LEARN Journal: Language Education and Acquisition Research Network
(ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 15, No:1, January – June 2022
Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

To What Extent May EFL Undergraduates with EMI Develop English Vocabulary? The Case of Civil Engineering

Wenhua Hsu

whh@isu.edu.tw, Department of Applied English, I-Shou University, Taiwan

APA Citation:
Hsu, W. (2022). To what extent may EFL undergraduates with EMI develop English vocabulary? The case of civil engineering. *LEARN Journal: Language Education and Acquisition Research Network,* 15*(1),* 469-494.

Received 27/05/2021

Received in revised form 02/11/2021

Accepted 11/12/2021

Keywords
EMI; BNC/COCA; lexical text coverage

Abstract

English-medium instruction (EMI) is gaining popularity among EFL higher education institutions. However, not all EMI programs provide the same English immersion as those in the Anglosphere. The researcher targeted English-medium university textbooks as a research focus, since they are first and foremost learning material of specialist knowledge and offer EFL non-English-major students a channel for exposure to English. A 6-million-token textbook corpus of civil engineering (CE) compulsory courses was compiled and the vocabulary level thereof along the word-frequency scale of the British National Corpus and the Corpus of Contemporary American English was measured. Then the researcher sought to estimate how many new words CE majors with EMI can encounter often enough to have an opportunity of learning them. Results show that CE textbooks reached the 5th 1000-word-family level at 95% text coverage and stretched to the 10th 1000 at 98% coverage. Beyond the first 3000 word families, only 3,433 word families occurred 12+ times. This frequency was assumed as a benchmark for incidental learning to occur.

	For EMI practitioners who are concerned with their students' vocabulary development, the results can serve as a reference for future investigations into other disciplines.

Introduction

In response to globalization, English as a/the medium of instruction or English-medium instruction (EMI) is gaining much popularity among higher education institutions in Taiwan, where Taiwanese Mandarin is the official language. Before the burgeoning of EMI, English has been one of the university-wise required General Education subjects for undergraduates regardless of their majors. English for General Purposes (EGP) courses have been provided to freshmen and sophomores two to three hours per week. In line with the goal of General Education, the content of EGP is geared toward general interest across the curriculum, covering versatile subject matter, so as to broaden students' horizon beyond their specialty. Along with EGP courses, some departments offer 2- to 3-credit hour elective English for Specific Purposes (ESP) courses to their students. The two types of English courses are mostly delivered in Taiwanese Mandarin, though. After taking EGP and ESP courses, non-English majors no longer take any English course and may have less exposure to English thereafter, which is often the case in the current English as a foreign language (EFL) setting.

In recent years, the number of EMI degree programs has been increasing by leaps and bounds, since it is one of the evaluation indices regulated by Taiwan's Ministry of Education's (MOE) for university internationalization (MOE, 2018). EMI is generally perceived as a means to enhancing domestic students' English abilities for global competition. For international visibility, more and more universities join this EMI current to raise their prestige in world academic rankings. One initiative for creating the context of internationalization at home is to provide international programs with EMI that enroll both international and local students. Therefore, there are currently international programs (i.e., EMI programs) and non-EMI programs running in parallel in many universities. International students mainly come from Southeast and Northeast Asia to take degree courses alongside Taiwanese students in English-taught programs (MOE, 2020).

Despite higher tuition fees for domestic students, the fast expansion of EMI programs reflects the widespread belief to varying extents that students can simultaneously acquire disciplinary knowledge and enhance English abilities as a result of studying in EMI contexts. It is generally held that EMI programs provide immersion in the target language (English), which facilitates incidental learning (Yeh, 2014). There is some literary evidence to manifest that long-term incidental exposure to English contributes to lexical gains over time (Brown et al., 2008; Pellicer-Sanchez & Schmitt, 2010; Vidal, 2011).

However, not all EMI programs provide the same English immersion as those in majority-native-English-speaking countries. In Taiwan, although there is a growing trend in recruiting international academics with a PhD in a specialized field, they are not necessarily native speakers of English. A significant portion of the faculty members in EMI programs are Taiwanese teachers, despite the fact that they have earned a PhD from one of the core Anglosphere and can speak English fluently. Meanwhile, outside of the EMI classrooms, students are exposed to their mother tongue most of the time. In view of these discrepancies, immersion in the current EMI programs should not be interpreted in the same light as that in English as a second language contexts where teachers and local students are mostly native speakers of English.

Ideally, subject teachers should have expertise in both content and language knowledge so that they can offer some guidance to help their students with EMI to develop disciplinary literacy (Wilkinson, 2013). However, suggesting that subject teachers be trained with linguistic knowledge seems a bit far-fetched. During an investigation into 70 European universities with EMI programs, O'Dowd (2018) discovered the phenomenon that the concern of English improvement usually plays second fiddle compared with disciplinary knowledge development. Constrained by limited class hours, subject teachers may be predisposed to viewing transmitting specialist knowledge adequately as their principal responsibility while delegating English teaching to EGP or ESP teachers (Jiang et al., 2019), leading to a lack of content and language integrated learning (Airey, 2012; Block & Moncada-Comas, 2019; Dearden & Macaro, 2016). In that case, the current EMI programs may largely depend on incidental learning of English, as opposed to EGP or ESP

courses where English language development is the aim of explicit instruction and intentional learning is required.

Accordingly, in EMI programs where there is not much attention paid to language per se, English-medium specialist textbooks may become an important source of input to offer non-English majors a channel for exposure to English. Meanwhile, they are also first and foremost learning materials of specialist knowledge. As such, the researcher targeted English-medium specialist textbooks as a research focus with a particular concern about potential vocabulary growth, for the reason that vocabulary plays a fundamental role in developing language proficiency.

