3.87
3	3_awl_570.txt	119865	3.36	78.9	2225	3.90	15.69	563	1.11	4.98
4	out of based list_freq 9.txt	591129	16.55	95.45	9372	16.43	32.12	9372	18.52	23.5
0	-	162721	4.56	100.01	38719	67.88	100	38719	76.50	100
TOTAL:		3571501			57040			50616		

Reference: AntWordProfiler software.

Conclusions and Discussion

Conclusions

L2 learners need to know at least 4,000-5,000 word families to comprehensibly read any texts [8] and besides GSL and AWL which are common word lists that students should know, other words are as important according to their fields. In this research, when combining the 2000 GSL, 570 AWL, and 3,043 word families of non-GSL and non-AWL, the words add up to 5,613 word families which confirms the previous research from Nation [8]. According to the Macmillan Dictionary [6], the 3,043 word families can be categorized into 3 groups which are 1) words with common meanings but do not appear frequently in general contexts, 2) words that appear common but include specialized meanings, and 3) words with specialized meanings. This word list can be used as a useful word list for Physical Education and Sport Science students in their English class.

Though Kongcharoen's [2] Basic Physical Education and Sport Science English Word List was created, there are only 97 non-GSL and non-AWL word families which co-occur with the Sports News word list. This makes the Sports News word list useful as another glossary in English classes for this

group of students. The observation from this research is that there are varieties of words used in newspaper language which makes the important words in the newspapers, especially sports news, low in frequency count. Despite the low frequency in the Sports News Corpus, all the words play important roles in helping students to read sports news comprehensibly. Proper nouns and acronyms are other groups of words with a large quantity in sports news. Since the news reports are about situations, places, or things, names/proper nouns appear in a large quantity and at a high frequency. From the total of 9,235 words above, 5,148 words are proper nouns and acronyms. However, students do not need to focus on these words since they can be changed according to situations, places, and time. However, students can extrapolate the meaning of those acronyms.

Teachers can use Sports News word list to create teaching materials such as vocabulary projects or to store the list as glossary. If students are not yet familiar with GSL, teachers can start their course with materials that use GSL to get students familiar with the basic word families first, followed by AWL. The non-GSL and non-AWL can be the latest list that students can acquire since some words are very field-specific words. For example, *ballpark* which is specifically used with baseball, *bullpen* which means an exercise area for pitchers in baseball, *cox* which means someone who directs people in boat racing, etc. Whereas some words can be found expressing their normal meaning, when appearing in sports news, they can have another meaning. For example, *bat*, which students might be familiar with the meaning as being an animal rather than the stick used in baseball or cricket games. Thus, the Sports News word list can assist students when reading sports news and teachers can use sports news as teaching materials.

Discussion

In order to reach 95% coverage of sports news, the criterion was adjusted to include any words which appeared more than 9 times in the whole corpus. From the table above, it can be noticed that there are 591,129 running words which are non-GSL and non-AWL. From those 591,129 running words, there are 9,372 word types which can be classified as 5,148 proper nouns and acronyms and 4,224 content words. Words from the 4,224 non-GSL and non-AWL content words were categorized into word families which accounted for 3,043 word families of non-GSL and non-AWL from Sports News corpus.

Regarding AWL, AWL coverage in the corpus is quite low at 3.36%. From the whole corpus, there are 1,336 AWL words which met the criterion. These 1,336 words have been categorized into word families, which appear 486 word families from the total of 570 AWL word families. Even though the AWL coverage is quite low, the words appearing in sports news are quite varied. As 486 AWL word families appear in Sports News Corpus which accounted for 85.26% of all AWL, students also need to know AWL to read sports news comprehensibly.

From the table above, students then should know 2,000 GSL word families first, 570 AWL word families and 3,043 word families of non-GSL and non-AWL to be able to read sports news comprehensibly. Even though the frequency count is quite low, the words are still very important in reading sports news comprehensibly.

To answer the research question: What are non-GSL and non-AWL word families from Sports News corpus that students need to pay attention to?

From 3,043 word families of the non-GSL and non-AWL, there are many word families which are worth paying attention to. After consulting with an online dictionary [6], this word list can be categorized into 3 categories, [1] words with common meanings but do not appear frequently in general contexts, [2] words that appear common but include specialized meanings, and [3] words with specialized meanings.

