Doctoral Thesis

Bridging the Vocabulary Gap for English as an Additional Language Learners

Gavin Brooks

Division of Integrated Arts and Sciences
Graduate School of Integrated Arts and Sciences

Hiroshima University

March 2023

Summary

This thesis investigates the vocabulary knowledge of English as an Additional Language (EAL) learners studying at international schools in Japan. My goal in writing the thesis is to provide these learners with the support that they need to be successful in the classroom. To do this, I investigated two different but connected strands of research: an investigation of the vocabulary knowledge that EAL learners currently possess and an examination of the vocabulary that these learners require to be successful in the classroom. In the final part of the thesis, to bridge the gap that exists between the vocabulary that EAL learners know and the vocabulary they need to succeed academically, I discuss the development of the International School Academic Vocabulary Lists, a set of domain-specific word lists designed specifically to support EAL learners in the classroom.

The first two experimental chapters of the thesis deal with measuring EAL learners' vocabulary knowledge and determining how this will influence their ability to succeed in the classroom. I did this in two stages. In the first stage, I used a battery of assessment tools to examine the importance of vocabulary for EAL learners' reading comprehension. This investigation showed that vocabulary was the single biggest predictor of EAL learners' ability to understand the content of written texts. I then examined the vocabulary knowledge of EAL learners and looked at the coverage this vocabulary knowledge would give them over the texts that they are expected to read in the classroom. This analysis made use of a large corpus of academic texts taken from a representative sample of the subjects that international school students would be likely to study. This analysis allowed me to determine two important facts. First, EAL learners do not possess the vocabulary knowledge necessary to comprehend the textbooks they are required to read without significant support from their teachers. Second, the gap between what EAL learners know and what they would need to know to be able to understand these texts is too great to be covered by existing word lists.

In the next two chapters, I addressed this issue by compiling a representative corpus of textbooks (the International School Corpus of Academic Texts, IS-CAT) and developing a set of domain-specific word lists from this corpus that can be used in the classroom. In Chapter Five, I detail my initial attempt to build a corpus of international

school textbooks and create a more appropriate word list from this corpus. For this chapter, I compiled these initial word lists following the traditional methodology that researchers have used to develop word lists. This methodology involves the use of frequency and range to select the words that would be most useful to know when reading in a specific domain. The word lists that I developed in this chapter were successful in providing much greater coverage of the IS-CAT than existing word lists such as the Academic Word List (AWL). However, I was able to identify some issues that needed to be addressed with these word lists, including my use of the outdated General Service List to remove high-frequency words, the use of words instead of lemmas as a unit of counting, and the exclusion of high-frequency words with academic meanings from the word lists.

I address these issues in Chapter Six by using more modern techniques to create a new set of IS-AVL. In this chapter, I show that these updated word lists not only provide greater coverage than the word lists I compiled in Chapter Five but also provide greater coverage over the IS-CAT than competing word lists such as the AWL, Middle School Vocabulary Lists, or Secondary Vocabulary Lists (SVL). Through an analysis of the coverage the IS-AVL word lists provide over the various subcorpora of the IS-CAT, a parallel corpus, and a corpus of non-academic texts, I am able to show that these word lists have the potential to provide a valuable tool that can be used to support EAL learners in the classroom.

I end the dissertation with a discussion of what still needs to be done to be able to effectively use these word lists to support EAL learners in the classroom. I also explain how the lists I developed in this dissertation will allow teachers to identify the vocabulary that EAL learners need to succeed in the classroom. I hope that the insights from this dissertation, and the word lists I have compiled, will provide additional tools that can be used to help teachers to support their learners in the classroom, making it easier for EAL learners to get the educational experience that they deserve.

Acknowledgements

I would like to thank everyone who has supported me during the process of researching and writing this thesis. I owe a debt of gratitude to all of the family members, teachers, mentors, and colleagues who have guided and supported me in this endeavour

Firstly, I would like to thank my supervisor Jon Clenton for taking me on as a student and providing me with the support that I needed to complete this project. I have learned a lot from his insight and patience, and I appreciate all his advice and all of the opportunities that he has given me over the course of my studies.

I would also like to thank Simon and all the others whose input made this thesis possible. I hope that I have been able to learn and improve from their advice.

I would also like to thank my friends for the support and encouragement they offered. And to thank my parents, for listening to my ideas and offering their advice and input and for making it possible for me to get to where I am now.

Finally, and most importantly, I would like to thank my family, Shizuka, Maya, and Amber, for supporting me, pushing me, and putting up with me during this process. Without your support, this thesis would not have been possible. I would like to dedicate this thesis to you

Table of Contents

Tables and Figures

List of Abbreviations

AVL	Academic Vocabulary List
AWL	Academic Word List
BNC	British National Corpus
COCA	Corpus of Contemporary American English
DP	Deviation of Proportions
EAL	English as an Additional Language
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EMI	English as a Medium of Instruction
ESL	English as a Second Language
FLE	First Language English
GSL	General Service List
IB	International Baccalaureate
IS-CAT	International School Corpus of Academic Texts
IS-AVL	International School Academic Vocabulary Lists
K	1,000
L1	First language
L2	Second language
MSVL	Middle School Vocabulary Lists
NLP	Natural Language Processing
NNS	Non-Native Speaker
NNSEAL	Non-native speaker EAL learner
NS	Native speaker
nVLT	New Vocabulary Levels Test

OCR	Optical Character Recognition
POS	Part of Speech
SLA	Second Language Acquisition
SVL	Secondary School Vocabulary Lists
SWRT	Single Word Reading Test
TOK	Theory of Knowledge
uVLT	Updated Vocabulary Level Test
UWL	University Word List
VLT	Vocabulary Levels Test
YARC	York Assessment of Reading Comprehension

Chapter One

Introduction

1.1 Introduction

This thesis investigates the importance of vocabulary for English as an additional language (EAL) learners studying at international schools in Japan. I hope that this thesis will allow me to develop a set of word lists that can help support these learners in the future. However, before we can look at how and why we would want to do this, it is first necessary to explain who EAL learners are and why a set of vocabulary lists would be invaluable to this group of learners. While EAL learners make up a diverse group of students, they are characterized as those learners who are studying in a classroom where the medium of instruction is English, but who speak a language other than English as their home language (Murphy, 2014). Sharples (2021) identifies two implications that stem from this definition: i) the primary place of learning for these learners is the mainstream classroom and ii) for these learners the learning of language is inseparable from content. While he correctly points out that the challenges faced by EAL learners are not unique, there is considerable research that suggests that the language difficulties that EAL learners face are significant and that these difficulties play a significant role in their academic struggles (e.g., Afitska & Heaton, 2019; Clegg & Afitska, 2011; Coxhead & Boutorwick, 2018).

However, to better support EAL learners in the classroom, we first need to have a clearer understanding of their academic needs (Hawkins, 2005). Something that has become even more critical considering the recent increases in the number of EAL learners, both in countries where English is spoken as a first language, and internationally. In countries where English is spoken as a first language, the inward migration of economic migrants, asylum seekers, and refugees (Sharples, 2021) has resulted in significant increases in EAL learners studying in the mainstream classroom. In the United Kingdom (UK), for instance, schools have experienced double-digit increases in the number of learners whose home language is one other than English over the last ten years (Strand et al., 2015). A similar situation has also been unfolding in the United States. For example, the number of EAL learners enrolled in public schools in the US in 2019 was estimated to be around 5.1 million students, or 10.4% of the total

students enrolled in the public school system (US Department of Education, National Center for Education Statistics, 2022). However, this growth is not uniform, and the percentage of EAL learners in some states, such as Texas, has reached as high as 19.6%. Such increasing numbers of EAL learners is not limited to English-speaking countries as a trend can be seen internationally with the expansion of international schools. The worldwide number of International Baccalaureate (IB) schools, for example, grew 33.3% between 2016 and 2020 (*International Baccalaureate Facts and Figures*, 2022).

Vocabulary is one aspect of linguistic proficiency with which EAL learners have been shown to struggle. The research shows that EAL learners typically enter school with significantly lower levels of vocabulary knowledge than their First Language English (FLE) peers (NALDIC, 2015, October 27). EAL learners also tend to take longer to master the vocabulary necessary for academic success (Coxhead & Boutorwick, 2018). Previous research supports the claim that EAL learners likely have lower levels of vocabulary knowledge than their FLE classmates (e.g., August et al., 2005; Coxhead & Boutorwick, 2018). What remains unclear, however, is how these differences in vocabulary knowledge influence EAL learners' ability to succeed academically in English as a Medium of Instruction (EMI) classrooms. Furthermore, the amount of vocabulary that EAL learners require in the classroom is manifestly subject and grade dependent (Coxhead, 2012).

What appears to have been established is that a lack of vocabulary knowledge can cause EAL learners to struggle with reading comprehension (Brooks et al., 2021). The reason for this is that vocabulary is an integral part of the reading process, and learners who cannot master the vocabulary that is used in the texts they are required to read for their classes often struggle to comprehend these texts (Coxhead et al., 2010; Droop & Verhoeven, 2003). While it is still not clear precisely how much vocabulary EAL learners need to succeed academically in an EMI context, research shows that EAL learners are likely to struggle with two types of vocabulary that are essential for academic success, high-frequency vocabulary and general academic vocabulary (Coxhead & Boutorwick, 2018). Research has consistently highlighted the importance of these high- and mid-frequency vocabulary items for learners studying in an EMI environment. For example, Laufer (1989) suggests that 5,000 words represents the lexical threshold beneath which other facilitating factors in reading comprehension may

not be effective. However, there is ample evidence to suggest that EAL learners begin their education with significantly less vocabulary than this (Brooks et al., 2021; Coxhead & Boutorwick, 2018). The implication of this is that vocabulary expansion should be one of the major goals of any EAL program.

While it is clear from the discussion above that there is a pressing need to support EAL learners in their vocabulary acquisition, it is still unclear what words these learners need to know to be successful in the classroom. The rest of this chapter will look at ways in which researchers have attempted to better understand the gaps that exist in EAL learners' vocabulary knowledge. After this overview, I will give a brief account of how research into the use of corpora and frequency lists can help bridge these gaps. However, before I begin this discussion, I want to start this dissertation by giving some background on why I chose this area of research for my thesis.

1.2 Background

The concern of this thesis is to provide EAL learners with the tools that they need to succeed in the classroom. This topic was motivated both by my teaching experience and my current situation as a father and educator in Japan. I have been fortunate enough to work in the international school context, teaching for almost five years at an International Baccalaureate school in South America. During this time, I taught several subjects, both in the sciences and the humanities. This experience allowed me to witness first-hand the struggles that EAL learners can experience with discipline-specific vocabulary and discourse. At this school, students who were often fluent and confident discussing daily events would find it extremely difficult to explain the simple steps that they had taken to conduct an experiment or explain the significance of a basic academic concept. Due to the international school context, the subjects were taught using textbooks and materials written for First Language English (FLE) learners. There was often little language support provided for EAL learners in the textbooks and materials they gave us to use in the classroom. Now that I have a child of my own, who will enter the international school system in the future, I hope that the ideas and tools that come out of this dissertation will help to provide her, and students like her, the support that they need to overcome these problems.

I believe, however, that this area of research is important not just for my personal reasons, but also at an international level. In November 1989, the United Nations Convention on the Rights of the Child adopted a resolution that enshrined the right of all children to receive an education that permits them to develop to the best of their abilities and talents (United Nations, 1989: articles 28–30). For EAL learners, if this is to be realised, it is first necessary to address their language needs. However, efforts to do this have been hampered by a lack of research in this area. This gap in the research is clear in vocabulary instruction and assessment. For example, to date, there are no assessment tools or vocabulary frequency lists that specifically target EAL learners. As a result, researchers who are investigating vocabulary knowledge in EAL learners are left using frequency lists and assessment tools developed for ESL or FLE learners (e.g., Brooks et al., 2021; Coxhead, 2012; Coxhead & Boutorwick, 2018). To provide these learners with the educational experience that they deserve, it is first necessary to develop the tools necessary to support them in the classroom.

1.3 The vocabulary in the EAL classroom

1.3.1 The importance of vocabulary

The reason for my focus on vocabulary in this dissertation is that researchers have consistently shown that vocabulary is an important predictor of a learner's ability to understand and produce both written (e.g., Daller & Phelan, 2013; Milton & Treffers-Daller, 2013) and spoken texts (e.g., Clenton et al., 2020; de Jong et al., 2013). Presently there are word lists that have been designed to provide learners with support at different levels and in different contexts. These include university level academic word lists (e.g., Coxhead, 2000; Gardner & Davies, 2014), a middle school word list (Greene & Coxhead, 2015), and word lists for specialized subjects such as medicine and pharmacology and linguistics (e.g., S. Fraser, 2007; Lei & Liu, 2016; Vongpumivitch et al., 2009). However, to date there is no word list that has been developed specifically for EAL learners studying in an international school context.

As I outlined above, for EAL learners, one area where vocabulary knowledge is critical is reading comprehension. Research has firmly established a link between vocabulary knowledge and reading comprehension and proficiency (e.g., Graves et al., 2012; Nagy & Townsend, 2012; Ogle et al., 2015; Pressley & Allington, 2014). The

reason for such importance is that learners need to know the words that are used in a text to understand what that text is saying. However, to properly support learners in their vocabulary acquisition, it is important to better understand what words they actually need to know. In relation to the EAL learners studying the EMI classroom, this involves a focus on academic and domain-specific technical vocabulary.

1.3.2 The importance of academic and technical vocabulary for EAL learners

One of the most important types of vocabulary for EAL learners, who are required to read academic books and listen to academic lectures as part of their classes, is academic vocabulary. Academic vocabulary refers to the type of words that one would need to function in an academic contexts, especially the vocabulary needed to read and understand academic texts. The reason for this is that the writers of academic textbooks tend to use language that is more precise and has a more specialized register that is often unfamiliar in meaning to EAL learners (Leung, 2014). Schmitt & Schmitt (2020) give the following example to help illustrate what academic vocabulary is and why it would be difficult for an EAL learner to understand:

> A. The company changed its marketing ideas to try to make more money.
>
> B. The company modified its marketing strategy to try to increase revenue.
>
> (p. 7)

Schmitt and Schmitt point out that, on the surface, the difference between these two sentences is that the first is easier to understand because it uses only high-frequency words, while the second has a more academic register and is more precise; because the second sentence is written in an academic style. The underlined words, while similar in meaning, are more precise than the high-frequency words used in the first sentence. For example, while *change* and *modify* both include the idea of changing something from one thing to another, *modify* also includes the concept of making a small change to improve something. However, while these words may make the sentence more academic and more precise, they also make it harder for EAL learners to understand. These sentences also highlight the importance of academic word lists, as these lists provide teachers with a tool that they can use to help support EAL learners with the academic vocabulary that they need to understand these texts.

Several studies have focused on identifying the academic vocabulary necessary to understand textbooks, articles, or other academic texts. One of the earliest studies in this area is Barber (1962), who noted that there were certain words that reoccurred regularly in academic texts and called for the creation of a list to facilitate the learning of these words. Starting in the 1970s, researchers answered this call and Campion and Elley (1971) and Praninskas (1972) both developed word lists by manually counting the words that occurred in a corpus of academic texts. The first computer-based study of academic vocabulary is Coxhead's (2000) academic word list. Since then, there have been several other studies that have tried to employ more modern techniques to develop an academic word list (e.g., Browne et al., 2013; Gardner & Davies, 2014). While these studies are important, the target audience for these lists are not EAL learners but students in pre-university English courses or in the first year of academic study at English-medium universities (P. Nation, 2016). As learners get older, the textbooks they are using in the classroom get more specialized, and contain more specialized words related to their content area. As a result, these learners need support with more than just academic vocabulary, they also need to know the domain-specific specialized words that are being used in their textbooks (Greene & Coxhead, 2015).