Amid a wide range of academic disciplines, the researcher set out from civil engineering. The selection was simply a random choice since civil engineering is one of the fields of study in many universities and colleges. This research sought to answer the following three questions:

1. What vocabulary level may EFL civil engineering undergraduates attain after finishing their degree program with EMI?
2. What vocabulary size would EFL civil engineering students with EMI need to start with?
3. Beyond the most frequent 3000 word families, how many words from English-medium specialist textbooks may civil engineering undergraduates encounter often enough for learning to occur?

Literature Review

BNC/COCA Word-Frequency Scale

To measure the vocabulary level of a text, a large word-frequency scale is needed. The key to creating a scale for measuring vocabulary levels is to rank words by frequency. In order to make word lists on a large scale, the corpus from which a frequency-ordered lexicon is generated must be genre-balanced and large enough to be a representative sample of spoken and written English (Nation, 2016). Using 100-million-word British National Corpus (BNC) (containing data from the 1980s to the early 1990s) and one-billion-plus-word Corpus of Contemporary American English (COCA) (from 1990 onwards), Nation (2020) has been devoted to the compilation of word lists for over many years. Thus far, he has identified the most frequent 2000 word families

derived from the specially selected corpus and developed the 3rd to 25th 1000 word-family lists based on occurring frequency and dispersion in the BNC/COCA, totaling 25,000 word families with 25 frequency bands. According to Nation (2020), low-frequency word families continue to be added to the existing word lists and the number of 1000-word-family lists have been expanded to 28.

The rationale behind vocabulary levels in association with frequency is that high-frequency words are more likely to be encountered and learned than low-frequency words (Nation, 2006). For example, the words *equipment* and *speed* appear more frequently than the words *apparatus* and *velocity* in the large BNC/COCA. *Equipment* and *speed* rank on the 2nd 1000-word-family level while *apparatus* and *velocity* belong to the 5th 1000. Learners who have some knowledge of *apparatus* may already know *equipment*. If learners know *velocity*, it is highly likely that they are already familiar with *speed*.

Along the BNC/COCA word-frequency scale, Schmitt and Schmitt (2014) labeled the first 3000 word families as high-frequency vocabulary and the first 4000 to 9000 as mid-frequency vocabulary as well as those after the first 9000 as low-frequency vocabulary. They further stressed the importance of mid-frequency words based upon Nation's (2006) estimate that knowledge of the first 9000 word families are essential to fluent reading of all sorts of texts.

The ranked BNC/COCA word lists have been utilized by some researchers to estimate lexical coverage and the vocabulary level of a text as well as the vocabulary amount necessary for adequate comprehension (Hsu, 2020; McQuillan, 2016; Webb & Rodgers, 2009). For an evaluation of the BNC/COCA word lists, Dang and Webb (2016) conducted a series of tests on different word lists and found that the BNC/COCA word lists performed better on a variety of written and spoken texts than the other word lists. In particular, the BNC/COCA 2000 word families were evaluated highly by English teachers (Dang et al., 2020). They are the most frequent general words that occur across all kinds of texts. In this research, the BNC/COCA word lists were used to identify the division amongst the diverse vocabulary levels contained in CE textbooks.

Lexical Text Coverage

Pertinent to the ranked BNC/COCA word lists is lexical text coverage. Nation (2006) defined lexical text coverage as "the percentage of running words in the text known by the reader" (p.61). Through a linear regression to examine the relationship between the text coverage that leaners' vocabulary provides and comprehension of that text, Laufer and Ravenhorst-Kalovski (2010) discovered that vocabulary size accounted for 64% of the variance in the comprehension scores. Applying statistical methods to their dataset, Schmitt et al. (2011) also found a positive linear relationship between the percentage of words known in a text and the comprehension level of that text, showing that better comprehension increases with more known words.

However, researchers have differed on the coverage percentage required for adequate comprehension. Hu and Nation (2000) reported that the density of unknown words had a markedly negative effect on text comprehension. In their study, no subjects with a vocabulary size merely providing 80% text coverage could read fluently. But at 90% coverage, a minority of their subjects were able to read well. When increasing to 95% coverage, a majority of them gained adequate comprehension. At 100% coverage, most of the subjects could read effortlessly. Overall, two putative coverage points 95% (no more than 5% unknown words for minimally acceptable comprehension) and 98% (no more than 2% unknown words for optimal comprehension) are considered to be lexical thresholds for assisted and unassisted reading respectively (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006).

Vocabulary levels and vocabulary size needed for adequate comprehension are two sides of the same coin. Setting a vocabulary threshold at 98% text coverage, English newspapers and novels measure up to the first 8000—9000 word-family levels, while at 95% coverage, they register the first 4000 (Nation, 2006). In other words, knowing 95% and 98% of the words of English newspapers and novels as vocabulary goals entails mastery of a minimum of the first 4000 word families as well as the first 8000—9000 respectively. Since measuring the vocabulary level of a text depends on the predetermined coverage percentage, this study also adopted 95% and 98% as two cutoffs for measuring the vocabulary levels of civil engineering textbooks.

Incidental Learning and Repetitions

Among various factors that are crucial to incidental learning is repetitions in that a new word is seldom acquired merely through a single encounter (Horst et al., 1998). When an unknown word recurs, the possibility of learning that word grows, as repeated occurrences stand a favorable opportunity of increasing the word's prominence and hence the learning thereof. Nevertheless, no consensus has been reached in regard to the number of repetitions that ensures the acquisition of a new word, since learners' proficiency, context clues and even the word itself may affect learning (Webb, 2008; Zahar et al., 2001).