1. Words with common meanings but do not appear frequently in general contexts such as *abdominal, abscess, abuse, activate, adjust, agility, arbitration, artifact*, etc. These words might not appear frequently enough to be in the GSL or AWL; however, these words appear frequently in sports news. The majority of the 3,043 word families falls into this group. Hence, in order to read sports news comprehensibly, students need to know quite a variety of words. Field-specific words alone may not help to comprehend the reading materials effectively.

2. Words that appear common but include specialized meanings appear quite frequently in this corpus. These words look normal with common meanings, but when looking into another meaning, these words contain specialized meaning which suits only sports or sports news, such as *backroom, bamboozling, bat, comeback, crossbar, fin, freshman, gaming, makeover, setback, showdown,* etc. For example, *bat*, which students might be familiar with the meaning as being an animal rather than the stick used in baseball or cricket games, *comeback* means a period when someone or something becomes successful or popular again, but in sports news, *comeback* means the situation when the athletes return to the fields or games after a long pause. *Misfiring* refers to the situation when the bullet does not come out, or the plan goes wrong. Interestingly, both definitions can be seen in sports news. *Showdown* normally refers to a big meeting, argument, or fight that finally settles a disagreement between people or proves who is the best, but in sports news, *showdown* means an important match between two great players or teams. Students may not know the true meaning of these words when coming across them in sports news. Therefore, this word list can assist students to understand the sports news better, and students can expand their vocabulary bank when they must read further texts related to sports or physical education.

3. Words with specialized meanings also appear frequently in the corpus. Another type of word which appeared frequently is words with a specialized meaning. Words such as *innings, footy, fairway, tryline, yardage*, etc. are placed in this group. These words contain specialized meanings. *Ballpark* is specifically used with baseball. *Bullpen* means an exercise area for pitchers in baseball. *Cox* refers to someone who directs people in boat racing. *Innings* has a specialized meaning which is only related to cricket. *Footy* is the game or sport of football, usually Australian rules football or rugby league, but not soccer. *Tryline* has a specialized meaning only used in rugby football. *Yardage* is only related to American football, etc.

In this article, only 2 groups of non-GSL and non-AWL can be seen in appendices due to the limit of journal capacity. Only words that appear common but include specialized meanings, and words with specialized meanings are placed in the appendices.

Here are some examples of non-GSL and non-AWL words with common meanings but do not appear frequently in general contexts (not be seen in appendices), etc., *adrenaline, affection/affectionately, agility, auction, audition, banish, banned, bolt, champion, cherish, clash, columnist, competitor, diet, drift,*

endorsement, endurance, footage, heatstroke, indoor, infringement, mentor, offence, outrage, podium, rage, showcase, sneaker, etc.

Implications and Limitations

This word list can be used as a glossary for reading comprehension activities in English classes when sports news is used as a reading source. Teachers can also use this word list to create some useful materials for Physical Education and Sport Science students such as vocabulary projects. Students can also put these words into their vocabulary bank in order to comprehensibly read sports news. However, this word list has never been tried out on students. The word list can be tested to see whether it is effective or not by examining the students. So further experiments with students need to be conducted.

Appendices: Supplementary data

The Following is the Supplementary Data to this Article:

Appendix 1: Words that appear common but include specialized meanings.

Non-GSL and non-AWL content words from sports news		
Achilles hill (Achilles' hill)	fin	rookie/rookies
attacker/attackers	freshman/freshmen	sellout
backroom	fritz	setback/setbacks
bamboozling	gaming	shark
bame	greats	shit
bat/bats/batted/batting	homesick	showdown
bloke	hulk	shutout
buck/bucks	lockout	slim
cabinet	makeover	spearhead/spearheaded/ spearheading
comeback/comebacks	midlands	underdog/underdogs
commonwealth	misfiring	upturn
crossbar	pissed	versus/vs
decider	placings	vet/vets
domino	podcast/podcasts	washout
downplayed	pregame	winnings
fallout	punter/punters	yanks
fightback		

Appendix 2: Words with a specialized meaning.