Specialized vocabulary, while similar in some ways to academic vocabulary, (for example, both have numerous words with Greek or Latin roots), has several important differences. Whereas academic words are common across a range of different academic texts and disciplines, technical words are domain specific. Technical words are those words that are important to know within a specific domain but are not as frequent outside of that domain (Nation, 2016). One characteristic of technical vocabulary that can make it hard to recognize when creating a word list is that some high-frequency words can have a technical meaning within a specific domain. Greene and Coxhead (2015) give the example of the word *solution*. This word has a technical meaning in both the sciences and maths, but the meaning is different in both domains. It also has another different meaning when used in conversation. For this reason, learners need to be able to identify and understand the technical vocabulary within each domain. Doing so makes it easier for learners to understand the meaning of that word in its specific domain, and teachers to teach the word in context.

What therefore needs addressing is determining how important such different words are for comprehension. In most academic textbooks, these words can make up over 25% of the running words of the text (Green & Lambert, 2018), which would make these texts difficult for EAL learners to understand without knowledge of these words. To provide an example of what this would look like for an EAL learner with knowledge of the first 2,000 most frequent words (which some EAL learners do not have, see Brooks et al., 2021; Coxhead & Boutorwick, 2018) the following shows a passage from a science textbook with the academic and technical words replaced with nonsense words of the same length.

Figure 1.1

An Example of a Passage Taken from a Biology Textbook with Academic and Specialized Words Replaced with Nonsense Words

> The **meotyla crachri** of soil depends on several **fritol**. One **fritol** is the rock from which the soil was formed. Another **fritol** is the size of the **pockmarf** which make up the soil. **Motylas** (and water) **drecu** rapidly through sand (large **pockmarf**) and are held by clay (small **pockmarfs**). Since **cuvi pockmarfs** carry a **nieceroo** charge, they bind positively charged **pockmarfs** such as Ca2+, K+ and Mg2+.
>
> The most important **fritols** affecting **meotyla crachri** of the soil are **bugachomem fritols**. An undisturbed environment will **rawerla nieceroos veo** soil, plants, animals and **microobtrosons**. **Doonfiskan** can quickly **ductua** the soil when the **nieceroos** removed are not replaced (**kayingraci**). In some countries too much **nidnesto** is put on the land resulting in a very high **meotyla crachri**. The **meotylas** will **leniun** into the streams and lakes and cause **eggetedlation** which is rapid growth of **idnesto** which **ductuas** the **ayingrac** in the water when they die and are broken down by **bayingra**.

As is clear from this passage, it would be difficult for a learner to understand the contents of this passage without knowing the specific academic and technical

vocabulary. What therefore needs determining is, if a teacher wants to support students in learning these words, what resources are currently available to help them do this.

1.3.3 What support is currently available to EAL learners?

Despite the increasing number of both academic and specific word lists, in fields as diverse as medicine, applied linguistics and agriculture (P. Nation, 2016), as well as lists for middle school age children (Greene & Coxhead, 2015), and university age students (Coxhead, 2000; Gardner & Davies, 2014), at the moment, there are no word lists that have been designed specifically for EAL learners. Because of this, teachers who want to support EAL learners with their vocabulary acquisition are forced to do so by either using existing general word lists (e.g., the General Service List (GSL), West, 1953), academic word lists designed for university learners (e.g., the AWL, Coxhead, 2000), or word lists compiled for texts written for FLE learners (e.g., Davies, 2002; P. Nation, 2020). While word lists, such as the GSL and AWL provide extremely good coverage over certain types of texts, for example Coxhead (2000) found that the GSL and the AWL provided approximately 90% coverage for university level academic texts, they do not provide the same coverage over the textbooks that EAL learners are required to read in the classroom (Coxhead et al., 2010). An EAL word list is, therefore, essential since it details the specific vocabulary that this group of learners needs to comprehend the texts that they are required to read to succeed academically.

1.4 The International School Word List: What can we expect?

There are several criteria that a word list designed for EAL learners would have to meet. It would have to provide better coverage of the textbooks that they are using than existing word lists, it would have to include high-frequency words with specialized academic meanings and domain specific words that this group of learners are likely not to know, and it would have to be discipline-specific. Looking at existing word lists for the middle- and high school levels (e.g., Green & Lambert, 2018; Greene & Coxhead, 2015), we can guess that these lists would contain between 400 and 800 headwords and provide coverage of close to 10% or higher of a representative corpus of texts in each of the disciplines being taught in this context. Ideally, we would want these lists, when combined with the words that EAL learners at this level are already likely to know, to provide 95% coverage (a number that will be discussed in more detail in the Chapter 2)

of the textbooks within a discipline. However, in a study that looked at the vocabulary profiles of science textbooks used in Grades 9-12 in New Zealand, Coxhead, Stevens, and Tinkle (2010) found that even Grade 9 learners needed to know over 9,000 words in order to read a science textbook. Such findings suggest it may be difficult to compile word lists for the different disciplines that provide coverage as high as 95%. However, if we can develop a set of lists that provide between 10% and 20% of coverage for each of the discipline specific lists, this would go a long way in addressing the gaps that exist in EAL vocabulary instruction.

1.5 Conclusion and objectives

The three obvious questions that can be inferred from the discussion above are:

1. What gaps exist in EAL learners' vocabulary knowledge?
2. How do these gaps affect the ability of EAL learners to succeed in the classroom?
3. Exactly what vocabulary do EAL learners need to bridge these gaps to succeed academically in an EMI learning environment?

In this thesis, I will try to bring these three research strands together to provide EAL learners with the support they evidently need in the classroom. I would first like to focus on the initial two questions, and to examine the importance of vocabulary for EAL learners' academic success, and to then look at the gaps that exist in their vocabulary knowledge. In doing so, I hope to set the foundations for the second part of this thesis, developing a set of domain-specific word lists designed to help EAL learners succeed in the classroom. To do this, I will need to identify the textbooks that these learners are required to read and to look at the vocabulary profiles of these textbooks.

In the next chapter, I lay the groundwork necessary to answer these three questions. I then follow this with an in-depth literature review that looks at the most important studies in the areas of vocabulary research and corpus and word list development. In doing so, I pay particular attention to research that is focused on better understanding the vocabulary requirements of the EAL classroom and the development of academic, rather than general or specialized, word lists.

Chapter Two

Literature Review

2.1 Introduction

This chapter follows on from Chapter One by looking at a selection of studies that highlight the work that has been done in the fields of developing and validating word lists in both the EAL and general academic context. It is widely acknowledged in the research that vocabulary knowledge is integral to success in all language skills (Daller et al., 2007; Meara, 1996; P. Nation & Webb, 2011). Vocabulary is critical for EAL learners studying in an EMI context and has been shown to be one of the major predictors of school performance (Long & Richards, 2007). However, because principled vocabulary instruction requires teachers to focus on the most useful words first (Schmitt & Schmitt, 2020) to provide support for EAL learners in the classroom, it is first necessary to identify what vocabulary this group of learners need to succeed. To do that, the corpus from which any EAL word list is constructed should represent the words EAL learners are likely to encounter in their daily studies (Greene & Coxhead, 2015; P. Nation, 2016)

Understanding the academic language this group of learners requires to succeed academically is vital. Cummins (1984) describes the distinction that exists between two types of linguistic skills, both of which are essential for learners in the classroom. Basic Interpersonal Communication Skills (BICS) are the language skills learners acquire in familiar situations where the situation provides a range of cues that can assist with understanding, while Cognitive Academic Language Proficiency (CALP) represents an understanding of the type of formal academic language that is commonplace in academic textbooks and lectures (Leung, 2014). Even proficient EAL learners can struggle with CALP for two reasons (Cummins & Yee-Fun, 2007): firstly because of the nature of CALP there are fewer opportunities for EAL learners to acquire it outside of the classroom setting and secondly the types of vocabulary and grammar that make up this type of language become progressively more difficult as learners progress through the grades. However, because CALP is strongly associated with academic progress (Cummins, 2008) it is important that EAL learners are given the support that they need to improve this skill in the classroom. An essential component of CALP is a familiarity

with academic vocabulary, especially the type of vocabulary being used in textbooks and lectures (Leung, 2014).

However, even though academic vocabulary has long been viewed as integral to EAL learners' academic success (Ardasheva & Tretter, 2017; Coxhead & White, 2012), there have been few studies that have looked at what comprises academic vocabulary in the EAL context. Traditionally, the word lists used to develop materials and assessment tools for EAL learners are often lists of general or academic vocabulary compiled from corpora of texts aimed at L1 English speakers, often university students (Green & Lambert, 2018). These general academic word lists are usually used for the sake of expedience. There are fewer academic word lists that focus specifically on the secondary school context, and the ones that exist have been compiled fairly recently. However, the use of general academic word lists as a pedagogical tool for secondary school EAL learners is problematic because they have not been validated in this context. Read (2000) and others caution against over-generalizing the use of pedagogical tools and stress the need to validate these tools before using them outside of the context within which they were developed. This is especially true for academic and specialized word lists, as what is identified as important by any given word list is likely to heavily depend on the corpora from which the word list was derived (P. Nation, 2016). If such general academic word lists are to be used in the EAL classroom, it is important to better understand how well these lists actually measure the vocabulary learners would be likely to encounter in the EAL context. There is also a pressing need to develop EAL-specific word lists to provide better lexical support for this group of learners (Coxhead & White, 2012).

To help us better understand what has been done to date to address these two needs, this chapter first examines how our knowledge of the relationship between vocabulary knowledge and academic proficiency in the EAL context has developed. It then describes how the corpora and word lists that are used in the EAL context have evolved. I have therefore divided the literature review in this chapter into two distinct sections. The first section reviews some important studies that have helped to illustrate why knowledge of vocabulary is important for both EAL and English language learners and to examine how existing word lists can help teachers and researchers better understand how to support EAL learners vocabulary development. The second section

traces the development of academic word lists that have been used in the EAL context, initially as general academic word lists and then as EAL-specific word lists.

To illustrate the influence that the early research has had on subsequent studies, I present the studies in each section in chronological order. This is of particular importance for the second section, which focuses on the development of word lists, as changes in corpus linguistic research brought about by advances in computer and software technology have had a significant impact on how word lists are developed (P. Nation, 2016). In the first section, I have reviewed studies that provide important information about the type of lexical threshold needed for reading including the vocabulary size and coverage required for comprehension. I also investigate what these studies tell us about how word lists can determine what level of vocabulary knowledge learners need to reach the threshold necessary for the comprehension of different texts. In the second section, the literature review first focuses on the gradual development and improvement of general academic English corpora and word lists and then shows how the techniques used for general academic word lists have been used to identify the lexis needed by learners in the junior high and high school classroom setting. The initial focus on general academic word lists is necessary for two reasons. First, as noted above, it is only recently that researchers have developed corpora and word lists that focus specifically on the type of texts that EAL learners are likely to encounter (Green & Lambert, 2018; Greene & Coxhead, 2015). Second, there are few word lists for learners studying in the middle school or high school contexts, and the ones that exist have only recently started to gain traction; for example, Green and Coxhead's (2015) Middle School Vocabulary List was only added to Lextutor (www.lextutor.ca), a popular online tool that enables users to find the vocabulary profile of texts for research and pedagogical purposes, in January 2018. General academic word lists are still commonly used for assessing and developing materials for EAL learners (Coxhead, 2011).

While not all of the ten papers covered in this section focus exclusively on the EAL context, all have helped to shape our understanding of what vocabulary EAL learners need to succeed and why this vocabulary is important. The first section begins with Nation's (2006) and Coxhead et al.'s (2010) investigation of the coverage provided by different word lists for different types of texts, including the text EAL learners are likely to encounter in the classroom. In the next paper, Laufer & Ravenhorst-Kalovski

(2010) investigate the question of what percentage of words in a text a learner would need to know to understand a given text. Finally, Coxhead and Boutorwick (2018) bring these two threads together in a study that investigates the vocabulary development of EAL learners and examines the lexical coverage that this vocabulary knowledge would provide them for the materials that they are reading in the classroom. The second section shifts the focus to the development of word lists. In this part of the chapter, I first examine the increasingly sophisticated techniques that Xue and Nation (1984), Coxhead (2000), Gardner and Davies (2013), and Lei & Liu (2016) used to develop first general academic, then specific academic word lists. I then look at how these techniques have been used to develop word lists for use in the middle school (Greene & Coxhead, 2015) and secondary school (Green & Lambert, 2018) classroom context. The first of these by Greene and Coxhead (2015), describes the development of a corpus and word list that focuses on middle school EAL learners. Finally, Green and Lambert's (2018) article describes the development of a set of EAL word lists compiled for learners studying in a secondary school EMI context in Singapore. For each of these papers, the review comprises a summary of the original study, followed by my commentary on the paper. The chapter finishes with an outline of the main findings of these studies and a discussion of the important issues raised, along with a synopsis of the questions that still need to be answered.

2.2 Review of selected studies on vocabulary coverage and comprehension

2.2.1 Nation (2006): How Large a Vocabulary Is Needed for Reading and Listening?

Introduction

One important area of research regarding the vocabulary needs of EAL learners is investigating how many words they need to know (Coxhead, 2012). In the past, studies have tried to answer this question by looking at either how many words there are in English (e.g., Goulden et al., 1990; Nagy & Anderson, 1984) or by trying to determine how many words we could expect an average native speaker of English to know (e.g., Goulden et al., 1990; Zechmeister et al., 1995). However, both approaches provide figures that are too large to be a reasonable goal for second language (L2) learners. A more appropriate measure would be to look at the number of words that learners will

need to be familiar with to understand the texts that they are regularly expected to engage with in the classroom, an approach taken by earlier studies in the field (Hirsh & Nation, 1992). However, these earlier studies were limited by the vocabulary lists that were available to researchers at the time. Nation's (2006) study, which is still the most-cited article outlining the lexical requirements of English (Schmitt & Schmitt, 2020), is unique because it is one of the first studies to use modern word lists to determine the size of vocabulary a learner would need to read different genres of texts. Nation (2006) used the different subcorpora of the British National Corpus (BNC, Leech et al., 2014) to determine how many word families there are in the three text types learners would commonly engage with in the L2, novels, newspapers, and movies. By looking at these three text types, Nation (2006) concludes that L2 speakers require knowledge of between 8,000 and 9,000 word families to understand these texts.

Summary

Building on the research by Hu and Nation (2000) and Kurnia (2003) showing that approximately 98% coverage was necessary for learners to understand a written text, and roughly the same amount of coverage was necessary for spoken texts (Adolphs & Schmitt, 2003). Nation (2006) uses the BNC in his study to investigate the number of word families required to achieve this level of coverage for three different text types. According to Nation, the focus of his research is two-fold: first, to use the BNC to better understand the vocabulary size needed to understand a variety of text types and second, to act as a trial for word-family lists developed from the BNC to see if they can accurately estimate the number of word families learners need to know.

In the first part of his study, Nation used the BNC to develop fourteen 1,000-word-family frequency lists. The word-family lists used in this study were based on the word type and lemma lists compiled from the BNC. As the original lists were ordered using the frequency of the lemmas, Nation checked the order of the items in the word-family lists by looking at the frequency data obtained by running the fourteen lists over a corpus made up of nine different spoken and written corpora: LOB, FLOB, Brown, Frown, Kohlapur, Macquarie, Wellington written, Wellington spoken, and LUND (these corpora are all freely available at http://gandalf.aksis.uib.no/icame.html). With one small exception, the process of running the fourteen lists over the nine different corpora

showed that each frequency band of the newly created BNC word-family lists accounted for increasingly fewer tokens in the nine corpora. The only exception was that the tenth band of the BNC word families levels provided slightly higher coverage of the LOB than the ninth band. However, the difference was tiny, with the tenth accounting for 3,328 tokens compared to 3,217 for the ninth, and this only occurred when looking at the tokens, and not families. The lists were further validated by looking at the total number of types in each list. Because lower-frequency words have fewer family members, the number of types should decrease as you move from high-frequency to low-frequency lists, which is what occurs in these lists. Finally, the word lists were compared with the nine corpora again to see if there were any missing high-frequency words or word families. While Nation added several family members to the lists, for example *reds* to *red*, no additional families were added.