Waring and Takaki (2003) provided some evidence that to have a 50% probability of recalling a newly-learned word after a lapse of three months, eight repetitions or more would be indispensable. They further recommended that a minimum of twenty repetitions be a prerequisite for incidental learning. In a similar vein, Webb (2007) maintained that at least ten repetitions would be required for learning the full meaning of a word. To gain rich knowledge of a word, Pellicer-Sanchez and Schmitt (2010) suggested ten to fifteen repetitions. However, in Vidal's (2011) experiments, the greatest number of new words being acquired appeared in two to three repetitions.

Based on prior results ranging from two to twenty repetitions, Nation (2014) took a middle ground and chose twelve repetitions as the cutoff to measure the amounts of input needed for incidental learning of most of the first 9000 word families. This research also adopted twelve times as the threshold for incidental learning to occur and calculated the number of word families with 12+ occurrences in civil engineering textbooks.

Incidental Learning and the Amount of Input

To support extensive reading and to evaluate its possibility for incidental learning, Nation (2014) selected 25 novels from Project Gutenberg and computed the amount of input needed to meet most of the words at a certain 1000-word-family level at least 12 times for incidental learning to occur. According to Nation (2006), knowledge of the first 9,000 word families would provide 98% coverage of a variety of unsimplified texts. In a later study, setting the first 9000 word families as

a vocabulary goal, Nation (2014) estimated that to encounter most of them over 12 times, learners would need to read a minimum of 3 million words of novels.

Replicating Nation's (2014) method, Hsu (2019, 2020) used VOA news articles and TED Talks English transcripts as input to measure the minimal amount needed for 12+ encounters with most of the first 9000 word families. Results show that learners would need to read 6 million words of VOA news or 4.8 million words of TED transcripts as opposed to 3 million words of novels.

Likewise, when EFL undergraduates with EMI complete their required coursework, they will have read a fairly large amount of English-medium specialist textbooks. In view of voluminous textbook reading, the researcher was concerned about EFL civil engineering majors' vocabulary development after university entry.

Method

Compiling a Textbook Corpus of Civil Engineering

Making reference to the civil engineering (hereafter CE) undergraduate programs in prestigious universities in Taiwan, the researcher identified 21 compulsory courses that most CE departments require their students to take regardless of their selected area of specialization in CE such as public works engineering, architectural engineering, environmental engineering, natural disaster prevention and wood construction. They include four foundation courses (engineering mathematics, calculus, general physics and foundation engineering) within the Faculty of Engineering, nine fundamental professional courses (engineering graphics, surveying, mechanics of materials, fluid mechanics, soil mechanics, mechanics of solids, reinforced concrete, hydrology and introduction to geomatics) as well as eight specialized CE courses (transportation engineering, construction engineering, architectural, structural, hydraulic, environmental, seismic and geotechnical engineering).

Based upon the catalogues of internationally reputable textbook publishers and the professional and technical examinations for engineers (civil engineers, structural engineers, hydraulic engineers, geotechnical engineers, survey engineers, environmental engineers) promulgated by

Taiwan's Ministry of Examination as well as the references list that CE professors provide for their students, the researcher noted down textbook titles. The same books were put in the priority list for later screening. Subsequent to this listing was the selection of textbooks with the 4th or higher edition to make certain their popularity. Furthermore, the textbook for inclusion in the candidate list must have 700+ pages in length for at least one semester use. This arbitrary decision was made after a series of comparison among textbooks in the amount of content and the number of editions.

When the candidate textbook list was finalized, six CE teachers offered their help to confirm reputable textbooks for each required course by ticking the appropriate box on a survey list. Next to each textbook were two questions (concerning if they recognize this book and if they would consider it for classroom use) with three options (yes, no and uncertain) for a check. The textbooks without a consensus of at least two of the six teachers ticking yes for either of the two questions were discarded. The number of candidate textbooks across 21 required courses was eventually reduced to 131.

From the shortlist, the researcher randomly picked one among six to seven candidate textbooks for each required course, totaling 21 textbooks (see Table 1). It was assumed that CE undergraduates in the EMI program would read these core textbooks at some stage within four years of university study. All the sampled textbooks in PDF were downloaded from digital databases, which were subscribed by the university libraries in Taiwan for educational and research purposes. They were saved as plain texts in UTF-8 with photos, graphs and tables being automatically removed. After front matter and references were deleted, the CE Textbook Corpus had approximately 6 million tokens.

TABLE 1

Composition of the CE Textbook Corpus

Required courses	Tokens	Required courses	Tokens
Engineering mathematics	167,161	Hydrology	260,624
Calculus	367,985	Introduction to geomatics	413,514
General physics	357,643	Transportation engineering	286,384

Foundation engineering	157,063	Architectural engineering	460,771
Engineering graphics	158,491	Structural engineering	233,921
Surveying	321,396	Hydraulic engineering	209,048
Mechanics of materials	427,420	Construction engineering	329,523
Fluid mechanics	266,124	Environmental engineering	266,439
Soil mechanics	277,441	Seismic engineering	187,304
Mechanics of solids	230,970	Geotechnical engineering	274,139
Reinforced concrete	365,895	Total:	6,019,256

Instrument and Data Processing

As discussed in Literature Review, vocabulary researchers have generally adopted 95% and 98% lexical text coverage as a lower and upper benchmark respectively for good comprehension to occur. By dint of the BNC/COCA word-frequency scale containing the ranked twenty-five 1000-word-family lists (Nation, 2020), the researcher was able to measure the vocabulary levels of CE textbooks. The rationale for the measurement was to count how many of the ranked 1000-word-family lists from the first 1000 were needed until text coverage accumulated to 95% or 98%. Meanwhile, the vocabulary level of that text was extrapolated according to which 1000-word-family list was the last one being added when the cumulative coverage arrived at 95% or 98%.