Non-GSL and non-AWL content words from sports news		
Aces	bogeyed/bogeys	esports
amateur/amateurs	bookies/bookmakers	fairway/fairways
apprentice	boxed/boxer/ boxers/boxing	fastball
archer/archery	bronze	featherweight
arena/arenas	buck/bucks	fielded/fielding/ fielder/fielders
ascot	bullpen	finalist/finalists
autosport	bunker	finisher/finishers
backcourt	caddie/caddies/caddy	fixture/fixtures
backfield	carded	flyweight
backhand	catcher/catchers	footballer/footballers /footballing
backheel	catchweight	footwork
backline	changeup	forehand
backrower	cockpit	fourballs
backup	cornerback	foursomes
badminton	cricket/cricketer/ cricketers/cricketing	freestyle
baller	cruiserweight	friendlies
ballet	curveball	frontrunner
ballon	dash	fullback/fullbacks
ballpark	defenseman	gameweek
baseball	derbies/derby	goalie/goalkeeper/ goalkeepers/ goalkeeping
baseline	doping	goalless
baseman	downfield	goalscorer/goalscorers /goalscoring
basketball/basketballer	downhill	golf/golfer/golfers/golfing
baton	dropbacks	grandmaster
batter/batters	dugout	grandstand/grandstands
birdie/birdied/birdies	dummy	grounder
bleacher	dunk/dunks	groundout
bloodstock	equaliser/equalizer	gymnast/gymnastics /gymnasts

Non-GSL and non-AWL content words from sports news		
halfback	onside	racetrack
halftime	openside	racket/racquet
handball	outfield/outfielder	rallied/rallies/rally/ rallying
keeper	outplayed	rebound/rebounded /rebounding/rebounds
kickboxer	outscore/outscored /outscoring	referee/refereeing/ referees
kickboxing	overmatched	rematch
kickoff/kickoffs	overs	roughing
knockout/knockouts	paceman	rounders
Layup	par	ruck
legspinner	paralympic/paralympics	rugby
linebacker/linebackers	parkrun	scoreless
lineman/linesman/linemen	peloton	scoreline
lineout/lineouts	penalties/penalty	scorer/scorers
lineup/lineups	pitcher/pitchers	scrimmage
marathon	playbook	scrum
matchday	playmaker/playmakers /playmaking	semifinal/semifinals
matchup/matchups	playoff/playoffs	shortstop
middleweight	polo	sideline/sidelined /sidelines
midfield/midfielder/midfielders	postseason	skateboarding
midseason	powerplay	ski/skier/skiing
Miler	preseason	slugger
miscued/miscues	presser	snooker
motorsport	pro/pros	soccer
netball	puck	sportsbooks
Nil	putted	sportsman/sportsmen/ sportspeople
oddsmakers	quarterback/quarterbacks	sportsmanship
offside	quarterfinal/quarterfinals	springboard
olympian	racegoers	sprint/sprinted/ sprinter/sprinters/ sprinting/sprints
olympic/olympics	racer/racers	

Non-GSL and non-AWL content words from sports news		
stadium/stadiums	tiebreak/tiebreaker /tiebreakers/tiebreaking	unseeded
stakeholders	timeout/timeouts	unsportsmanlike
stance	touchdown/touchdowns	uppercut
strikeout/strikeouts	touchline	volleyball
strokes	tournament/tournaments	welterweight
surf/surfer/surfers/ surfing		
swimmer	triathlon	wicketkeeper/ wicketkeepers/ wicketkeeping
swingman	tryline	wicketless
takedown	undercard	wideout/wideouts
tee/teed/teeing	undrafted	wrestle/wrestled/ wrestler/wrestlers/ wrestling
tennis	unplaced	yoga

Acknowledgements

This research was funded by Department of Foreign Languages, Faculty of Humanities, Kasetsart University.

References

[1] Biber, D. (2006). *University Language*. John Benjamins.

[2] Kongcharoen, P. (2018). Basic Physical Education and Sport Science English Word List for Physical Educaiton Students. *rEFLections, 25*(2), 120-147.

[3] Mozaffari, A., & Moini, R. (2014). Academic words in education research articles: A corpus study. *Procedia - Social and Behavioral Sciences, 98*, 1290-1296.

[4] Coxhead, A. (2021). Vocabulary in English in tertiary contexts: connecting research and learning. *LEARN Journal: Language Education and Acquisition Research Network, 14*(1), 1-14.

[5] Gavioli, L. (2005). *Exploring Corpora for ESP Learning.* John Benjamins.

[6] Macmillan, (n.d.). *www.macmillandictionary.com.* Retrieved April 15, 2021, from https://www.macmillandictionary.com/dictionary/british

[7] Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

[8] Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review, 63,* 59-82.