In the second part of the study, the word lists developed using the BNC were used to determine the lexical knowledge required to understand three different types of texts, novels, newspapers, and movies. For novels, Nation looked at the text coverage of five different novels (see Table 2.1). While the amount of coverage of the different frequency bands varied slightly, from 7,000, for *Turn of the Screw*, to 9,000, for *Lady Chatterley's Lover*, words were needed to reach the 98% coverage necessary for understanding (Hu & Nation, 2000). By combining all the novels into a single corpus, Nation found that the first 2,000 word families provided 87.83% coverage, 4,000 plus proper nouns provided 94.8% coverage, and 9,000 plus proper nouns provided 98.24% coverage.

Table 2.1

The Lexical Coverage of the BNC in Several Novels

Word list	Lord Jim (%)	Lady Ch. (%)	Screw (%)	Gatsby (%)	Tono-Bungay (%)	Average
2,000	87.29	88.09	91.71	87.71	86.95	88.35
4,000 + proper nouns	94.24	95.06	96.08	95.02	94.36	94.95
9,000 + proper nouns	98.06	98.22	98.52	98.47	98.00	98.25
Proper nouns	1.04	2.05	0.50	2.12	1.55	1.45

Note. Adapted from "How large a vocabulary is needed for reading and listening," by P. Nation, 2006, *Canadian Modern Language Review*, *63*(1), p. 71. (https://doi.org/10.3138/cmlr.63.1.59)

Nation next looked at the vocabulary coverage of newspapers. To do this, he compiled a corpus of newspaper articles using the section entitled Reportage from the LOB, FLOB, Brown, Frown, and Kolaphur corpora. In each of these corpora, the Reportage section comprised forty-four 2,000-token collections of news articles. The five resulting newspaper corpora that Nation compiled each comprises 88,000 tokens. Nation then used Range to determine what coverage the word lists developed using the BNC provided for each of the newspaper corpora. I show the resulting coverage figures in Table 2.2.

Table 2.2

Text Coverage in of Five Newspaper Corpora by the BNC Word-Family List

Word List	LOB (%)	FLOB (%)	Brown (%)	Frown (%)	Kolaphur (%)	Average
2,000	84.33	83.07	81.54	81.79	84.15	82.98
4,000 + proper nouns	95.39	95.10	94.14	93.93	94.64	94.64
8,000 + proper nouns	98.31	98.03	97.60	97.28	98.05	97.85
Proper nouns	5.29	5.66	6.12	5.43	4.55	5.41

Note. Adapted from "How large a vocabulary is needed for reading and listening," by P. Nation, 2006, *Canadian Modern Language Review*, *63*(1), p. 72. (https://doi.org/10.3138/cmlr.63.1.59)

Using the average from both tables, the results show that the most common 2,000 word families provide better coverage in novels than newspapers, 88.35% compared to 82.98%. However, there were significantly more proper nouns in the newspaper corpora, between 4.55% and 6.12% of all tokens compared to between 0.5% and 2.12% for the novels. As a result, the combination of proper nouns and the 4,000 and 8,000 most common words provided comparable coverage across both types of corpora. As with novels, readers would have to be familiar with the first 8,000 most frequent word families to reach the 98% coverage necessary for comprehension.

The next genre that Nation looked at was graded readers. Graded readers are books that have been specifically written for learners of English as a foreign language and are different from other books because they are written in a way that strictly controls the level of vocabulary and grammar used in the books (P. Nation & Waring, 2019). Nation looked at the coverage of the BNC word lists in a single graded reader, Oxford Bookworms Series Level 3 reader The Picture of Dorian Gray. He determined that knowledge of the 3,000 most frequent word families, along with proper nouns, provided a coverage of 98.86%, or a sufficient coverage to understand the book. Of the

total 10,578 words in the book, there were only 20 words from the 4,000- to 9,000-word bands and none from the 9,000-word band onward.

The final genre included in Nation's study was children's movies. He included movies to allow for the analysis of spoken as well as written texts. We know from other studies that spoken English can have a different vocabulary profile from written English (Dang et al., 2017) so it was important for Nation to also look at this type of text. For his analysis, Nation chose the children's movie *Shrek*. This was done by converting the script into a text file, removing the stage directions, which would not have been spoken in the actual movie, and then analyzing the remaining words against the BNC word list. The corpus consisted of almost 10,000 tokens, which comprised about 1,100 word families (see Table 2.3).

Table 2.3

Cumulative Percentage Coverage Figures for Shrek by Word Families From the BNC

Word list (1,000)	Coverage including proper nouns (%)	Coverage without proper nouns (%)
1	81.54	83.01
2	86.44	87.91
3	92.81	94.28
4	95.27	96.74
5	96.15	97.62
6	96.48	97.95
7	96.61	98.08
8	96.74	98.21
9	96.90	98.37
10	96.98	98.45
11	97.07	98.54
12	97.14	98.61
13	97.45	98.92
14	97.78	99.25
Not in the lists	98.53	100.00

Note. Adapted from "How large a vocabulary is needed for reading and listening," by P. Nation, 2006, *Canadian Modern Language Review*, *63*(1), p. 74. (https://doi.org/10.3138/cmlr.63.1.59)

Nation found that with knowledge of the first 7,000 most frequent words, along with proper nouns, which account for 1.47% of the running words in Shrek, one could understand 98.08% of the words in the movie. This means that there would be one unknown word for every 50 words in the movie. While this is low enough to allow for

learners to guess the meaning from context, that a movie comprises spoken text means learners do not have as much time as they would have with a written text to guess the meaning of the word. Of course, as Nation points out, movies provide strong visual support so it would be possible to enjoy a movie like Shrek even if one did not understand all the words. Nation also found that when he compared the word families found in Shrek to those found in Toy Story, the words outside of the most frequent 3,000 were very different between the two movies. This means that the mid to low-frequency words learners would need to watch one movie would probably not help them understand a different movie.

Nation concludes that to achieve 98% coverage of spoken and written text learners would be required to understand the first 6,000 to 7,000 most frequent word families from the BNC corpus for spoken texts and the first 8,000 to 9,000 for written texts. There is also a lot of variation in the coverage of the first 1,000 word families plus proper nouns between spoken and written texts, with a greater percentage of spoken texts being within the first 1,000 frequency band. He also notes that for reading some genres, such as newspapers, learners would benefit from knowledge of the Academic Word List (Coxhead, 2000); there is, however, a significant overlap between the BNC and AWL (Neufeld et al., 2011).

Comment

Nation's (2006) study contributes towards our understanding of the importance of vocabulary for second language learners. It represents one of the first studies to use frequency lists developed from a large corpus using modern techniques to determine how many words a learner would have to know to understand a text. Nation's examination of different text genres is useful because it serves to highlight the similarities and differences between these text types, including the difference in the coverage provided by high-frequency words in spoken compared to written texts. It also helps show why certain authentic texts, such as newspapers and novels, may be extremely difficult for EFL, or EAL, learners to understand. However, despite the importance of the study and its pedagogical implications, the method and the conclusions Nation draws from his findings need to be examined.

Primarily, Nation's corpus of graded readers and movies comprises only a single text; even at the time of the study, this would have been an extremely limited sample size, making it difficult to generalize from the results. It is also unclear why Nation chose a Level 3 graded reader from the Oxford Bookworm collection for his analysis of this genre. Because only one level from one publisher was analyzed it is hard to determine how graded readers from other publishers, or at different levels, would differ from the one analyzed in the study.

While the newspaper and novel corpora are larger and more diverse, each subcorpora in these genres was examined individually, and no affordance was given in the statistical analysis to examine similarities and differences between the subcorpora to justify the generalizations made. In addition, the analysis presented in the paper is extremely limited: for example, although Nation examined a second movie as means of comparing the mid to low-frequency words between that movie and Shrek, he did not present any other statistical analysis about the coverage the BNC provided for this movie. Also, while Nation produced concordances for 43 items, he only discussed three items.

While the above discussion illustrates several potential weaknesses with Nation's study, it does not show that the conclusions he reached are incorrect. The importance of this study is that it shows how modern word lists can analyze the comprehensibility of different texts. It also has significant pedagogical implications, both in the texts teachers should choose for a certain group of learners, and the vocabulary items that these learners need to focus on to understand what they are reading. However, more data are needed to better understand the lexical coverage of different word lists for different genre types.

2.2.2 Coxhead, Stevens, & Tinkle (2010): Why Might Secondary Science Textbooks Be Difficult to Read?

Introduction

Nation's (2006) study of the lexical coverage of different frequency levels of the BNC word lists across different texts is important, since it shows that word lists can be used to assess spoken and written texts to determine the amount of vocabulary a learner would have to know to understand these texts. There have been several follow-up

studies that have looked in more depth at the lexical coverage of these word lists across different texts. This section will review one of these studies (Coxhead et al., 2010) that looks specifically at the lexical coverage of high school level science textbooks. While the corpus size of this study is quite small at only 279,733 tokens, it is the only study to date that has examined the lexical knowledge that learners studying high school science in the EMI context would require to understand the textbooks that they are reading using word lists compiled with modern techniques.

Summary

Coxhead et al. (2010) focussed on the vocabulary of four high school level textbooks that were commonly used in classrooms in New Zealand. The study uses four different word lists to examine the difference in coverage between these lists. These lists are West's (1953) GSL, Coxhead's (2000) AWL, Coxhead and Hirsh's (2007) pilot science list for EAP, along with Nation's (2006) BNC frequency lists that were described in the previous review. Building on Nation's (2006) study, Coxhead et al. use these lists to ascertain how much vocabulary a learner would need to understand a high school level science textbook and to determine which word lists provide the best coverage of these texts. In doing so, their aim was to help classroom teachers make pedagogical choices as to how much and what vocabulary to teach their learners. Because textbooks are such an important part of the learning environment of the EMI classroom and because textbook makers often provide little or no vocabulary support for non-FLE learners (Harmon et al., 2000), it is important for teachers to provide the lexical scaffolding that EAL learners need to understand these texts. Taylor's (1979) previous corpus-based study of secondary textbooks in Australia examined textbooks across six different subjects, maths, science, history, commerce, social studies, and geography to determine what about these texts would make them hard for migrants and non-FLE learners to understand. While she found that science books are likely to be more difficult to understand in terms of vocabulary, she did not have access to the types of vocabulary lists that Coxhead et al. were able to use in their study.

In the first part of their study, Coxhead et al. scanned and used optical character recognition (OCR) to extract the text from grades nine, ten, eleven, and twelve of the New Zealand Pathfinder science textbook series (Hook, 2004; 2005; 2006; Relph,

Croucher, & Castle, 2006). Because of the nature of the textbooks, there were several issues with the OCR process, including some colours of text not being recognized properly and complex page layouts leading to the text being in the incorrect order, which they fixed manually. They also removed any hyphenated words by replacing hyphens with a space. The total running-words of the four corpora was 279,733, with the subcorpora running from 56,058 running-words, for the grade twelve textbook, to 88,685 running-words, for the grade eleven textbook. They then analyzed the textbooks using the Range Program (Heatley et al., 2004), using two different sets of word lists. The first analysis comprised the GSL, the AWL, and Coxhead and Hirsch's pilot science-specific word list, while the second comprised the first 20,000-word families of Nation's (2006) BNC frequency list. The authors noted that, because there was a significant difference in the total of running words between the different textbooks, it made some of the analysis difficult.

With the first analysis, Coxhead et al. (2010) found that the GSL, AWL, and Science-specific list, along with proper nouns, provided just over 90% coverage (see Table 2.4) of the corpus of science texts. About 76.96% of this coverage came from the GSL, while 7.05% and 5.90% came from the AWL and Science-specific word lists, respectively. The 76.96% coverage provided by the GSL of the corpus of science textbooks was almost 6% higher than the coverage the GSL was found to provide for university-level science texts (Coxhead & Hirsch, 2007). A lot of this coverage was provided by a tiny portion of high-frequency vocabulary that has science-specific meaning, something that will be covered in more detail later when discussing the benefits of including specialized meanings of high-frequency words in a domain-specific corpus (see, for example, Green & Lambert, 2018; Lei & Liu, 2016). The science-specific word list was also found to provide greater coverage of the scientific texts in this corpus than it did on a tertiary level corpus of scientific texts (Coxhead & Hirsch, 2007). The AWL only provided 7.05% coverage, which is lower than the 9.1% coverage the AWL provided for the science section of the AWL corpus. This supports Greene and Coxhead's (2015) finding that secondary school textbooks do not contain the same percentage of general academic words as university texts, and supports the need for an academic word list that caters specifically to this group of learners.

Table 2.4

Coverage of the Four Science Textbooks by the GSL, AWL, and Science-Specific Word Lists

Word list	Tokens/%	Running Coverage	Families
GSL 1000	70.10	70.10	886
GSL 2000	6.86	76.96	645
AWL	7.05	84.01	412
Science-specific list	5.90	89.91	264
Proper nouns	0.48	90.39	269
Not in the lists	9.61	100	N/A

Note. Adapted from "Why might secondary science textbooks be difficult to read," by A. Coxhead et al., 2010, *New Zealand Studies in Applied Linguistics*, *16*(2), p. 44.

One can see from this analysis that even for learners who are familiar with the vocabulary from all these word lists, secondary science textbooks would be difficult to read, because 10% of the words in the textbooks were not found in any of the four word lists used in the analysis. Many of these off-list words were not general low-frequency vocabulary, but words that are connected to science, such as *crust*, *planet*, *friction* and *genes*, and would have been important for learners to know in order to understand the textbooks. In total, these four lists provided 90.39% coverage of the text, which falls well below the 98% coverage necessary for understanding (Hirsh & Nation, 1992).

Coxhead et al. next looked at the coverage provided by the BNC word families. Table 2.5 summarizes the coverage provided by different frequency bands of the BNC. The first two 1,000 frequency bands provide 81.03% coverage of the textbooks, which is comparable to the 83% coverage Nation (2006) found for newspapers, but lower than he found these high-frequency words provided for both graded readers (91%) and novels (88%). Furthermore, it is not until one reaches the 14,000-frequency band that the BNC, along with proper nouns, provides the 98% coverage needed for

comprehension, which is much lower than the coverage Nation (2006) found the BNC provided for both novels and newspapers, which only required the first 9,000 and 8,000 frequency bands, along with proper nouns, to achieve at least 98% coverage.

Table 2.5

Coverage of All the Science Textbooks by the BNC Lists

Word list	Tokens/ Percentage	Families
2,000	81.03	1602
4,000 + proper nouns	92	2851
9,000 + proper nouns	96.5	3842
14,000 + proper nouns	98.07	4274
Proper nouns	0.48	269
Not in the lists	1.23	????

Note. Adapted from "Why might secondary science textbooks be difficult to read," by A. Coxhead et al., 2010, *New Zealand Studies in Applied Linguistics*, *16*(2), p. 47.

Coxhead et al. also investigated the differences in the coverage provided by the different word lists for the different years. They found that the GSL/AWL/Science-specific word lists provided decreasing coverage across the different grade levels (see table 2.6) with the best coverage provided by these lists over the grade 9 subcorpora with 92.64% coverage and the lowest coverage over the grade 12 subcorpora where they only provided 88.50%. One interesting point to note is that with the GSL, the lowest coverage was found with the grade 11 subcorpora, where the GSL 1000 and 2000 provided only 69.64% and 6.16% coverage compared to the 70.08% and 6.81% coverage they provided over the grade 12 subcorpora.

Table 2.6

Coverage of the GSL/AWL/ Science-Specific Lists Over the Four Individual Textbooks

Text	Year 9	RT	Year 10	RT	Year 11	RT	Year 12	RT	Avg.
GSL 1000	71.03	71.03	69.88	69.88	69.64	69.64	70.08	70.08	70.1
GSL 2000	7.82	78.85	6.91	76.79	6.16	75.80	6.81	76.89	6.9
AWL	7.25	86.10	6.98	83.77	7.48	83.28	6.25	83.14	6.9
Science-specific list	6.14	92.24	6.45	90.22	6.13	89.41	4.58	87.82	5.8
Proper nouns	0.37	92.61	0.38	90.60	0.44	89.85	0.78	88.50	0.49
Not in any list	7.39	100	9.40	100.00	10.15	100.00	11.50	100.00	10.1

Note. RT = Running Totals. Adapted from "Why might secondary science textbooks be difficult to read," by A. Coxhead et al., 2010, *New Zealand Studies in Applied Linguistics*, *16*(2), p. 46.