The vocabulary analysis program AntWordProfiler (Anthony, 2014) was administered to analyze the lexical profiling of CE textbooks. Apart from BNC/COCA base word lists, another three ever-growing lists of proper nouns, acronyms and compounds complied by Nation (2017) were also installed to the program. During the implementation of AntWordProfiler, the words that were sorted out as the 'Words NOT Found In Base Lists' (hereafter called off-list) were further examined.

If off-list words were personal or place names, they were supplemented to the existing proper noun list. The full forms of acronyms are usually glossed in appendix and can also be found in context (e.g., AC for asphalt concrete, RC for reinforced concrete, BRW for brick retaining wall, ELCB for earth leak circuit breaker, TOB for top of beam). The acronyms in the off-list were added to the existing acronym list. The hyphens of hyphenated compounds were deleted so that they would not be mistaken by AntWordProfiler for off-list words (e.g., tension-saturated, wetting-induced, well-ventilated, weigh-in-motion,

trailing-counterweight-rigged). Closed compounds may be classified as off-list words as well (e.g., crawlspace, weatherproof, switchboard, groundwater, sinkhole, topsoil). They were thus supplemented to the existing compound list.

In calculation of text coverage, excluding the coverage of proper nouns, acronyms and transparent compounds would overestimate the vocabulary level of a text (Nation, 2006; Nurmukhamedov & Webb, 2019). It may not be difficult to recognize a proper noun from its spelling or transliteration. One can infer the meaning of a transparent compound from its constituent words without much effort. It may also not be arduous to look up an acronym in context or in appendix. Consequently, the text coverage of these three types of words were added to that of the base word lists until the coverage percentage totaled 95% or 98%.

Frequency Count

As aforementioned, repeated exposure to an unknown word may raise the likelihood of incidental learning of that word. In the context where college entrance exams include the English test like ours, matriculated undergraduates already have some knowledge of English inflectional and derivational morphemes, which contributes to new word learning (Nagy et al., 1989). Word families were hence decided as a counting unit. The occurrence of any member of a word family in a range of contexts may help students to guess its meaning from contexts, consolidate knowledge from multiple encounters and thus allows for opportunities for learning that word. For instance, the headword *amplify* and its family members *amplifies, amplified, amplifying, amplifier, amplifiers, amplification, amplifications* rank on the BNC/COCA 5th 1000. The occurrences of *amplify* plus its inflected and derived forms were counted together so that the combined frequency would signify the number of repeated exposure to that word family.

Results and Discussion

Vocabulary Levels of the CE Textbooks

Since English-medium specialist textbooks are the major source of input for EFL undergraduates with EMI, Research Question (RQ) 1 'What

vocabulary level may CE undergraduates attain after finishing their degree program with EMI?' can be addressed from the vocabulary levels of English-medium textbooks of CE courses.

Table 2 provides a snapshot of (1) the overall vocabulary levels of CE textbooks at 95% and 98% text coverage and (2) the lexical distribution along the BNC/COCA word-frequency scale as well as (3) the vocabulary demand of CE textbooks as a whole.

TABLE 2

Coverage of Each of the BNC/COCA Word Lists in CE Textbooks

BNC/COCA word lists	Occurrence in tokens	Coverage in tokens (%)	Cumulative coverage (%)
Proper nouns	228,732	3.80%	3.80%
Acronyms	91,493	1.52%	5.32%
Compounds	105,337	1.75%	7.07%
1st 1000	3,476,120	57.75%	64.82%
2nd 1000	882,423	14.66%	79.48%
3rd 1000	632,624	10.51%	89.99%
4th 1000	199,237	3.31%	93.30%
5th 1000	121,589	2.02%	**95.32%**
6th 1000	60,193	1.00%	96.32%
7th 1000	33,708	0.56%	96.88%
8th 1000	29,494	0.49%	97.37%
9th 1000	19,262	0.32%	97.69%
10th 1000	29,494	0.49%	**98.18%**
11th 1000	12,640	0.21%	98.39%
12th 1000	22,873	0.38%	98.77%
13th 1000	7,825	0.13%	98.90%
14th 1000	4,815	0.08%	98.98%
15th 1000	6,621	0.11%	99.09%
16th 1000	6,019	0.10%	99.19%
17th 1000	11,437	0.19%	99.38%
18th 1000	2,408	0.04%	99.42%
19th 1000	3,612	0.06%	99.48%
20th 1000	3,010	0.05%	99.53%
21st 1000	1,806	0.03%	99.56%
22nd 1000	1,204	0.02%	99.58%
23rd 1000	1,204	0.02%	99.60%
24th 1000	602	0.01%	99.61%
25th 1000	1,204	0.02%	99.63%

Off-list	22,271	0.37%	100%
Total	6,019,256	100%	

Note: The bolded figures indicate the level at which the cumulative text coverage has already reached 95% (at the 5th 1000) and 98% (at the 10th 1000).

As can be seen in Table 2, the CE Textbook Corpus contained 6,019,256 tokens (running words). The BNC/COCA 1st 1000 word families accounted for 57.75% of the total words of the corpus, the 2nd 1000 word families 14.66% and the 3rd 1000 word families 10.51% and so forth. This manifests the relative significance of knowing the most frequent 3000 word families because the text coverage of the first 3000 was larger than that of the other 1000 word families by a large margin. At the 7th 1000, the coverage dropped to less than 1%, and from that level onwards, each additional 1000 word families covered a small percentage of the text.

Despite the importance of the first 3000 word families, CE students with this vocabulary size may have great difficulty in reading their specialist textbooks. The cumulative 89.99% coverage provided by the firs 3,000 word families plus proper nouns, acronyms and compounds means unfamiliarity with 10.01% of the total words in a CE text. Encountering one unknown word in every 10 words (in less than one line) may make CE students feel frustrated.