[9] Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

[10] Laufer, B. (1989). What percentage of text lexis is essential for comprehension?. In Lauren, & M. Norman (Eds.), *Special Language: Form Humans Thinking to Thinking Machines. Multilingual Matters,* pp. 316-323.

[11] Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics, 30*(3), 407-427.

[12] 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.

[13] Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure?. *Reading in a Foreign Language, 8*(2), 689-696.

[14] Yu, M., & Renandya, W. A. (2021). A Corpus-based Study of the Vocabulary Profile of High School English Textbooks in China. *LEARN Journal: Language Education and Acquisition Research Network, 14*(1), 28-49.

[15] van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*, 547-479.

[16] Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, England:Cambridge University Press.

[17] Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.

[18] Coxhead, A. (2018). *Vocabulary and English for Specific Purposes Research.* Taylor & Francis.

[19] Coxhead, A. (2010). What can corpora tell us about English for Academic Purposes? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics.* Taylor and Francis.

[20] Davies, Mark. (2004). *British National Corpus*. Oxford University Press. Retrieved from https://www.english-corpora.org/bnc

[21] Davies, Mark. (2008). *The Corpus of Contemporary American English (COCA)*. Retrieved from https://www.english-corpora.org/coca

[22] Davies, Mark. (2010). *The Corpus of Historical American English (COHA)*. Retrieved from https://www.english-corpora.org/coha

[23] Connor, U., & Upton, T. A. (2004a). *Discourse in the Professions: Perspectives from Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins.

[24] Adamson, B. (2004). *China's English : A history of English in Chinese education.* Hong Kong University Press.

[25] Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27-38.

[26] West, M. (1953). *A general service list of English Words.* Longman, Been and Co.

[27] Flowerdew, L. (2004). The argument for using English specialized corpora, in Connor, U., & Upton T. A. (Eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins, pp. 11-33.

[28] Mackenzie, J., Benham, R., Petri, M., Trippas, J.R., Culpaper, J.S., & Moffat, A. (2020). *CC-News-En: A Large English News Corpus.* Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Association for Computing Machinary, pp. 3077-3084.

[29] Bostan, L.A.M, Kim, E., & Klinger, R. (2020). *GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotinals, Semantic Roles, and Reader Perception.* Proceedings of the 12th Language Resources and Evaluation Conference. (pp. 1554-1566). Marseille, France: European Language Resources Association.

[30] Kutter, A., & Kantner, C. (2012). Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention. *International Relations Online Working Paper, 2012*(01), 1-38.

[31] Chotimongkol, A., Saykhum, K., Chootrakool, P., Thatphithakkul, N., & Wutiwiwatchai, C. (2009). *LOTUS-BN: A Thai broadcast news corpus and its research applications.* 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, pp. 44-50.

[32] Al-Khawaldeh, A.A., Mat Awal, N., & Zainudin, I.S. (2014). A Corpus-Based Description of Discourse Markers in Arabic Sport Journalistic Texts. *Journal of Islamic and Human Advanced Research*, *4*(4), 200-215.

[33] Jones, M., & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics.* Taylor and Francis.

[34] Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

[35] Garnder, D., & Davies, M. (2013). A New Academic Vocabulary List. *Applied Linguistics, 35*(3), 305-327.

[36] Moon, R. (2010). What can a corpus tell us about lexis? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics.* Taylor and Francis.

[37] Nation, I. S. P. (2013). *Learning vocabulary in another language (second edition).* Cambridge: Cambridge University Press.

[38] Hyland, K. (2006). *English for Academic Purposes.* Taylor & Francis.

[39] Evison, J. (2010). What are the basics of analysing a corpus? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics.* Taylor and Francis.

[40] Laosrirattananchai, 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.

[41] Kruawong, T., & Phoocharoensil, S. (2020). Developing an English zoology academic word list: A corpus-based study. *Thoughts, 2020*(2), 63-78.

[42] Laosrirattananchai, P., & Ruangjaroon, S. (2020). The word lists of hospitality service review construction. *JSEL, 15*(1), 107-158.

[43] It-Ngam, T., & Phoocharoensil, S. (2019). The development of science academic word list. *Indonesian Journal of Applied Linguistics*, *8*(3), 657-667.

[44] Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes, 39*, 1-11.

[45] Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academy word list. *English for Specific Purposes*, *27*(4), 442-458.