The coverage provided by the BNC word lists followed a similar pattern (see Table 2.7). Coxhead et al. found that the BNC frequency lists provided gradually decreasing coverage over the four grade levels, with the highest coverage at the grade nine level and the lowest at the grade 12 level. One interesting point that they found was that even the 15,000-frequency band plus proper nouns would not have provided the 98% coverage required for comprehension. At this grade level, 2.08% of the words were off-list, meaning that they were not frequent enough in the BNC corpus to be included in Nation's BNC word-family frequency lists, which cover the 15,000 most frequent word families in English. These included words like *gondwana* and *tuatara* that even first-language English speakers would be likely to struggle with.

Table 2.7

Text Coverage of the Four Science Textbooks by the BNC

Word list	Year 9	Year 10	Year 11	Year 12
2,000	82.27	81.29	80.53	80.09
4,000 + proper nouns	93.07	92.27	91.67	90.96
9,000 + proper nouns	97.28	96.81	96.32	95.52
11,000 + proper nouns	98.08	98.17	97.54	96.60
15,000 + proper nouns	98.72	98.74	98.17	97.18
Proper nouns	0.37	0.38	0.44	0.78
Not in the lists	0.87	0.75	1.33	2.08

Note. Adapted from "Why might secondary science textbooks be difficult to read," by A. Coxhead et al., 2010, *New Zealand Studies in Applied Linguistics*, *16*(2), p. 47.

In their study, Coxhead et al. found that the vocabulary content of textbooks would make them difficult for EAL learners, and even for FLE learners who rarely reach mastery of the 15,000 and higher word frequency bands until late in secondary school, if at all (Coxhead & Boutorwick, 2018; Coxhead et al., 2015). To read and understand a science textbook, learners would have to know at least 3,000 more words than they would need to read a novel in English. There is also a considerable increase in the number of words required to reach 98% coverage. Learners could reach this coverage with 11,000 words plus proper nouns in the first two years but would be required to know over 15,000 words plus proper nouns in the final year. Given the amount of vocabulary needed to understand these texts, it is not surprising that EAL learners have been found to struggle with understanding science textbooks and, consequently, struggle academically in these classes (Ardasheva & Tretter, 2017; Miller, 2009).

Comment

Coxhead et al.'s (2010) study is important because it is the first, and still one of the few, studies that have used modern frequency lists to examine why EAL learners may have difficulty reading and understanding classroom textbooks. The authors made use of the main word frequency lists available to them at the time and were able to correlate their findings to the findings of other studies, such as Nation (2006), to show that textbooks not only contain many low-frequency words but also that they may even be more difficult to read than other genres, like newspapers or novels. It also helps to indicate why just focusing on existing word frequency lists may not be enough. None of the existing frequency lists, for the time, could provide the 98% coverage necessary for learners to understand the grade 12 science textbook used in the study. This has important pedagogical and research implications. However, despite the importance of this study, there are several problems that need to be addressed in future studies.

The most obvious issue is the focus of this study on a single subject and a single set of textbooks. This issue is especially important to address if one wants to extend this study to schools outside of New Zealand. The authors themselves noted that some of the off-list words in the grade 12 textbook included the names of native New Zealand animals, which are words that would probably not be important for reading textbooks written for a different context. Including science books aimed at an international, or North American audience, may have resulted in a lower percentage of off-list words. On the other hand, Coxhead et al.'s use of general science textbooks, instead of domain-specific textbooks, such as biology or chemistry textbooks, may have resulted in including fewer domain-specific technical words in the corpus, which may have increased the coverage of high and mid-frequency vocabulary items.

Using word lists that focus on word families, as opposed to lemmas, may have also led to an under-counting of the words required to understand these texts. Word families, especially word families that include technical words, can often exhibit a large difference between the various family members, making it impossible for a learner to recognize the technical words as being part of the word family. For example, the word *particle*, as it is used in a physics textbook to refer to a particle of matter, is a member of the word family that has *part* as its headword. However, just because a learner can

recognize *part* in a sentence such as "I am part of this class." does not mean they would be able to recognize the word *particle* in a science textbook. If learners cannot recognize these technical words as members of a specific word family, knowledge of that word family would not be sufficient for them to understand the technical word, resulting in lower coverage of the corpus than Coxhead et al.'s analysis using the GSL, AWL, or BNC would seem to indicate.

While the above discussion illustrates several weaknesses with Coxhead et al.'s study, their study helps to illustrate why EAL learners may have difficulty understanding science textbooks. This study is also important because the low levels of coverage achieved by the AWL and BNC help to illustrate the need for word lists developed specifically for this type of text. While the AWL provides approximately 10% coverage of university-level academic texts, it only provides 7.05% coverage of the secondary school textbooks examined in this study. Given the importance of this subject for EAL learners, and the difficulties they have been demonstrated to have understanding science texts (Ardasheva & Tretter, 2017; Coxhead & Hirsch, 2007; Miller, 2009) there remains a need for a set of word lists to bridge this gap.

2.2.3 Laufer and Ravenhorst-Kalovski (2010): Lexical threshold revisited: Lexical text coverage, learners' vocabulary size

Introduction

Laufer and Ravenhorst-Kalovski's (2010) study is one of the most extensive to date that tries to answer the question of how many words of a text learners would have to know to understand that text. This study builds upon earlier studies (Hu & Nation, 2000; Laufer, 1989, 1992) and uses measures of reading comprehension, along with an estimate of the number of words in the text the learners would be likely to know, to calculate the coverage necessary for language comprehension. In this study, Laufer and Ravenhorst-Kalovski investigated how the learners' vocabulary size and the coverage that this vocabulary provides relates to reading comprehension. The two thresholds necessary for understanding, 95% for minimum coverage and 98% for optimal coverage, that Laufer and Ravenhorst-Kalovski calculated from this extensive study are often cited in the research (e.g., Qian & Lin, 2019; Zhang & Zhang, 2022). While there are several potential issues with the methodology used in this study, Laufer and

Ravenhorst-Kalovski's study is important because it provided researchers and teachers with a valuable starting point when trying to determine just how many words a certain group of learners needs to know.

Summary

Laufer and Ravenhorst-Kalovski (2010) examined the relationship between the lexical coverage of academic texts, learners' vocabulary knowledge, and their English reading comprehension. The vocabulary scores of 745 students enrolled in an English for Academic Purposes class at an academic college in Israel were correlated with the scores the same students received on a standardized reading comprehension test. The authors used the scores from the reading section of the English Psychometric Exam, a standardized exam required for college entrance in Israel, as a measure of the participants' reading comprehension. The reading section of this exam consists of 60 multiple-choice questions. The maximum score on this section of the test was 150 and the scores of the participants ranged from 75 to 133. Laufer and Ravenhorst-Kalovski measured the participants' vocabulary size using Schmitt et al.'s (2001) revised version of Nation's (1983) Vocabulary Levels Test. The authors used this test to measure the participants' knowledge of the 2,000, 3,000, and 5,000 frequency bands of the BNC. They also used the participants' scores from the 2,000-frequency band to estimate their knowledge of the 1,000 most frequent word families and the scores from the 3,000 and 5,000 frequency bands to estimate their knowledge of the 4,000 frequency band. These scores were added together to provide an overall score for vocabulary proficiency for each of the participants: "for example (if) a learner received 28 on the second 1,000, 22 on the third, and 8 on the fifth, his or her score would be 28+28+22+15+8=101" (Laufer & Ravenhorst-Kalovski, 2010, p. 21). The first 28 in this calculation came from the fact that the learner could master the 2,000-frequency band, so the researchers assumed they could also master the 1,000 level. The 15 in the calculation is the average of the scores the learner received from the 3,000- and 5,000-frequency bands.

Laufer and Ravenhorst-Kalovski then measured the lexical coverage of the reading comprehension test. As the actual tests the participants sat were not available to the researchers, they measured the lexical coverage of the test using three previous versions of the test. The authors created a corpus of 19,037 words using these three tests

and then used the Range frequency analysis software program (Heatley et al., 2004), and Cobb's Lextutor website (http://lextutor.ca) to determine the vocabulary profile of the corpus. The 20,000 most frequent families of the BNC were used for this analysis (see Table 2.8 for the coverage of the different frequency lists).

Table 2.8

Coverage of the English Psychometric Tests by BNC Frequency Lists (Proper Names in the "Off List" for the Initial Coverage Bands)

Frequency level	Coverage % Test 1	Coverage % Test 2	Coverage % Test 3	Average cumulative coverage
K1	80.15	75.91	79.58	78.58
K2	9.39	10.04	7.92	87.67
K3	2.54	3.11	3.24	90.56
K4	2.21	2.58	2.35	92.81
K5	0.74	1.09	1.27	94.00
K6	0.80	1.13	0.66	94.80
K7	0.32	0.48	0.77	95.40
K8	0.46	1.27	0.90	96.30
K9	0.11	0.44	0.14	96.53
K10-K20	1.10	1.00	0.88	97.50
Off list	2.19	2.97	2.32	100.00
Proper names	2.00	2.48	1.8	2.1

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

From this analysis, Laufer and Ravenhorst-Kalovski calculated that when proper nouns were included learners would achieve 95% coverage with knowledge of 4,000 word families and 98% coverage with knowledge of the 7,000–8,000 frequency bands. They then used the participants' scores on the vocabulary levels test to determine which word frequency bands each participant was likely to know and correlated these measures of lexical coverage to the scores that participants received on the reading comprehension test (Table 2.9).

Table 2.9

Vocabulary Size, Lexical Coverage and Reading Comprehension (Maximum Reading Comprehension Score = 150)

Approximate vocabulary size	Lexical coverage	Percentile on the psychometric test	Reading score: Mean (SD)	No. of students
1,000	78.58	50%	83 (6)	109
2,000	87.67	53%	90 (7.8)	199
3,000	90.56	66%	102 (8.9)	204
4,000	92.81	72%	111 (9.4)	200
5,000	94	83%	122 (8.3)	23
6,000	94.8			
7,000	95.43	91%-99%	138 (4)	10
8,000	96.3			

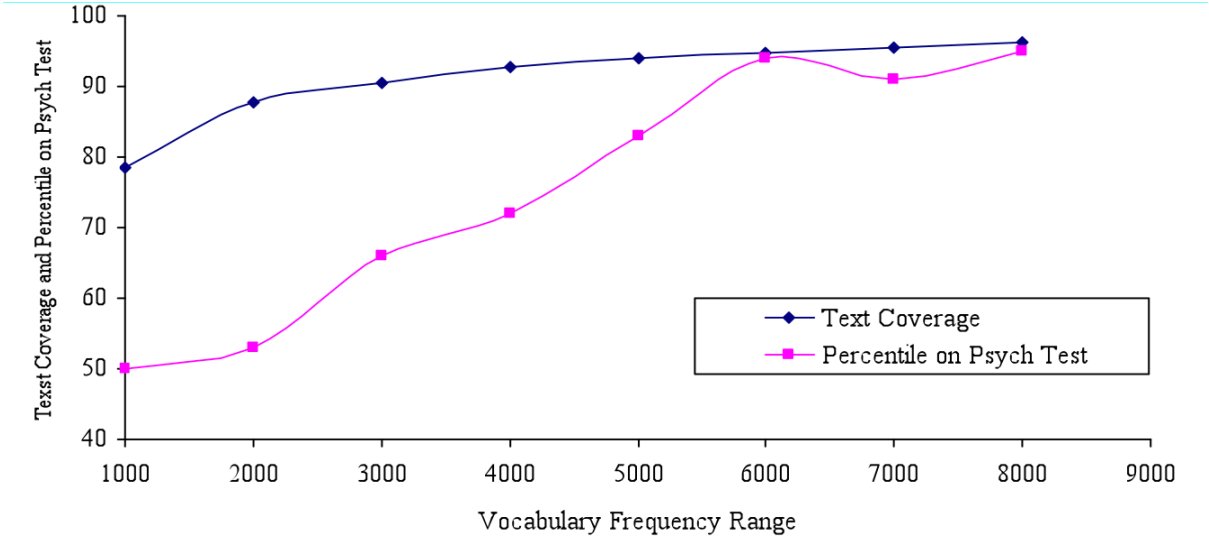
Note. Adapted from "Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension," by B. Laufer and G.C. Ravenhorst-Kalovski, 2010, *Reading in a Foreign Language*, *22*(1), p. 23. (https://doi.org/10125/66648)

They also used a linear regression analysis to determine what percentage of variance in the participants' reading comprehension scores could be attributed to their

vocabulary knowledge. When they excluded the top 10 participants, their analysis showed that 64% of the variance in reading comprehension could be attributed to vocabulary knowledge (see Figure 2.1).

Figure 2.1

Text Coverage and Reading Scores in Relation to Vocabulary Frequency Range



Note. Adapted from "Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension," by B. Laufer and G.C. Ravenhorst-Kalovski, 2010, *Reading in a Foreign Language*, *22*(1), p. 24. (https://doi.org/10125/66648)

Their analysis showed that the lexical coverage provided by knowledge of between 6,000- and 8,000-word families, approximately 98% coverage, was required for learners to achieve a score of 134 or higher, the score necessary for learners to show that they have the ability to read academic material independently. Participants who had knowledge of between 4,000 and 5,000 word families, or about 95.5% coverage with proper nouns, received between a 116-133 score on the reading comprehension test. This score indicates that these students would be able to read academic English texts with support from the teacher. Learners who only had knowledge of the first 3,000 word families, or fewer, were most likely to receive scores of below 116 on the reading

comprehension test, showing that they were likely to struggle reading academic texts. From this, Laufer and Ravenhorst-Kalovski determined that a lexical coverage of 95% was necessary to read an academic text, with assistance, while 98% was the threshold of lexical coverage necessary for learners to read independently. This closely matches the findings of Hu and Nation (2000), who suggested that 98% coverage is the lexical threshold necessary for most learners to read, while 95% coverage is the minimum threshold for acceptable comprehension.

Comment

While an important study, there are some obvious issues with Laufer and Ravenhorst-Kalovski's (2010) methodology. To begin with, they calculated lexical coverage by looking at older versions of the test and extrapolated from those scores to determine the lexical profile of the text that the participants actually sat. However, these previous tests showed significant differences in how much coverage each frequency band provides. For example, in Test 1, 95% coverage was provided by the first 5,000 frequency bands but these same word families only provided 92.73% coverage of Test 2. Even if the authors had used the coverage from the actual test the participants sat, this still would not have given the actual coverage of the participants' vocabulary knowledge, since the participants' vocabulary sizes, and the coverage of the text that these provided, were estimated using a VLT developed from the BNC. The lexical coverage the participants' vocabulary knowledge would have provided was not actually measured directly, which makes it difficult to make any statements about how much of the variance seen in the reading comprehension test can be explained by text coverage.

A second issue with this study is how the authors calculated the participants' vocabulary scores. The VLT they used did not measure above the 5,000-frequency band, so it would have been insufficient to indicate which participants had the vocabulary knowledge necessary for 98% coverage. They also did not measure the 1,000 or 4,000 frequency bands directly, but rather used the participants' scores on the other frequency bands to estimate how many words they would be likely to know of the 1K and 4K frequency bands. As the VLT only contains 30 words for each 1,000 word band, it is unclear how accurate these estimates would have been.

2.2.4 Coxhead & Boutorwick (2018): Longitudinal Vocabulary Development in an EMI International School Context: Learners and Texts in EAL, Maths, and Science

Introduction

From the papers reviewed so far in this chapter, we know learners need to understand between 95% to 98% of a text to understand that text, either with assistance or independently (Laufer & Ravenhorst-Kalovski, 2010). We have also seen that 95% coverage of a high school science textbook requires the first 9,000 word families, while 98% coverage requires 14,000 word families (Coxhead et al., 2010). The question that now needs to be asked is "do EAL learners have the vocabulary knowledge necessary to understand the textbooks that they are using in the classroom?" One large-scale longitudinal study that examines this question is Coxhead and Boutorwick's (2018) paper, which investigates the effects of EAL learners' vocabulary knowledge at an international high school in Germany where English is the Medium of Instruction (EMI). This study looks at the development of EAL learners' vocabulary over time to better understand what effect this may have on their academic performance.