By the 5th 1000-word-family level, 95% coverage was met (specifically 95.32%). In other words, knowledge of the most frequent 5000 word families plus proper nouns and so on would suffice to provide 95% text coverage (circa 5 unknown words in every 100 words). To put it in perspective, if CE freshmen have a vocabulary size of the first 5000 word families, they may read English-medium CE textbooks well in terms of tolerable interruptions (e.g., guessing unknown words or dictionary look-up) to the flow of reading. With the guidance and instruction of CE teachers, the frequency of consulting a dictionary may be fewer than 5 times per 100 words.

By the 10th 1000 level, slightly over 98% coverage was attained (98.18%). This implies that CE textbooks would even rival newspapers and novels, of which the vocabulary levels reach the 8th to 9th 1000 at 98% coverage as per Nation (2006). Table 3 provides further information regarding the vocabulary level of each compulsory course.

TABLE 3

Vocabulary Levels of CE Textbooks Along the BNC/COCA Scale

CE Textbooks	Vocab Level at 95%	Vocab Level at 98%	CE Textbooks	Vocab Level at 95%	Vocab Level at 98%
engineering mathematics	4,500	11,000	calculus	8,000	11,000
foundation engineering	4,500	8,000	general physics	6,000	10,000
engineering graphics	4,500	10,000	surveying	4,500	10,000
fluid mechanics	5,500	10,000	soil mechanics	5,000	10,000
mechanics of solids	6,000	10,000	mechanics of materials	4,500	9,000
hydrology	5,000	11,500	reinforced concrete	5,000	10,000
introduction to geomatics	4,500	10,000	architectural engineering	5,000	10,000
transportation engineering	4,000	8,000	structural engineering	5,000	10,000
construction engineering	4,500	9,000	hydraulic engineering	5,000	11,000
environmental engineering	6,000	12,000	seismic engineering	5,500	11,000
geotechnical engineering	5,000	10,000			

An examination of Table 3 reveals that 14 out of the 21 required subjects reached 95% text coverage at the 4500—5000 word-family levels with the remaining 7 subjects ranging from 4000 to 8000. If CE students have a vocabulary of the first 5000 word families, they would attain adequate comprehension of the 14 English-medium specialist textbooks with similar ease in terms of the frequency of consulting a dictionary or guessing unknown words. Beyond the first 4500—5000 word families, six subjects demanded higher vocabulary thresholds (a minimum of 5500 word families for fluid mechanics and seismic engineering, 6000 for mechanics of solids, general physics and environmental engineering as well as 8000 for calculus). That is, in terms of vocabulary load, the six subjects involving higher vocabulary levels may be more difficult to read than the other subjects. At 98% coverage, 11

subjects extended to the 10,000-word-family level with the other 10 subjects ranging from 8000 to 12,000.

Among the 21 compulsory courses, transportation engineering required the least vocabulary (4000 word families at 95% coverage and 8000 at 98% coverage). At 95% coverage, calculus reached the 8000-word-family level, surpassing the average 4500—5000 levels by a lot. CE students may need to make strenuous efforts to read calculus since it contains more higher-level words. After calculus, environmental engineering was the second heaviest in terms of vocabulary levels (6000 at 95% and 12,000 at 98%). In contrast, transportation engineering may be much easier to read, having a dense distribution of words within the range of the first 4000 word families. Moreover, between 95% and 98% coverage (only a difference of 3%), engineering mathematics and hydrology had the widest dispersion of words, scattering from the 4500 to the 11,000-word-family level and from the 5000 to the 11,500-word-family level respectively. It was conjectured that studying these two subjects would result in CE students working on a wide variety of vocabulary.

The apparent discrepancy between the 4000—8000 levels at 95% coverage and between the 8000—12,000 levels at 98% coverage indicates that majoring in CE entails different vocabulary loads for different compulsory courses. Specifically, when reading with teacher instruction, a minimum of 4000 word families is required for transportation engineering versus 8000 for calculus (i.e., setting at 95% coverage for assisted reading). However, when some chapters are designated as reading assignments, a threshold of 8000 word families is suggested for foundation engineering versus 12,000 for environmental engineering (i.e., setting at 98% coverage for unassisted reading). As can be seen in Table 3, either at 95% or 98% coverage, there was a striking difference in vocabulary levels among the compulsory courses. This hints that some courses entail the learning of a much larger vocabulary than the others.

In answer to RQ1, Table 2 shows that CE textbooks generally reached the 5th 1000-word-family level at 95% coverage and stretched to the 10th 1000 at 98%. The latter gives us a beacon of hope concerning what vocabulary level CE undergraduates may attain after finishing their degree program with EMI. If CE students ideally increase their vocabulary to 9000 word families during four years of study, with this vocabulary size

they would be able to read a diversity of authentic texts according to Nation (2006). However, the results for RQ3 show a different picture.

Starting Vocabulary for CE Students with EMI

RQ2 'What vocabulary size would EFL civil engineering students with EMI need to start with?' can be tackled from 95% coverage under assisted reading conditions. At this coverage point, EFL students planning to study CE undergraduate program with EMI may need to have a minimal vocabulary of the first 5000 word families and optimally the first 10,000 word families (providing 98% coverage) in order to perform all sorts of reading tasks well in their field (see Table 2).

However, for many FEL learners, having a grasp of the first 5000 word families is still not an easy job, let alone 10,000 word families. It is worth noting here that after the cumulative coverage reached 95%, by the 7th 1000 level, each additional 1000 word families provided only a small increase in coverage (less than 1%). In the light of less than 1% text coverage from each of the 7th to 25th 1000-word-family lists, not all the words would be equally significant for CE majors. Some words may appear only once or twice throughout the entire textbooks. It shows a potential need for a CE-specific word list with high frequency, distinct from the BNC/COCA word lists in the upper scale. A CE-specific word list would be worth developing if it provides a better learning return in terms of greater text coverage compared with less than 1% coverage from the 7th to 25th 1000. The issue of creating a more restricted CE-based word list is worth pursuing but is beyond the present research focus. For discussion of the selection criteria in the development of a science-specific or an engineering-specific word list with high text coverage, see Coxhead and Hirsh (2007), Hsu (2014), Mudraya (2006) and Ward (1999, 2009).