Summary

In their 2018 study, Coxhead and Boutorwick used the VLT (Schmitt et al., 2001) to analyze how the vocabulary knowledge of 467 students studying at an international school in Berlin developed over time. The participants came from over 50 different countries with those of German nationality making up the largest percentage at 43% and English-speaking countries, such as the U.S. and Canada, making up 12% of the participants. They gave the participants their first VLT when they entered Grade 6 and were tested once a year until they graduated. The authors divided the participants into three groups. They first categorized the participants as either Native Speakers (NS) or Non-native Speakers (NNS) based on the participants' nationality. They then created a separate Non-native speaker EAL learner (NNSEAL) group that consisted of those participants in the NNS group that required additional support in the classroom. The students in the NNSEAL group were identified as having lower proficiency in English by their classroom teachers and were enrolled in special classes designed to improve their language skills, including their vocabulary knowledge.

Coxhead and Boutorwick found that the native speaker cohorts were able to master the first 2,000 and 3,000 words by the time they entered Grade 6 and attained mastery of the 5,000-word and AWL by Grade 8 (see Table 2.10). They also were almost able to attain mastery of the 10,000-word level by Grade 10. The NNS came into Grade 6 without having mastered even the first 2,000 most frequent word families. While these participants were able to attain mastery of these words by the end of Grade 6 and the 3,000-word level in Grade 8, they could not master the 5,000-word level and the AWL until Grade 10, and never achieved mastery of the 10,000-word level. The NNSEAL participants, as expected, received the lowest scores on the VLT. This group of learners could not achieve mastery of the 2,000-word level until the start of Grade 9 and took until Grade 11 to master the 3,000-word level and AWL. They also never achieved mastery of the 10,000-word level. A mixed-effects model with the VLT scores as the dependent variable and cohort, year enrolled, and the VLT level as the fixed independent variables, and student and cohort as the random variables, was used to determine trends in the data. This analysis indicated that while all three groups improved their scores over time, they plateaued at different points depending on what group they were in with the NS vocabulary knowledge plateauing first, followed by the NNS and then the NNSEAL.

Table 2.10

Average VLT Scores with Standard Deviations of the Three Groups

Grade and Group	VLT 2,000	VLT 3,000	VLT 5,000	VLT 10,000	VLT AWL
6 NNS	26.0 (3.2)	21.2 (5.8)	15.9 (6.3)	6.5 (5.3)	16.1 (6.1)
6 NNSEAL	16.0 (7.2)	10.2 (6.2)	6.8 (5.2)	2.1 (2.5)	7.3 (5.7)
6 NS	28.8 (1.4)	26.9 (3.3)	23.0 (6.2)	10.2 (7.0)	20.9 (5.2)
7 NNS	28.0 (2.0)	25.6 (3.7)	20.8 (5.1)	9.7 (5.5)	21.8 (4.6)
7 NNSEAL	23.3 (5.1)	17.0 (6.3)	12.5 (5.2)	4.0 (3.4)	12.8 (7.0)
7 NS	29.2 (1.1)	28.7 (2.2)	25.3 (4.8)	16.6 (7.4)	25.3 (4.4)
8 NNS	28.7 (1.3)	27.5 (2.7)	23.3 (4.4)	12.5 (6.0)	24.6 (4.3)
8 NNSEAL	25.2 (4.5)	21.4 (6.2)	15.6 (5.0)	6.2 (4.7)	18.0 (7.0)
8 NS	29.2 (1.1)	28.6 (2.2)	26.9 (3.5)	18.2 (6.5)	27.0 (2.6)
9 NNS	29.3 (1.1)	28.1 (2.4)	24.4 (4.5)	14.1 (6.3)	25.7 (3.7)
9 NNSEAL	27.4 (2.9)	22.9 (5.7)	19.1 (6.2)	8.7 (5.6)	22.3 (5.3)
9 NS	29.8 (0.5)	29.5 (0.8)	28.0 (3.0)	21.7 (6.4)	28.3 (2.5)
10 NNS	29.4 (1.0)	28.7 (1.9)	26.7 (2.9)	16.2 (6.1)	27.8 (2.2)
10 NNSEAL	28.8 (2.2)	25.7 (4.5)	22.9 (3.8)	10.2 (5.2)	25.8 (2.5)
10 NS	29.9 (0.2)	29.8 (0.4)	29.6 (1.0)	25.0 (4.4)	28.9 (1.4)
11 NNS	29.7 (0.8)	29.2 (1.3)	27.5 (2.1)	18.9 (5.6)	28.7 (1.4)
11 NNSEAL	29.6 (0.9)	27.9 (1.9)	25.5 (2.6)	14.4 (4.6)	27.2 (2.6)
11 NS	29.8 (0.5)	29.9 (0.4)	29.4 (0.9)	25.0 (3.7)	29.6 (0.5)

Note. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 598. (https://doi.org/10.1002/tesq.450)

Coxhead and Boutorwick then examined what level of coverage the different group's vocabulary knowledge would give them on the academic texts they were likely to encounter in the classroom. To do this, the authors compiled a corpus of just under 500,000 words made up of textbooks being used at the school. This corpus of textbooks was comprised of three different subject area subcorpora (see Table 2.11): English, Science, and Maths. They scanned these texts into the computer and converted into text files using optical character recognition, a process whereby images of printed or written text are changed into editable text files using a computer. The resulting text files were then checked against the originals and any errors were corrected. Coxhead and Boutorwick also developed a parallel corpus of 3,119 tokens compiled from a collection of learning materials developed by the teachers for the classroom (see Table 2.12).

Table 2.11

Textbooks by Subject, Approximate Level in the International School, and Running Words

Subject	Grade	Texts	Running words	Totals
English	6	The Hunger Games (Chapters 1 and 2)	24,826	147,642
	10	Pride and Prejudice	122,816	
Maths	8	Gamma Mathematics	106,005	249,117
	11	Delta Mathematics	143,112	
Science	8	Pathfinder Year 9 (NZ)	8,829	64,795
	11	Pathfinder Year 12 (NZ)	55,966	
Total			461,554	461,554

Note. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 596. (https://doi.org/10.1002/tesq.450)

Table 2.12

Running Words and Texts of the Learning Materials by Subject

Subject	Number of texts	Total running words
English	6	845
Maths	2	320
Science	4	1954
Total	12	3119

Note. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 598. (https://doi.org/10.1002/tesq.450)

Coxhead and Boutorwick then carried out a frequency analysis using the Range program (Heatley et al., 2004). This was done using the word families from the 1,000 to 25,000 frequency bands of Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) (Davies, 2008) which they supplemented with lists of proper nouns, abbreviations, and marginal words. For the learning materials, over 95% coverage, the coverage necessary to understand a written text with support, was achieved for all three subject areas with knowledge of the 8,000 most frequent word families (see Table 2.13). For the English learning materials, knowledge of the first 8,000-word families would have provided over 98% coverage. For the corpus of textbooks (see Table 2.14) the first 8,000 word families provided over 98% coverage of the two novels, something that was expected based on the results of previous studies (Coxhead, 2012; Hirsh & Nation, 1992; P. Nation, 2006), and over, or close to, 95% coverage of the Grade 8 and Grade 11 Maths texts and the Grade 11 Science text. However, the first 8,000 word families only provided 92.95% of the Grade 8 Science text, well below the lexical threshold needed for understanding.

Table 2.13

Vocabulary Profile of Learning Materials in EAL, Maths, and Science (%)

Word Lists	EAL	Maths	Science
BNC/COCA 1,000-3,000	92.43	88.13	87.31
BNC/COCA 4,000-8,000	3.56	5.63	5.42
BC/COCA Supplementary Lists	2.97	1.56	4.04
Total	98.96	95.32	96.77

Note. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 601. (https://doi.org/10.1002/tesq.450)

Table 2.14

Coverage of the BNC/COCA High, Mid, and Supplementary Lists Over English, Maths, and Science Textbooks (%)

Frequency Bands	English Grade 6	English Grade 10	Maths Grade 8	Maths Grade 11	Science Grade 8	Science Grade 11
High BNC/COCA 1,000-3,000	92.62	92.28	85.32	73.19	85.75	83.79
Mid BNC/COCA 4,000-8,000	4.39	3.22	5.45	4.58	3.60	7.28
Supplementary Lists	1.08	3.42	7.20	16.85	3.60	4.97
Total	98.09	98.92	97.97	94.62	92.95	96.04

Note. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 601. (https://doi.org/10.1002/tesq.450)

Finally, Coxhead and Boutorwick looked at the coverage that the AWL (Coxhead, 2000) provided over the two corpora. They found that the AWL provided the greatest coverage over the Grade 11 Math corpus (7.6%), followed by the Grade 8 (6.41%) and Grade 11 (6.25%) science corpora, and the Grade 6 Math corpora at 5.61%. The AWL provided the least coverage over the novels at only 1.4%, which is consistent with what has been seen in other studies (Coxhead, 2000, 2012).

The results of this analysis show that the materials being used in the EAL classroom are likely to be difficult for EAL learners, even EAL learners that have mastered a lot of the high-frequency vocabulary used in general English texts. English, Maths and Science are likely to be the most difficult, with NNSEAL learner group potentially being able to understand fewer than 73.19 to 83.79 words in this context. While these may seem like high numbers of coverage, this would result in learners not being able to understand around two or three out of every ten words. Given that these words are likely to be the more technical, discipline-specific words (Green & Lambert, 2018), learners are likely to struggle to understand these texts.

Comment

While Coxhead and Boutorwick's paper is a comprehensive study that provides a detailed investigation of the vocabulary development of EAL learners, it is not without its flaws. Some of these are the same as the issues described in the studies I have reviewed thus far. Probably the biggest failing of the paper is that the corpus upon which the analysis of the threshold of coverage provided by the different word lists is based is extremely small, and only draws from a limited number of texts. While this is understandable, given the difficulties associated with compiling this type of corpus, it makes it hard to make any generalizations about how much coverage the different word lists would actually provide over all the textbooks that the participants would be likely to encounter during their secondary schooling. For example, both the Math and Science subcorpora comprised only two texts, each from a different grade level, and the entire Science subcorpora comprises less than 65,000 running words. The materials corpora are even smaller, 845 words for English and 320 for Maths.

Another issue with the study is that, while it looks at the vocabulary coverage different learners would be likely to achieve at different times in their schooling and

speculates about the effects that this may have based on previous studies, this is not correlated to the actual academic performance of the learners. For this reason, while it is clear that the NNSEAL and NNS groups would face significant challenges understanding the vocabulary found in their textbooks, it is difficult to tell if this actually translates to these learners struggling academically.

Finally, as the authors themselves acknowledge, the VLT they used to determine the vocabulary knowledge of the learners is an older assessment tool, and the word lists on which it is based may not provide an accurate picture of the actual vocabulary profiles of academic textbooks. Because the BNC/COCA are based on a corpus consisting of texts that are likely to be read by FLE speakers in the UK and the USA (P. Nation, 2020) respectively, the frequencies found in these lists may not match the frequencies found in academic textbooks at the secondary school level (Green & Lambert, 2018; Greene & Coxhead, 2015).

2.3 Review of selected studies on academic and EAL word lists

2.3.1 Xue and Nation (1984): A University Word List

Introduction

Xue and Nation's (1984) article describes one of the first word lists explicitly developed for learners studying English as an academic subject. Previous word lists either focused on the vocabulary that first language English speakers would need for the classroom (e.g., Thorndike & Lorge, 1944) or were developed for English Language learners, but were not academic (e.g., West, 1953). The word list described in this article was developed specifically for non-English-speaking students studying at a university where English was the primary medium of instruction.

Summary

Xue and Nation (1984) created the University Word List (UWL) as a tool for teachers to use in their classrooms. Their goal in creating this list was to identify which words were relevant to students at the university level and then develop a list of these words that teachers could use in their classrooms. In doing so, they hoped to provide teachers with a way to focus on the type of vocabulary that ESL learners would need to understand the texts that they were likely to encounter in the university classroom.

The UWL itself was developed using four existing word lists. First, an initial list of words was derived by combining two existing word lists, Campion and Elley's (1971) and Praninskas' (1972) word lists, the result was a word list comprising 678 word families. They then compared the combined word list to two other existing word lists that were compiled using a different set of principles, Lynn's (1973) and Ghadessy's (1979) word lists.

Campion and Elley's (1971) word list was published by the New Zealand Council for Educational Research. They originally developed their word list for the vocabulary subtest of the Language Achievement Test of Overseas Students (LATOS), a test that was instituted in 1970 in New Zealand at the request of the Grants Committee. As part of the LATOS, the original purpose of Campion and Elley's word list was to measure the English language proficiency of international students to determine if it was at the level required by universities in New Zealand (Movick, 1977). They compiled the list using a corpus of 301,800 words selected from 23 textbooks and 19 lectures published in journals, along with a selection of university exam papers. The corpus included texts from the 19 academic disciplines with the largest enrolment in New Zealand universities. The first 5,000 words from Thorndike and Lorge's (1944) list were excluded from the final list. Campion and Elley also used range as a criterion for selecting words for their list and removed all words that occurred in only one discipline. They then selected the remaining words based on the frequency of their occurrence in three of more of the 19 disciplines, along with a set of subjective criteria, including how familiar they thought those words would be for a native speaker.

Praninskas (1972) used a corpus of 10 first-year university textbooks that students at the American University of Beirut were required to read for their arts and science classes to create the American University Word List (AUWL). The resulting corpus covered ten disciplines and consisted of 272,466 tokens. Words that appeared in West's (1953) General Service List (GSL) were excluded. Like Campion and Elley (1972), it also used range as a criterion for including words in the AUWL. However, because the range of subject areas was so small, and the corpus had so few tokens, the resulting list was small and there was little variety in the words included (Coxhead, 2000).

After they had compiled the initial word list, the 678-word families in this list were checked against two different word lists that had been compiled differently. Lynn's (1973) and Ghadessy's (1979) word lists were both compiled by counting the annotations that international students made above words in textbooks. The rationale behind these lists was that international students would likely write a gloss (a short summary or the word that they were unfamiliar with), usually by translating the word into their own language, above vocabulary items that they found problematic. Lynn's (1973) corpus was derived from 52 textbooks taken from 50 students of accounting, business administration, and economics. The resulting list consisted of 197 word families, which Lynn organized by frequency of occurrence. Ghadessy's (1979) corpus was derived from 20 textbooks from the disciplines of chemistry, biology, and physics. The resulting list contained 795 items, which were alphabetized. Of the word families in Xue and Nation's list 70% were compiled by combining Campion and Elley's (1972) and Praninskas' (1972) word lists, overlapped with the words found in Lynn and Ghadessy's lists. A further 59 high-frequency non-overlapping words found in Lynn's (1973) and Ghadessy's (1979) lists, but not in the combined list, were added to the 678 word families found in the base list.

With the combination of these four different word lists, Xue and Nation (1984) arrived at a final UWL comprised of 737 base words. Xue and Nation then broke the UWL into sub-lists based on the range and frequencies of the words in the lists from which they were initially taken (see Table 2.15). To make it more accessible to teachers, the sub-lists were made available to teachers, a vocabulary test was developed, and a list of collocations for the words found in the UWL was prepared.

Table 2.15

Sources and Size of the Sub Lists of the UWL

Sub-lists	Number of words	From the Praninskas List		From the Campion & Elley list		From the Lynn list		From the Ghadessy list	
		range	number of words	range	number of words	frequency	number of words	frequency	number of words
1	75	10	35	8-12	37	8-12	2	13	1
2	88	9	39	6-7	44	19	3	12	2
3	76	8	44	5	26	8-12	4	11	2
4	70	7	49	4	16	16	5		
5	70	6	47	4	19	15	4		
6	79	5	55	3	20	14	4		
7	79	4	49	3	23	13	7		
8	74	3	44	3	25	12	5		
9	59	2	26	3	27	11	6		
10	64	1	5	3	41	10	18		

Note. Adapted from "A university word list," by G. Xue and P. Nation, 2018, *Language Learning and Communication*, *3*(2), p. 217.