Amount of Vocabulary Beyond the First 3000 Word Families Occurring 12+ Times

Table 4 provides a listing of the number of word families from the BNC/COCA word lists appearing 12+ times in CE textbooks as opposed to in novels and news.

TABLE 4

Number of Word Families from the BNC/COCA Word Lists Occurring 12+ Times

	3 million words of novels (Nation 2014, p.6)	6 million words of VOA news (Hsu 2019, p.417)	6 million words of CE textbooks across 21 core courses
Vocabulary level at 98% coverage	8000—9000	6,000	10,000
1st 1000	N.A.	N.A.	906/ (984)
2nd 1000	994	N.A.	831/ (970)
3rd 1000	972	N.A.	840/ (975)
1st—3rd 1000 subtotal	N.A.	N.A.	2,577/ (2,929)
4th 1000	945	996	628/ (898)
5th 1000	929	985	490/ (816)
6th 1000	904	962	366/ (746)
7th 1000	857	898	295/ (661)
8th 1000	817	851	246/ (584)
9th 1000	805	805	204/ (523)
4th –9th 1000 subtotal	5,257	5,497	2,229/ (4,228)
10th—25th 1000	N.A.	N.A.	1,204/ (3,828)
Total	N. A.	N. A.	6,010/ (10,985)

Note: The figures in parentheses indicate the total number of word families at a particular 1000 word-family level appearing in CE textbooks, including the number of words occurring fewer than 12 times.

The bottom right column in Table 4 shows that the 6-million-token CE Textbook Corpus across 21 compulsory courses contained 10,985 word families from the BNC/COCA 1st to 25th 1000, including 2,929 word families from the first 3000. Following Nation (2014), twelve repetitions were adopted in this research as a threshold for incidental learning to occur. The figures without parentheses in the right column of Table 4 are bona fide the numbers of word families occurring 12 times or more.

This research aimed at vocabulary development beyond the first 3000 word families in that CE matriculated students already know the first 3000 word families after passing the college entrance exam. Table 4 answers RQ3 'Beyond the most frequent 3000 word families, how many

words from English-medium specialist textbooks may CE undergraduates encounter often enough for learning to occur?', demonstrating that a total of 10,985 out of the 25,000 word families (44%) would be encountered when CE undergraduates complete their compulsory courses. The 10,985 word families included words occurring fewer than 12 times and quite a few one-timers (words appearing once) such as *riot, tumor, agony, venter, shigellosis*, to name but a few. Only 6,010 out of the 10,985 families (55%) appeared 12 times or more. However, among the 6,010 word families, 2,577 word families came from the first 3000, with only 3,433 word families belonging to the 4th to 25th 1000.

As mentioned earlier, in support of extensive reading, Nation (2014) used novels as input to estimate a minimal amount to read to gain 12+ encounters with 800+ word families from each of the mid-frequency bands (4th to 9th 1000). He gauged that continual reading up to 3 million words of novels would help learners to meet most of the first 9,000 word families often enough for incidental learning (see Table 4). In contrast, Hsu (2019) reported her finding that although VOA news only reaches the 6th 1000-word-family level at 98% coverage, by voluminously reading VOA news up to 6 million words, EFL learners would get sufficient input to encounter most of the first 9000 word families enough times.

Compared with Nation (2014) and Hsu (2019), even though CE undergraduates finish reading circa 6 million words of English-medium specialist textbooks, they would still not meet most of the mid-frequency vocabulary at least 12 times (see Table 4 for 628 < 800 word families at the 4th 1000). With the advance towards the 9th 1000-word-family level, the number of mid-frequency vocabulary occurring 12+ times becomes smaller and smaller, assuming that 12 repetitions are the necessary threshold for incidental learning to occur. There was a reduction from 628 to 204 word families at the 9th 1000, which were a lot fewer than 800 word families.

This may be because, in most novels and news articles, a large number of different words are used. It may also be the diversity of topic areas involved in novels and news that result in the richness of vocabulary. In contrast to novels and news, the vocabulary recycling of CE textbooks is strong, which may reduce students' vocabulary load when they go on to another compulsory textbook. But CE majors may make small progress in mid-frequency vocabulary learning (totaling 2,229

word families occurring 12+ times) in comparison with novels (5,257) and news articles (5,497) (see Table 4).

TABLE 5

Number of Word Families Beyond the First 3000 Occurring in CE Textbooks 12+ Times

Textbooks	Vocab level at 95%/ 98% coverage	4th to 9th 1000	10th to 25th 1000	Total
CE textbooks as a whole	5,000/ 10,000	2,229	1,204	3,433
Engineering mathematics	4,500/ 11,000	376	152	528
Calculus	8,000/ 11,000	599	182	781
General physics	6,000/ 10,000	616	337	953
Foundation engineering	4,500/ 8,000	686	235	921
Engineering graphics	4,500/ 10,000	347	113	460
Surveying	4,500/ 10,000	846	346	1,192
Fluid mechanics	5,500/ 10,000	915	366	1,281
Soil mechanics	5,000/ 10,000	1,093	399	1,492
Mechanics of solids	6,000/ 10,000	1,499	542	2,041
Mechanics of materials	4,500/ 9,000	1,674	724	2,398
Hydrology	5,000/ 11,500	1,043	435	1,478
Reinforced concrete	5,000/ 10,000	735	188	923
Introduction to geomatics	4,500/ 10,000	1,202	348	1,550
Architectural engineering	5,000/ 10,000	1,790	610	2,400
Transportation engineering	4,000/ 8,000	952	332	1,284
Structural engineering	5,000/ 10,000	1,061	371	1,432
Construction engineering	4,500/ 9,000	1,624	557	2,181
Hydraulic engineering	5,000/ 11,000	732	201	933
Environmental engineering	6,000/ 12,000	1,302	793	2,095
Seismic engineering	5,500/ 11,000	778	257	1,035
Geotechnical engineering	5,000/ 10,000	1,536	547	2,083