By highlighting 737 word families that occur frequently in university textbooks, the UWL was able to provide teachers with a tool that they could use to help prepare learners for their university level classes. The UWL represents a bridge between general high-frequency words that have a wide range and are probably already familiar to students hoping to enter EMI classes at a university level and the low frequency domain

specific vocabulary that they are likely to learn when they study specific subjects at university. By focusing on a more general type of academic word, i.e., those academic words that occur regularly across different fields of study, the UWL provides teachers with a tool that they can use to help students in both EAP and more general first year university classes.

Comment

Xue and Nation's work on the university word list is important because it represents a first attempt to develop a word list designed specifically for L2 speakers of English studying in an EMI classroom. Since the UWL provided teachers with a tool that they could use in the classroom, it represents one of the first examples of how a word list can be beneficial as a pedagogical tool.

However, this list is not without its shortcomings. The first of these shortcomings is that, because of the technological limitations of the time when the list was developed, the researchers compiled a list using lists developed from existing corpora rather than develop their own corpora. As a result, the UWL inherits many of the weaknesses of the word lists that it was compiled from, primarily that they are small and the texts they are comprised of do not come from a balanced range of topics. Campion & Elley's (1971) and Praninskas' (1972) word lists, which represent the base of Xue and Nation's word lists, were derived from a combined corpus of 574,266 running words. While this number is respectable for the time at which they developed the list, compared with recently developed word lists, it is small. For example, Coxhead (2000) used a corpus of 3.5 million words to develop the Academic Word List, and Gardner and Davies (2014) used a 120 million-word corpus to develop the New Academic Vocabulary List. The problem with using this small a corpus is that it could make it difficult to accurately represent the full range of subjects and texts that we would expect students to engage with at the university level.

There are also some potential methodological issues with how these lists were combined. The first issue is that the authors of the original lists did not use the same criteria for removing high-frequency vocabulary from their lists. Campion & Elley used the first 5,000 words from Thorndike and Lorge's (1944) Teacher's Word Book of 30,000 Words, while Praninskas used West's (1953) General Service List. The

difference in procedure could have resulted in discrepancies between the high-frequency items removed from either list and the ordering of the remaining words. Furthermore, the range and type of texts were different between the two lists, which could cause a lack of balance in the combined list. One (Campion & Elley, 1971) made use of journal articles and exams, while the other (Praninskas, 1972) was compiled using first-year textbooks. The former list also included material from 19 different academic subjects, while the latter was drawn from only ten subjects. These could have caused differences in the frequency and range of the vocabulary that appeared in these two lists. Combining the two lists into a single one, without considering the discrepancies in how the original lists and corpora were derived, might have resulted in inconsistencies in how the words in the final list were ordered.

Despite concerns regarding the methodology of how the UWL was compiled, this paper is important because it represents one of the first attempts to develop a word list that focuses on general academic vocabulary that students studying in an EMI classroom would be expected to know. Subsequent studies using this particular list highlight the advantages of using this type of academic word list in the classroom (Hirsh & Nation, 1992) instead of more general word lists such as West's (1953) GSL or the subsequent word lists compiled from the BNC or COCA.

2.3.2 Coxhead (2000): A New Academic Word List

Introduction

Coxhead's (2000) article describes her seminal Academic Word List (AWL) creation and validation. The AWL was unique for its time both because of the size of the corpus from which it was compiled and the amount of coverage the 570 word families provided for academic texts. Coxhead compiled the AWL from a well-designed corpus of 3.5 million words that she compiled from a selection of academic English texts explicitly chosen to produce the AWL. The AWL soon replaced the UWL as the word list of choice for EAP students and teachers and is still the most commonly used academic word list in ESL and EAP classes today.

Summary

To develop the AWL, Coxhead compiled a corpus of 3.5 million running words from 28 subjects grouped into four disciplines (see Table 2.16). While other researchers, such as

Martin (1976) and Xue and Nation (1984), had previously produced lists of academic vocabulary, by using a corpus that represented the domain of texts that the word list was supposed to reflect Coxhead's AWL provided greater coverage of academic texts with fewer word families than existing lists. Coxhead could achieve this coverage because she compiled the AWL from a corpus that was representative, well organized, and of sufficient size.

Table 2.16

Texts and Subject Areas of the Academic Word List

	Discipline				
	Arts	Commerce	Law	Science	Total
Running words	883,214	879,547	874,723	875,846	3,513,330
Total Texts	122	107	72	113	414
Long Texts	18	18	23	19	78
Medium Texts	35	37	22	37	142
Subject areas	Education, History, Linguistics, Philosophy, Politics, Psychology, Sociology	Accounting, Economics, Finance, Industrial relations, Management, Marketing, Public policy	Constitutional, Criminal, Family and medicolegal, International, Pure commercial, Quasi-commercial, Rights and remedies	Biology, Chemistry, Computer science, Geography, Geology, Mathematics, Physics	

Note. Adapted from "A new academic word list," by A. Coxhead, 2000, *TESOL Quarterly*, *34*(2), p. 220. (https://doi.org/10.2307/3587951)

To avoid a biased and non-representative corpus, Coxhead included multiple texts from various authors. Research in corpus linguistics (P. Nation, 2016) shows that texts' linguistic features can differ significantly within the same genre. It is essential that the texts included in a corpus represent the variety of the texts in the domain that they are supposed to reflect. If a corpus relies too much on texts from a single author, that author's idiosyncratic style may result in that corpus having different high-frequency lexical and grammatical features than other texts from that same genre.

To allow her to compare the subcorpora directly when looking at the frequency and range of the vocabulary in the corpus, Coxhead divided her corpus into four discipline-specific subcorpora of 875,000 tokens each (see Table 2.16). Each of the disciplines was further divided into seven subject areas. Coxhead balanced the number of short and long texts in each of the four discipline-specific subcorpora with each section containing a roughly equal number of short texts (2,000 to 5,000 words long), medium texts (5,000 to 10,000 words long) and long texts (over 10,000 words long, see Table 2.16 for a breakdown of the length of texts in the corpus).

The final criterion that Coxhead had to consider when developing her corpus was the size. Following Francis et al.'s (1982) findings that a corpus of 3.5 million words would be necessary to identify 100 occurrences of each member of a word family, the corpus she compiled the AWL from was just over 3.5 million tokens long. The final academic corpus consisted of 414 academic texts by over 400 authors and contained 3,513,330 tokens and 70,377 types.

After compiling the corpus, Coxhead used the Range corpus analysis program (Heatley et al., 2004) to count and sort the words. She selected the words for inclusion into the AWL based on three criteria: the word must be academic, it must occur with sufficient frequency in the corpus, and it must appear with sufficient frequency across the various discipline-specific subcorpora. To ensure that the words selected were academic, Coxhead included only words outside the 2,000 most frequent words listed in West's (1953) GSL. For frequency, Coxhead only included word families that appeared at least 100 times in the Academic Corpus. Finally, members of the word family had to occur at least ten times in each of the four major sections of the corpus and in 15 or

more of the 28 subject areas. The resulting word list consisted of 570 word families, which was then divided into ten sublists based on frequency.

After compiling the AWL, Coxhead then had to determine what coverage the list provided. With the academic corpus that she compiled the AWL from, Coxhead found that the AWL provided roughly 10% coverage. This amount of coverage was twice as much as the second 1,000 words from West's (1953) GSL and almost 20% more than the UWL, which contains 266 more word families than the AWL (836 to 570). When combined with the GSL, the AWL and GSL together provided 86.1% coverage of the academic corpus. By itself the AWL provided between 9.1% and 12% coverage of the four subcorpora (see Table 2.17).

Table 2.17

Coverage of the AWL for the Academic Corpus and Its Four Subcorpora

		General Service List		
Corpus or subcorpus	Academic Word List	First 1,000 words	Second 1,000 words	Total
Complete Academic Corpus	10.0	71.4	4.7	86.1
Arts Subcorpus	9.3	73.0	4.4	86.7
Commerce Subcorpus	12.0	71.6	5.2	88.8
Law Subcorpus	9.4	75.0	4.1	88.5
Science Subcorpus	9.1	65.7	5.0	79.8

Note. Adapted from "A new academic word list," by A. Coxhead, 2000, *TESOL Quarterly*, *34*(2), p. 224. (https://doi.org/10.2307/3587951)

The final corpus with which Coxhead checked the coverage of the AWL against was a non-academic corpus made up of fictional texts compiled from a collection of 50

books taken from Project Gutenberg's (http://www.gutenberg.net) collection of public domain books. Coxhead checked the AWL coverage against a non-academic corpus to determine if the words included in the academic corpus were indeed academic. The AWL accounted for only 1.4% of the tokens in this non-academic corpus. The differences in the coverage of the two corpora suggest that the majority of the word families included in the AWL are indeed academic.

Coxhead then compared the AWL with a second academic corpus consisting of those texts she had prepared for the inclusion in her academic corpus, but which she did not include for a variety of reasons. The second academic corpus contained 678,000 tokens, and Coxhead found that the AWL provided 8.5% coverage of this corpus and 79.1% coverage when combined with the GSL. Coxhead speculated that this corpus's lower coverage was because it contained more science texts than the original academic corpus. The GSL and AWL together provided 79.8% coverage of the science subcorpora of the academic corpus.

Coxhead concludes her article by discussing how the AWL can be used and making suggestions for future research. In the classroom, she believes teachers can use the AWL to help identify which vocabulary is useful for their learners. In doing so, they will be able to set achievable vocabulary goals and help learners focus on the words that are the most valuable and relevant to their context. She also believes that the AWL can be used when writing materials and textbooks. To make the AWL easier for teachers to use, she divided it into ten sublists ordered by decreasing frequency (see Table 2.18). She also notes that, because many of the words that make up the AWL are of Greek or Latin origin, teachers can help students understand these words by having them study the prefixes, suffixes, and stems that make them up. However, Coxhead does warn that teachers should not just have their students memorize words from the list, but rather consider the teaching context and approach vocabulary instruction in a language-and-message-focused way.

Table 2.18

Sublists of the Academic Word List

Sublist	Items	Coverage of the Academic Corpus (%)	Cumulative coverage (%)	Pages per repetition in the Academic Corpus
1	60	3.6	3.6	4.3
2	60	1.8	5.4	8.4
3	60	1.2	6.6	12.3
4	60	0.9	7.5	15.9
5	60	0.8	8.3	19.4
6	60	0.6	8.9	24.0
7	60	0.5	9.4	30.8
8	60	0.3	9.7	49.4
9	60	0.2	9.9	67.3
10	30	0.1	10	82.5

Note. Adapted from "A new academic word list," by A. Coxhead, 2000, *TESOL Quarterly*, *34*(2), p. 228. (https://doi.org/10.2307/3587951)

Comment

The AWL represents a significant improvement over the UWL both in terms of the number of word families needed to achieve a reasonable coverage of academic text and in terms of the methodology of its construction, but it is not without its problems. Despite its importance, there are several methodological problems with how the AWL was compiled, including the choices of texts, that the AWL itself was built upon the GSL, and how words were counted.

While Coxhead attempted to create a well-balanced corpus from which to compile the AWL, there are several potential issues with the texts that she included (or ignored) in her corpus. The first of these is with the subjects that she decided not to

include in the 28 subjects that make up the four subcorpora. These include essential academic fields such as statistics, biochemistry, astronomy, electronics, and ecology. Particularly surprising was her not including any medical texts in the academic corpus (S. Fraser, 2010). Furthermore, an over-reliance on legal texts resulted in words such as *legal*, *economy*, *policy* and *legislate* being included in the first sublist of the AWL even though these words are of questionable importance for students outside of specific fields.

Another issue that Coxhead's AWL suffers from is that it was built upon West's (1953) GSL. While an important word list, the GSL was compiled in 1953 and is now considered by numerous scholars to be out of date and to no longer be an accurate representation of the high-frequency words of English (Gardner & Davies, 2014; P. Nation, 2016; Neufeld & Billuroğlu, 2005). Because it was created in the 1950s, the GSL contains outdated words such as *shilling* but does not include high-frequency modern words such as *computer* or *internet*. A further complication related to building the AWL on top of the GSL is that many academic words, such as *business, capital*, and *exchange*, were excluded from the AWL because they occur in the GSL (Nagy & Townsend, 2012).

One final criticism of the AWL is how words were counted. The first issue is that Coxhead chose the word family as the unit for counting words for the AWL. While there is evidence that word families are important units in the mental lexicon (Xue & Nation, 1984) recent research has shown that, for L2 speakers of English, lemmas may be a more accurate count of what vocabulary knowledge is necessary to be able to understand a text (Brown et al., 2020; Schmitt & Zimmerman, 2002). Another issue with how words were counted in the AWL was that the Range program that Coxhead used to count the words in the corpus could not take into account polysemous words. Because of this, Coxhead would not have been able to take into account how the same word can have a very different meaning when it is used in a different academic field (Hyland & Tse, 2007). For example, the word *function* has a different meaning in biology than it does in mathematics. These polysemous words would have been included in the AWL, even though the word's meaning may vary depending on the field in which it is used.

However, despite these criticisms, the AWL provided (and still provides) an invaluable tool for both researchers and teachers. It has been used extensively to create materials, develop tests, and has helped to inform vocabulary teaching in the classroom for over twenty years. It demonstrated the importance of academic vocabulary and set the stage for developing subsequent discipline-specific word lists.

2.3.3 Gardner and Davies (2014): A New Academic Vocabulary List

Introduction

Gardner and Davies' (2014) article describes the 3,000 lemma Academic Vocabulary List (AVL) that they compiled from the academic subcorpus of the Corpus of Contemporary American English (COCA, Davies, 2002). They found that the list that they developed from the 120-million word subcorpus provides approximately 14% coverage of the running words in the academic corpus of COCA and the academic corpus of the British National Corpus (BNC, Leech et al., 2014). The AVL differs from previous academic word lists in two critical ways. First, the AVL was developed using lemmas, not word families. Second, the AVL is not built on top of a general high-frequency word list; instead, Gardner and Davies identified academic words statistically using the frequency, range, and dispersion of lemmas across both a corpus of academic texts and a corpus of general English texts.

Summary

Gardner and Davies used the academic section of the COCA to develop their AVL. The entire COCA corpus contains 425 million words, while the academic subcorpus comprises over 120 million words. This subcorpus comprises texts taken from academic journals (85 million words), academically oriented magazines (31.5 million words), and the financial and economic sections of newspapers (7.5 million words) (see Table 2.19). The only section of the corpus to contain text taken from newspapers is the Business and Financial discipline. They added newspapers to this section because of the difficulty involved in accurately separating text from the formulas and tables of articles published in these fields. The COCA academic subcorpus comprises nine disciplines: Education; Humanities; History; Social Science; Philosophy, Religion, and Psychology; Law and Political Science; Science and Technology; Medicine and Health; Business and Finances.

Table 2.19

Texts and Subject Areas of the Academic Vocabulary List

Disciplines	Total Size	Journals/ Magazines	Representative Titles
Education	8,030,324	J: 8,030,324	Journals: Education, J Instructional Psychology, Roeper Review, Community College Review; Magazines: (none)
Humanities	11,111,225	J: 11,111,225	Journals: Music Educators Journal, African Arts, Style, Art Bulletin, Hispanic Review, Symposium; Magazines: (none)
History	14,289,007	J: 11,792,026 M: 2,496,981	Journals: Foreign Affairs, American Studies International, J American Ethnic History; Magazines: American Heritage, Military History, History Today
Social science	16,720,729	J: 15,782,359 M: 938,370	Journals: Anthropological Quarterly, Geographical Review, Adolescence, Ethnology; Magazines: National Geographic, Americas
Philosophy, religion, psychology	12,463,471	J: 6,659,684 M: 5,803,787	Journals: Theological Studies, Humanist, Current Psychology, Church History, Psychology; Magazines: Psychology Today, Christian Century, U.S. Catholic
Law and political science	12,154,568	J: 8,514,782 M: 3,639,786	Journals: ABA Journal, Perspectives on Political Science, Harvard J of Law & Public Policy, Michigan Law Review; Magazines: American Spectator, National Review, New Republic
Science and technology	22,777,656	J: 13,363,151 M: 9,414,505	Journals: Bioscience, Environment, Mechanical Engineering, Physics Today, PSA Journal; Magazines: Science News, Astronomy, Technology Review
Medicine and health	9,660,630	J: 5,714,044 M: 3,946,586	Journals: J Environmental Health, Orthopaedic Nursing, American J Public Health; Magazines: Prevention, Men's Health, Total Health
Business and finance	12,824,831	M: 5,256,801 N: 7,568,030	Journals: (none); Magazines: Forbes, Money, Fortune, Inc., Changing Times. Newspapers: 'finance' section.
Total	120,032,441	Academic journals: 84,914,694, Magazines: 31,496,816, Newspapers: 7,568,030 (Business and finance only)	

Note. Adapted from "A new academic vocabulary list," by D. Gardner and M. Davies, 2014, *Applied Linguistics*, *35*(3), p. 314. (https://doi.org/10.1093/applin/amt015)

Because the COCA had already been tagged for parts of speech using the CLAWS tagger program from Lancaster University (http://ucrel.lancs.ac.uk/claws/), Gardner and Davies could count the lemmas in the corpus accurately. From the academic corpora, they selected the core academic vocabulary words based on four criteria: ratio, range, dispersion, and a discipline measure. Gardner and Davies used the first of these four criteria to exclude general high-frequency words. The other three criteria were used to exclude technical and discipline-specific words.

To remove general high-frequency lemmas from the AVL, Gardner and Davies only considered words that were over 50% more frequent in the COCA's academic subcorpus than in the non-academic section of the corpus. The authors chose the ratio of 1.5 after examining which words were included or excluded from the corpus at different ratios. Lower ratios (between 1.3 and 1.5) resulted in too many general high-frequency words such as *work* (n), *large*, and *most* being included in the final word list. Conversely, higher ratios resulted in important academic words such as *system*, *political*, and *create* being excluded from the final word list.

Gardner and Davies first used the criteria of Range to remove discipline-specific words from the AVL. Lemmas had to occur in at least seven of the nine disciplines with the expected frequency to be included in the AVL. As with ratio, a range of 20% was determined by trying different ranges to determine which range resulted in the most appropriate core academic vocabulary list.

Gardner and Davies also used dispersion, a measure of how evenly a word is spread across the corpus, to identify and remove discipline-specific words from the AVL. A dispersion of .01 means that the word only occurs in a small part of the corpus, while a dispersion of 1.00 means that the word is evenly dispersed throughout the corpus. Dispersion eliminated words that are technical and appear much more frequently in one discipline. For example, they removed *taxonomy*, *microcosm*, and *filial*, which are relatively technical and discipline-specific, from the AVL because those words have a low dispersion in the COCA academic corpus. To be included in the AVL, lemmas had to have a dispersion of at least .80.

Gardner and Davies' final criterion for identifying discipline-specific words was discipline measure. The discipline measure is a measure of how frequent words are in a

specific discipline compared to their frequency in the other eight disciplines. Gardner and Davies excluded lemmas that appeared with over three times the expected frequency in one of the nine disciplines. Discipline measure was effective at removing words such as *student* (Education), *ministry* (Philosophy, Religion, and Psychology), and *software* (Science and Technology) from the AVL.

By applying these four criteria to their academic subcorpus, Gardner and Davies identified 3,000 lemmas that met the criteria they had set. To compare the AVL with Coxhead's (2000) AWL, they converted their lemma based AVL into word families. The resulting word family list contains the 570 most frequent word families from the AVL, allowing Gardner and Davies to more fairly compare it with the AWL, which is also made up of 570 word families. Gardner and Davies first compared this word family list to the three different genres of texts in the BNC and COCA (see Table 2.20). They found that the AVL provided the greatest coverage of the academic corpus and the least amount of coverage with the fictional corpus when compared with both the COCA and the BNC. The AVL provides good coverage of both the newspaper and academic subcorpora of both the COCA and BNC. The coverage the AVL provides for these sections in both corpora is similar. That the AVL can give this amount of coverage of an academic corpus that is not the one they created it from shows that the list is a good representation of the core academic vocabulary.

Table 2.20

The Coverage of the AVL of the Different Genres of the COCA and BNC

Genres	COCA			BNC		
	Genre size	# Words AVL	Coverage	Genre size	# Words AVL	Coverage
Academic	120,847,709	16,633,796	13.8%	32,828,961	4,507,211	13.7%
Newspaper	77,553,000	6,229,359	8.0%	10,638,034	740,065	7.0%
Fiction	83,369,907	2,862,093	3.4%	16,194,885	548,708	3.4%

Note. Adapted from "A new academic vocabulary list," by D. Gardner and M. Davies, 2014, *Applied Linguistics*, *35*(3), p. 322. (https://doi.org/10.1093/applin/amt015)

Gardner and Davies (2014) then compared the coverage that the AVL and the AWL give of the academic section of the BNC and COCA. They found that the AVL provided nearly twice the coverage for both the BNC and COCA (see Table 2.21). They acknowledge that some of this may be because Coxhead (2000) built the AWL upon the GSL meaning that some high-frequency academic words may have been excluded from the AWL list because they were in the GSL. However, for the same reason, the AWL includes many high-frequency words that are more general than academic.

Table 2.21

The Coverage of the AVL and AWL in COCA Academic and BNC Academic

List	COCA academic			BNC academic		
	Genre size	# Words AVL	Coverage	Genre size	# Words AVL	Coverage
AVL (570)	120,847,709	16,633,796	13.8%	32,828,961	4,507,211	13.7%
AWL (570)	120,847,709	8,601,839	7.2%	32,828,961	2,261,469	6.9%

Note. Adapted from "A new academic vocabulary list," by D. Gardner and M. Davies, 2014, *Applied Linguistics*, *35*(3), p. 323. (https://doi.org/10.1093/applin/amt015)

By leveraging advancements in technology, corpus construction, and corpus size, Gardner and Davies (2014) developed an academic word list that responds to some criticisms of the AWL. Their list uses lemmas instead of word families, is not built upon the outdated GSL, and uses a more robust set of statistical criteria to determine which words to include and exclude from the AVL. Since they develop their word family list from the list of lemmas, Gardner and Davies were able to provide additional information about the members of each word family, including which members were the most frequent and, when individual members were technical, the discipline where that member is most likely to be used. This additional information provides teachers

with more tools that they can use to help them decide what vocabulary to teach and is part of what makes the AVL an invaluable tool for EAP and university classes across academic disciplines.

Comment

The AVL is important because it represents some significant changes in how word lists are compiled. While it is possible to argue that some choices, such as the use of lemmas instead of word families, may not be the most appropriate for the type of higher proficiency learner that this type of list is aimed at (see Dang, 2021; P. Nation, 2021; Webb, 2021 for a discussion on this issue) the AVL represents a progression in how word lists are developed. That it is not built upon the old GSL and instead used statistical criteria to remove non-academic words from the list represents a significant improvement over previous lists. However, it is not without its problems, especially for researchers and practitioners that intend to use this list with EAL learners.

The biggest issue with Gardner and Davies' (2014) AVL is with the size and breadth of the list. In total, the AVL contains 3,014 lemmas; however, a large number of the lemmas in this list may not actually be that useful even for university level EAL learners. Durrant (2016) examined essays written by students at British universities across 32 different disciplines. What he found was that only 427 of the total 3,014 lemmas were frequently used in over 90% of the disciplines he examined. In other words, as with the AWL (Hyland & Tse, 2007) the frequencies of the AVL are highly skewed and they achieve the majority of the coverage provided by the list with only a few items. In fact, over half of the items on the list are so infrequent that they may be better considered specialized vocabulary rather than general academic vocabulary.

Another issue with Gardner and Davies' (2014) AVL comes from one of the strengths of the list, the fact that they compiled it using nine discipline-specific subcorpora. However, many of these subcorpora, such as *Law and political science* or *Medicine and health* are not studied in a high school or junior high school context. Because the statistical analysis used to compile the final word list rests upon the frequency of the words in these subcorpora, the words identified by Gardner and Davies may not be the most appropriate for EAL students studying in an international school context (Greene & Coxhead, 2015; P. Nation, 2016).

The final criticism of the AVL as a list for the international school context is with the texts from which it was compiled. While academic journals, academically oriented magazines, and the financial and economics sections of newspapers may represent the texts learners are likely to encounter at the university level, they are not appropriate for high school students, even for FLE speakers. For a word list to be useful, it is important that corpora from which it is developed represent the texts that the group of learners the word list is being developed for is likely to encounter (Coxhead, 2019; Nation & Sorell, 2016). Given that the type of vocabulary found in academic journals is different from the words found in secondary school textbooks (Coxhead, 2012; Coxhead et al., 2010) and that the corpus includes discipline-specific texts unlikely to be seen at this level it may not be the best list to use with this group of learners.

However, despite these issues, the AVL represents a positive change in how word lists are developed. Its use of lemmas, the reliance on statistics rather than outdated general word lists to remove non-academic vocabulary, and the size and scope of the subcorpora have been used as a template for the development of several modern general and discipline specific word lists (e.g., Green & Lambert, 2018; Lei & Liu, 2016).

2.3.4 Greene and Coxhead (2015): Academic vocabulary for middle school students

Introduction

This book is the first of two articles that focus on developing frequency-based academic word lists explicitly for secondary and middle-school students. In their book, Greene and Coxhead (2015) describe the methodology that they used to develop a set of academic word lists for middle-school students. The word lists described in the book are discipline-specific and span five content areas: English grammar and writing, health, mathematics, science, and social studies. The corpus from which Greene and Coxhead created their word lists contains just over 18 million tokens, which they compiled from 109 textbooks spanning grades 6 to 8 and covering five content areas. Greene and Coxhead used the GSL (West, 1953) and the AWL (Coxhead, 2000) as a basis for compiling the five Middle School Vocabulary Lists. These lists contain between 435 word families, for the science list, and 321 word families, for the mathematics list and

provide between 5.9% and 10% coverage of the vocabulary found in middle school textbooks, depending on the content area.

Summary

To identify the words that middle school students need to know to understand the texts they encounter in the classroom, Greene and Coxhead used a methodology similar to the one Coxhead (2000) used to develop her seminal AWL. However, rather than develop a single academic word list, the Middle School Vocabulary Lists consist of five different content-specific lists. They collected the textbooks for these lists from five content areas being taught in US-based public schools (Table 2.22 summarizes the number of textbooks they used and the content areas these textbooks are from).

Table 2.22

The Number of Textbooks by Content Area Used to Create the Middle School Content-Area Textbook Corpus

	Content area					Total
	English grammar and writing	Health	Math	Science	Social studies and history	
6th Grade	6	3	12	9	9	39
7th Grade	7	2	9	7	8	33
8th Grade	5	2	14	7	9	37
Total	18	7	35	23	26	109

Note. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 143. Paul H. Brookes Publishing.

Greene and Coxhead designed their corpus based on Hunston's (2002) criteria regarding the size, content, balance and representativeness, and permanence of a corpus. They purposefully excluded one subject area from the corpus, reading and literature,

based on their assertion that the textbooks used in this content area are predominately filled with fictional texts, not academic texts. The 109 textbooks they selected to include in their corpus were then scanned and saved as text files, which were numbered and filed according to grade level and content area. The resulting Middle School Content-Area Textbook (MS-CAT) Corpus is 18,202,382 words long and covers five subject areas. I show the number of words in each of the content areas in Table 2.23. Differences in the subcorpora sizes were considered in the word selection process.

Table 2.23

Running Word Composition of the Middle School Content-Area Textbook Corpus by Grade and Content Area

	Number of Running Words by Content Area and Grade Level					Total
	English grammar and writing	Health	Math	Science	Social studies and history	
6th Grade	811,647	333,468	1,680,737	1,330,662	1,754,229	5,910,743
7th Grade	1,164,185	232,062	1,283,934	1,181,148	1,744,301	5,605,630
8th Grade	957,525	347,313	2,067,089	1,213,612	2,100,470	6,686,009
Total	2,933,357	912,843	5,031,760	3,725,422	5,599,000	18,202,382

Note. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 144. Paul H. Brookes Publishing.

After they had saved the textbooks as individual text files, Greene and Coxhead then used Range 1.32 (Heatley et al., 2004) to analyze the corpus. Their preliminary analysis showed that 79.56% of the running words in the MS-CAT Corpus come from the GSL, compared to the 76.1% of Coxhead's (2000) Academic Corpus, while a further 5.37% of the words are covered by the AWL, compared to 10% for the Academic Corpus. Together the GSL and AWL cover 84.93% of the MS-CAT Corpus, compared to 86.1% of the Academic Corpus. Greene and Coxhead also found that all but two of

the word families in the AWL were present in the MS-CAT Corpus (intrinsic and paradigm were not present).

After their preliminary analysis, Greene and Coxhead used several steps to identify words from the corpus to include in the Middle School Vocabulary Word Lists. First, following the AWL methodology, they used the GSL, which was compiled in 1953, to exclude general English vocabulary by removing any words found on the GSL from the list of words under consideration for inclusion in the Middle School Vocabulary Lists.

The AWL was then used to select the first two groups of words. Greene and Coxhead first identified words from the AWL that were present across all five subjects. This was done by identifying those members of the AWL word families with a range of 11.4 per million words in each of the five content-area subcorpora and a minimum frequency of 28.5 times per million within the MS-CAT Corpus. Greene and Coxhead added any words that met these criteria to all five Middle School Vocabulary Lists. They then identified a second group of subject-specific words from the AWL that occurred in the MS-CAT Corpus. The subject-specific words were made up of those members of the AWL families that had a minimum frequency of 28.5 times per million within the whole MS-CAT Corpus and 11.4 times per million words in a single content-area subcorpora. Greene and Coxhead added these words to their related content area vocabulary list.

Greene and Coxhead next identified two groups of words that were common in the MS-CAT Corpus, but were not present in the AWL. The first of these two groups were words that occurred frequently across all five subject areas in the MS-CAT Corpus but are not in the AWL. Greene and Coxhead identified these words using the same range and frequency cuts that Coxhead used in the AWL. They added any words that were not in the AWL but had a range above 11.4 times per million words in each of the content-area subcorpora and 28.5 times per million words in the whole MS-CAT Corpus to each of the five vocabulary lists. The fourth and final group of words Greene and Coxhead identified were frequently occurring technical vocabulary from each of the content areas. They selected these technical words by identifying the words that occurred over 100 times per million in a specific subcorpus. They added these subject-

specific words to the content area vocabulary list in which the words frequently occurred.

The final Middle School Vocabulary Lists consist of five lists, one for each of the content areas: English grammar and writing, health, mathematics, science, and social studies. Greene and Coxhead then organized the qualifying words into word families. When necessary, the family's headword was added to the list (e.g., because only *depression* and *depressed* were present in the MS-CAT Corpus, they added the headword *depression* to the list). Table 2.23 summarizes the number of words in each of the content-area lists. One interesting departure from Coxhead's AWL is that, except for the missing headwords, all the words included in each of the word families are those that meet the range and frequency criteria necessary for inclusion in the Middle School Vocabulary Lists. They added no additional family members to the list. Unlike the AWL, which includes all members of the word families included in the AWL, regardless of whether they were present in Coxhead's (2000) Academic Corpus or not.

Table 2.24

Word Types and Word Families in the Different Content Area Middle School Vocabulary Lists

Content area list	Number of types	Number of word families
English grammar and writing	722	374
Heath	802	406
Mathematics	616	321
Science	859	435
Social studies and history	809	394

Note. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 147. Paul H. Brookes Publishing.