Table 5 demonstrates the numbers of word families beyond the first 3000 occurring 12+ times in each CE textbook and in the CE textbooks as a whole. If a minimum of 12 repetitions for incidental

learning to occur is true, Table 5 reveals the amount of potential vocabulary growth for CE undergraduates as a result of constant exposure to English-medium specialist textbooks.

For instance, when studying calculus, EFL students would meet 599 mid-frequency word families (4^{th}—9^{th} 1000) and 182 low-frequency word families (10^{th}—25^{th} 1000) 12+ times, if they finish reading all the chapters in the calculus textbook. Likewise, other CE textbooks would also offer incidental learning opportunities of mid-frequency and low-frequency words with the architectural engineering textbook providing the most (totaling 2,400 word families), followed by the mechanics of materials textbook (2,398 word families), and with the engineering graphics textbook providing the least (460). It is worth noting that the words appearing fewer than 12 times in a single textbook may be met a few more times later when reading another textbook. When more and more CE textbooks are read, the number of word families beyond the first 3000 occurring 12+ times would gradually increase, as a total of 3,433 word families have shown in the CE textbooks as a whole versus 460 to 2,400 in a single textbook (see Table 5).

Despite the potential learning of 2,229 mid-frequency word families from CE textbooks, it is still far below the target of learning most of the 4^{th} to 9^{th} 1000 word families, namely at least 4,800 out of 6000 mid-frequency word families as per Nation (2014) and Hsu (2019). If knowledge of the first 9000 word families (providing 98% coverage of a variety of texts) is the vocabulary goal, the limited lexical provision by CE textbooks is apparent, compared with 3 million words of novels and 6 million words of VOA news containing 5,257 and 5,497 mid-frequency word families respectively occurring over 12 times.

Through a statistical analysis of the dataset in Table 5, the results demonstrate that there was no significant relationship between the vocabulary level of a textbook and the number of word families beyond the first 3000 occurring 12+ times ($r=0.079$, $p=.734$). This signals that a specialist textbook using higher-level vocabulary does not necessarily have a greater variety of words than we have expected of from that textbook. For instance, the calculus textbook arrives at the 8000-word-family level at 95% coverage and the 11,000 level at 98% coverage, because it contains many computation-related words, ranking in the upper bands of the BNC/COCA word-frequency scale. But only a very

small vocabulary is frequently used, as 781 word families out of the 21,000 (4^{th}—25^{th} 1000) appearing 12+ times have shown (see Table 5).

This may be true of other CE subjects. Initially, CE freshmen with limited vocabulary may be daunted by many so-called mid-frequency and low-frequency words such as *lateral, shear, deflect, modulus, flange, impervious, girder, torsion* and the like. Fortunately, most of them appear very frequently, functioning as lay-technical, sub-technical or highly-technical words. Over time, CE majors may become familiar with these new words with the help of textbook illustrations and with teacher guidance. As with Table 4, Table 5 reconfirms that CE novices with EMI may need a list of the most frequent CE-specific words for imminent learning. This decision can be left to EMI subject teachers or ESP teachers. How adjunct ESP teaching can facilitate CE study with EMI is another issue worthy of exploration but is beyond the scope of this paper. For ways to fostering team teaching and to boosting the quality of EMI programs by content and language integrated learning (CLIL), see Airey (2012), Block and Moncada-Comas (2019), Dearden and Macaro (2016), Lasagabaster (2018, 2021) and Stewart (2018).

Conclusion

This lexical research was a preliminary study on an EMI civil engineering undergraduate program in an EFL setting. It had a dual purpose: to measure (1) the vocabulary levels of CE textbooks and (2) the amount of vocabulary beyond the most frequent 3000 word families contained in CE textbooks. Generally, CE textbooks involving different specialist knowledge reached the 5^{th} 1000-word-family level at 95% text coverage and extended to the 10^{th} 1000 at 98% coverage. Results show that in the 6-million-token CE textbook corpus, 3,433 word families from the 4^{th} to 25^{th} 1000 levels occurred 12+ times, including 2,229 mid-frequency word families (4^{th}—9^{th} 1000). Even though EMI CE majors complete their core courses, continual reading of English-medium CE textbooks will still not help them to encounter most of the 6000 mid-frequency word families often enough for incidental learning to occur. This hints that an EMI program like civil engineering does not necessarily warrant the highest inclusion of mid-frequency words. It is highly likely that CE majors' vocabulary size would level off at the first 3000 word families plus one-third of the mid-frequency vocabulary (2,229/6,000), if

they do not read English texts outside of their specialist domain, which may often be the case in EFL settings. The value of this study has been to raise this awareness. When EMI has become a nationwide trend at tertiary education, the present results contribute to an understanding of what expectations we may reasonably have of the lexical development in an EMI program.

Concerning the vocabulary goal of the first 9000 word families for reading a variety of authentic texts, one advice to EFL undergraduates may be to continually read English newspapers, novels and all sorts of English articles. Or for this vocabulary goal, EGP teachers can encourage their students to do extensive reading when they take freshman English courses.

Although this paper contributes to the literature of EMI research, it has been worked within a narrow focus on the field of civil engineering. The findings may serve as a basis of comparison for investigations into other disciplines in the EMI mode. It is hoped that this study may provide some inspirations for future qualitative analyses of EFL learners' perceptions of English-medium specialist textbooks regarding reading difficulty.

Acknowledgements

This research was supported by a grant (MOST 110-2410-H-214-003) from Taiwan's Ministry of Science and Technology.

About the Author

Wenhua Hsu: A professor at I-Shou University, Taiwan. Her research interest includes frequent academic/sub-technical and lay-technical vocabulary as well as lexical bundles in specialized fields.

References

Airey, J. (2012). "I don't teach language." The linguistic attitudes of physics lecturers in Sweden. *AILA Review, 25*(1), 64–79. https://doi.org/10.1075/aila.25.05air

Anthony, L. (2014). *AntWordProfiler (Version 1.4.1)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software

Block, D., & Moncada-Comas, B. (2019). English-medium instruction in higher education and the ELT gaze: STEM lecturers' self-positioning as NOT English language teachers. *International Journal of Bilingual Education and Bilingualism.* Advance online publication. https://doi.org/10.1080/13670050.2019.1689917

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language, 20*(2), 136–163. https://doi.org/10125/66816

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue française de Linguistique Appliquée, 12*(2), 65–78. https://www.cairn-int.info/article.php?ID_ARTICLE=E_RFLA_122_0065

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

Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168820911189

Dearden, J., & Macaro, E. (2016). Higher education teachers' attitudes towards English medium instruction: A three-country comparison. *Studies in Second Language Learning and Teaching, 6*(3), 455-486. https://doi.org/10.14746/sllt.2016.6.3.5

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207–23. https://doi.org/10125/66953

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*, 54-65. https://doi.org/10.1016/j.esp.2013.07.001

Hsu, W. (2019). Voice of America (VOA) news as voluminous reading material for mid-frequency vocabulary learning. *RELC Journal, 50*(3), 408-421. https://doi.org/10.1177/0033688218764460

Hsu, W. (2020). Can TED talk transcripts serve as extensive reading material for mid-frequency vocabulary learning? *TEFLIN Journal,*

31(2), 181-203. http://dx.doi.org/10.15639/teflinjournal.v31i2/181-203

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. https://doi.org/10125/66973

Jiang, L., Zhang, L.J. & May, S. (2019). Implementing English medium instruction (EMI) in China: Teachers' practices and perceptions, and students' learning motivation and needs. *International Journal of Bilingual Education and Bilingualism, 22*(2), 107-119. https://doi.org/10.1080/13670050.2016.1231166

Lasagabaster, D. (2018). Fostering team teaching: Mapping out a research agenda for English-medium instruction at university level. *Language Teaching, 51*(3), 400 – 416. https://doi.org/10.1017/S0261444818000113

Lasagabaster, D. (2021). Team teaching: A way to boost the quality of EMI programmes? In F. D. Rubio-Alcalá, & D. Coyle (Eds.), *Developing and evaluating quality bilingual practices in higher education* (pp. 163-180). Multilingual Matters. https://doi.org/10.21832/9781788923705-011

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30. https://doi.org/10125/66648

McQuillan, J. (2016). What can readers read after graded readers. *Reading in a Foreign Language, 28*(1), 63-78. https://doi.org/10125/66715

Ministry of Education (2018). *Implement in full scale bilingualization of Taiwan's educational system; cultivate bilingual talents to bring Taiwan to the world.* [Data file]. Retrieved from https://english.moe.gov.tw/cp-13-17790-80201-1.html

Ministry of Education. (2020). Foreign student number in 2020. [Data file]. Retrieved from https://stats.moe.gov.tw/

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*(2), 235–256. https://doi.org/10.1016/j.esp.2005.05.002

Nagy, W., Anderson, R., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 262–282. https://doi.org/10.2307/747770

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

Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16. https://doi.org/10125/66881

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

Nation, I.S.P. (2020). The BNC/COCA word family lists. Retrieved from https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf

Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching, 52*(2), 188-200. https://doi.org/10.1017/S0261444819000028

O'Dowd, R. (2018). The training and accreditation of teachers for English medium instruction: An overview of practice in European Universities. *International Journal of Bilingual Education and Bilingualism, 21*(5), 553–563. https://doi.org/10.1080/13670050.2018.1491945

Pellicer-Sanchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language, 22*(1), 31–55. https://doi.org/10125/66652

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

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

Stewart, T. (2018) Expanding possibilities for ESP practitioners through interdisciplinary team teaching. In Y. Kırkgöz, & K. Dikilitaş (Eds.), *Key Issues in English for specific purposes in higher education* (pp. 141-156). Springer, Cham. https://doi.org/10.1007/978-3-319-70214-8_9

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258. http://dx.doi.org/10.1111/j.1467-9922.2010.00593.x

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language, 12*(2), 309–323. https://doi.org/10125/66967

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170–182. https://doi.org/10.1016/j.esp.2009.04.001

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language,15*(2), 130–163. https//doi.org/10125/66776

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics,28*(1)*,* 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20*(2), 232–245. https://doi.org/10125/66826

Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics, 30*(3), 407–427. https://doi.org/10.1093/applin/amp010

Wilkinson, R. (2013). English-medium instruction at a Dutch university: Challenges and pitfalls. In A. Doiz, D. Lasagabaster, & J. M. Sierra (Eds.), *English-medium instruction at universities: Global challenges* (pp. 3–24). Multilingual Matters. https://doi.org/10.21832/9781847698162-005

Yeh, C. C. (2014). Taiwanese students' experiences and attitudes towards English-medium courses in tertiary education. *RELC Journal, 45*(3), 305–319. https://doi.org/10.1177/0033688214555358

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review, 57*(4), 541–572. http://dx.doi.org/10.3138/cmlr.57.4.541