To test the word lists' coverage, Greene and Coxhead developed a smaller parallel corpus of 8,897,011 words that also covered each of the five content areas (see Table 2.25). They found that the content-area lists of the Middle School Vocabulary Lists provide between 5.95% (for social studies and history) to 9.48% (for science) coverage of the running words in the parallel middle school academic corpus. They also found the coverage of the Middle School Vocabulary Lists and the GSL to be above 86% for all but one of the subject areas, social studies and history (see Table 2.26).

Table 2.25

Running Words for the Subcorpora of the Parallel Corpus

	Subcorpora of the parallel corpus					Total
	English grammar & writing	Health	Math	Science	Social studies & history	
Number of running words	390,080	566,654	1,643,061	1,608,059	1,689,157	8,897,011

Note. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 150. Paul H. Brookes Publishing.

Table 2.26

Coverage of the Parallel Subcorpora by the Middle School Vocabulary Lists and the GSL

	Subcorpora of the parallel corpus				
	English grammar and writing	Health	Math	Science	Social studies and history
GSL (%)	82.41	84.00	79.45	79.36	78.53
MSVL (%)	6.08	8.17	9.41	9.48	5.95
Total	88.49	92.17	88.86	88.84	84.48

Note. GSL = West's (1953) General Service List, MSVL = Middle School Vocabulary Lists. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 150. Paul H. Brookes Publishing.

The final check that Greene and Coxhead (2015) performed on the Middle School Vocabulary Lists was to determine if these word lists were genuinely academic. Following Coxhead (2000), they checked the coverage of the Middle School Vocabulary Lists against a corpus of fictional literature. The corpus of fictional literature that they used was a 5,678,676-word corpus they had compiled from a collection of reading and literature textbooks. The Middle School Vocabulary Lists' coverage of this fictional corpus ranged from 1.73%, for the mathematical word list, to 2.89%, for the English grammar and writing word list. The low level of coverage provided by the Middle School Vocabulary Lists of this corpus indicate that the words in these lists are academic (coverage of the content-specific lists is given in Table 2.27).

Table 2.27

Coverage of the Fictional Corpus by the Middle School Vocabulary Lists and the GSL

	Middle School Vocabulary Lists				
	English grammar and writing	Health	Math	Science	Social studies and history
GSL (%)	83.75	83.75	83.75	83.75	83.75
MSVL (%)	2.89	2.11	1.73	2.09	2.48
Total	88.64	85.86	85.48	85.84	86.23

Note. GSL = West's (1953) General Service List, MSVL = Middle School Vocabulary Lists. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 150. Paul H. Brookes Publishing.

Greene and Coxhead's Middle School Vocabulary Lists are unique in that they are the first principled attempt to apply the methodology used to develop academic word lists for university students to create a set of word lists for younger EAL learners. Based on the coverage that Greene and Coxhead found these lists provided on their two sets of parallel corpora, we can say that the Middle School Vocabulary Lists provide a good representation of the vocabulary found in content area textbooks written for middle school students. This gives teachers at the middle school level an invaluable tool that they can use to support EAL learners in their classrooms.

Comment

Because it uses the same approach to compiling a word list as the AWL, the Middle School Vocabulary Lists suffer from many of the same methodological problems as the AWL. The two main issues are the use of the GSL to remove high-frequency vocabulary and the use of word families as the unit of counting. While I covered these issues in my previous discussion of the AWL, it would be helpful to summarize them again briefly. There are two main problems associated with using the GSL to remove high-frequency words. First, the GSL was developed in 1953 and contains many out-of-

date vocabulary items and is missing numerous words that would be considered high-frequency words today. Second, by using a general word list to remove high-frequency vocabulary it is possible that certain important words, such as *capital*, will be removed from the final word list, even though their usage in an academic text may be different from how they are used in general English. This is further compounded by using word families as the unit of counting. As a result, words such as *particle*, which is commonly used in science textbooks to describe a small piece of matter, were not included as they are members of a word family that is in the GSL, even if the more frequent member of the word family has a different meaning. For example, the headword for the word *particle* described above is *part*, some but not all of something, which has a different meaning from the word *particle* in a scientific context.

As with the AWL (Coxhead, 2000), Greene and Coxhead used word families as the unit of counting for the Middle School Vocabulary Lists. The issue with this is that the use of word families may be too broad. A single family can result in the inclusion of many semantically distant words under the same headword in a word family such as *please* and *unpleasantly*, *part* and *particle*, or *value* and *invaluable* (Brezina & Gablasova, 2015). Greene and Coxhead (2015) mitigate this problem by removing any infrequent members of the word family from the list. As a result, the family members are more likely to include the most common inflected or affixed forms and many of the more semantically different members of each family have been removed. However, the use of word families may have resulted in some words being over counted in the text and other words, such as the word *particle* as described above, being excluded.

One additional criticism of the Middle School Vocabulary Lists is its reliance on the AWL to select family members for inclusion into the lists. For this reason, the criticisms directed at the AWL regarding the need for discipline-specific word identification (Gardner & Davies, 2014; Martínez et al., 2009) can also be directed toward the sub-lists of the Middle School Vocabulary Lists.

However, despite these criticisms, the Middle School Vocabulary Lists are important because they represent the first attempt to develop a comprehensive academic vocabulary word list for the EAL context. It is a useful tool for both researchers and teachers, as it can identify the types of words that learners would need in that context.

As it becomes more widely used, it will be useful for the creation of materials, the development of tests, and as a reference tool for the pedagogical choices being made in the middle school classroom.

2.3.5 Lei and Liu (2016): A new medical academic word list: A corpus-based study with enhanced methodology

Introduction

While this study focuses on technical rather than general academic vocabulary, the technique that the authors used to compile their word list makes it important study for future academic word lists. This paper describes the creation of a medical academic word list, the Medical Academic Vocabulary List (MAVL). Lei and Liu (2016) developed the list to better support medical students and medical professionals whose first language is not English. Prior to the MAVL, discipline-specific word lists (e.g., Liu & Han, 2015; Wang et al., 2008) were developed primarily using the procedures used by Coxhead (2000) when she developed her Academic Word List (AWL). Lei and Lui's (2016) study deviates from these previous studies in two important ways. First, Coxhead (2000) used West's (1953) General Service List (GSL) to exclude high-frequency words from the AWL. Recent improvements in corpus linguistics and word list development (e.g., Gardner & Davies, 2014) have shown that there are several potential problems with this approach. One problem specific to the development of technical and academic word lists is the difficulty related to differentiating high-frequency vocabulary from technical vocabulary (Chung & Nation, 2004). Second, Lei and Liu use natural language processing (NLP) to lemmatize and tag the corpus for part-of-speech (POS). Improvements in NLP software had made lemmatization easier, which opens up the possibility of lemmatizing large corpora and using these corpora to build word lists around lemmas rather than word families. The word list that Lei and Liu were able to compile using these techniques was able to provide better coverage of medical texts with fewer words than previous technical lists (e.g., Wang et al., 2008).

Summary

Lei and Lui's (2016) paper describes the creation of a medical word list using a combination of the techniques employed by Gardner and Davies (2014) in creating the new Academic Vocabulary List (AVL) and the procedures Coxhead (2000) used to

develop the AWL. Lei and Lui hoped that by combining the insights from these two studies they could develop a modern medical vocabulary list that better serves the needs of medical students and non-native English-speaking professionals than existing word lists (e.g., Wang et al., 2008).

To develop their list, Lei and Liu first constructed two corpora: one 2.79 million word corpus compiled from articles from medical journals they called the medical academic English corpus (MAEC), and an additional 3.5 million-word corpus compiled from English medical textbooks they named the medical textbook English corpus (MTEC). They compiled the word list itself from the MAEC, while the MTEC was used to ensure that the words they identified in the MAEC were also present in the types of texts medical students would be likely to use in their classes. The MAEC consists of 760 articles taken from 38 medical journals randomly selected from the 176 SCI-indexed medical journals found in the Elsevier database (see Table 2.27 for a list of the specialist areas and journals included in the corpus). They divided these journals into 21 subcorpora with each subcorpora selected to cover a specialist area such as cardiology, dermatology or surgery. The tables, figures, notes, endnotes and footnotes, references, and appendices were not included in the corpus. According to Lei and Lui, their corpus differs from the corpora used in previous studies in three important ways: i) it is much larger than those used by previous studies; ii) it includes both research and review articles, as opposed to earlier medical word lists which have focused more on research articles; and, iii) unlike previous medical corpora, Lei and Liu did not exclude articles that were not written by FLE speakers. The MTEC is composed of a 3-volume textbook of medicine (Warrell et al., 2010) that was chosen as it serves as the assigned textbook of many foundation courses in medicine. This second corpus was used to ensure that all the words extracted from the MAEC also occur with sufficient frequency in medical textbooks.

Table 2.27

List of specialist areas and journals included in the MAEC

Discipline	Journal Name
1. Bone	Bone
	Osteoarthritis and Cartilage
2 Cancer/General oncology	Cancer Letters
	European Journal of Cancer
3 Cardiology	American Journal of Cardiology
	American Heart Journal
	International Journal of Cardiology
4 Dermatology	Journal of the American Academy of Dermatology
5 Drug & Alcohol dependence	Drug and Alcohol Dependence
6 Epidemiology	Journal of Clinical Epidemiology
7 General Surgery	Journal of Surgical Research
	Journal of the American College of Surgeons
	Surgery
8 Haematology	Critical Reviews in Oncology Haematology
	Blood Reviews
	Transfusion Medicine Reviews
9 Hepatology	Digestive and Liver Disease
	Journal of Hepatology
10 Infection	Journal of Hospital infection
11 Metabolism & Gastroenterology	Best Practice & Research Clinical Endocrinology & Metabolism
	Nutrition Metabolism and Cardiovascular Diseases
	Best Practice & Research in Clinical Gastroenterology
12 Medical informatics & Biomechanics	International Journal of Medical Informatics
	Journal of Biomechanics
13 Ophthalmology	American Journal of Ophthalmology
14 Pathology	Human Pathology
15 Preventive Medicine	Preventive Medicine
	Journal of Adolescent Health
16 Psychiatry	Journal of Affective Disorders
	Journal of Psychiatric Research
	Psychiatry Research-neuroimaging

Discipline	Journal Name
17 Schizophrenia Research	Schizophrenia Research
18 Specialized Oncology	Oral Oncology
	Gynaecologic Oncology
	Radiotherapy and Oncology
19 Specialized Surgery	Annals of Thoracic Surgery
20 Transplantation	Transplantation Proceedings
21 Virology	Journal of Clinical Virology

Note. Adapted from "A new medical academic word list: A corpus-based study with enhanced methodology," by L. Lei and D. Liu, 2016, *Journal of English for Academic Purposes*, *22*, p. 50. (https://doi.org/10.1016/j.jeap.2016.01.008)

After compiling the two corpora, Lei and Liu used four steps to identify words to include in the final MAVL. In the first step, Lei and Liu used the Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/) to lemmatize and tag the text for part-of-speech (POS). After the corpus had been lemmatized and tagged for POS, a Python (https://www.python.org/) script was used to extract the POS-tagged lemmas from the MAEC corpus. In the second step, a series of other Python scripts were used to identify lemmas to include in the list using the five criteria listed below.

The five criteria Lei and Liu used in the second step were: minimum frequency, frequency ratio, range ratio, dispersion, and a discipline measure. The first criterion they used was frequency. Lei and Liu followed Coxhead's (2000) frequency ratio and identified all the words in the MAEC that had a frequency of 28.57 times per million words (PMWs) over the entire corpus. The second criterion they applied was a frequency ratio. The purpose of this criterion was to remove general high-frequency lemmas from the medical word lists. Using the same ratio as Gardner and Davies (2014), Lei and Liu removed all the lemmas that were not at least 50% higher in the medical corpus than in a corpus of general English. The authors used the non-academic subcorpora of the British National Corpus (BNC) for this purpose and calculated the normalized frequency of the lemmas in the MAEC against the normalized frequency of the same lemmas in the BNC to determine which words should be included. The third criterion that Lei and Liu used was a range ratio. Because their goal was to identify

general medical vocabulary, the range ratio was used to ensure that the lemmas selected occurred across a wide range of subcorpora. Range ratio compares how frequently a word occurs within a single subcorpora compared to its frequency across the whole corpus. Lei and Lui stipulated that a lemma should have a range ratio of at least 50%, so it would have to occur with at least 20% of the expected frequency in at least 12 of the 21 subcorpora.

The fourth criterion that Lei and Liu used to identify words for inclusion in their list was a dispersion measure. Dispersion is a measure that combines frequency and range and measures how balanced word frequencies are across different subcorpora (P. Nation, 2016). It does this by measuring how frequently a word occurs in each of the different parts of the corpus. This is important, because merely measuring frequency may lead to including certain words simply for the reason they are used extensively in only one of the documents in the corpus, because of the subject of that document or because of the author's writing style. A dispersion measure is also more effective than range, which only looks at if a document contains a certain word, not at how many times that word appears in the document. This ensures that words are more or less evenly spread across the corpora. Similar to Gardner and Davies (2014), Lei and Liu used Juilland's D (Juilland et al., 1970). Juilland's D is a useful measure of dispersion because it measures the amount of times a word appears in a subcorpus by calculating the appearances of that word as a percentage of the total words in the corpus, as opposed to just counting how many times the word occurs (Gries, 2020). This is useful as it is better able to accommodate subcorpora of different lengths. While the authors acknowledge that there are several studies that question whether Juilland's D is the best measure of dispersion to use for a corpus (e.g., Burch et al., 2017) they conclude that this measure is considered to be the most reliable measure of diversity and is widely used in corpus linguistics (Rayson, 2003). While Gardner and Davies (2014) used .80 as the minimum Julliand's D in their study, Lei and Liu set their minimum Juilland's D score at .50.

The fifth and final criterion Lei and Liu used to create their initial word list was a discipline measure that focused on identifying and removing unnecessary discipline-specific and technical words from the list. Following Gardner and Davies (2014), Lei and Liu excluded words that were over three times more frequent in any three of the

discipline specific corpora than they were in the other 18 subcorpora. Their rationale was that these words were likely to be technical vocabulary as opposed to general medical vocabulary.

By applying these five criteria to the list of lemmas extracted from the MAEC, Lei and Liu were able to identify a preliminary list of 1,234 lemmas that had the potential to be included in the MAVL. They then checked this list against the MTEC.

The third step in creating the MAVL was to check the preliminary words extracted from the MAEC against the MTEC. In this step, Lei and Liu checked how frequently the 1,234 lemmas in the preliminary list occurred in the MTEC. This resulted in 269 lemmas (21.8%) of the lemmas that did not meet the 28.57 PMWs frequency requirement being removed from the list. In the fourth and final step, they checked the remaining 965 lemmas against Brezina and Gablasova's (2015) New GSL. They checked their list against the New GSL to ensure that only general high-frequency words with a special medical meaning were included in the MAVL. The 459 words from the preliminary MAVL word list that also appeared in the New GSL were checked against two well-known medical dictionaries (*Merriam-Webster's medical English dictionary, New Edition*, 2006; *Taber's Cyclopedic medical dictionary*, 2013). They removed high-frequency words that did not appear in either of those two dictionaries from the word list. Of the 459 words that were on the New GSL, 146 were not listed in any medical dictionary and were removed. After these general high-frequency words had been removed, the resulting list was composed of 819 lemmas.

After identifying the 819 lemmas to include in the MAVL Lei and Liu checked their list against general, academic, and medical corpora. The MAVL provided more than twice the coverage of the two medical corpora that it was checked against, 19.44% for the MAEC and 20.18% for the MTEC (see Table 2.28), compared to the BNC academic corpus (6.64%) and the non-academic BNC subcorpora (3.69%). That the MAVL provides over twice the coverage for medical corpora as it does for academic and general corpora shows it is a viable and representative list. To compare the coverage of the MAVL to a previously compiled medical word list, the MAWL (Wang et al., 2008), Lei and Liu first used 20k Familizer Pro (https://www.lextutor.ca/familizer/) to convert the 623 headwords from the MAWL

word list into 618 word families, which were then converted into 3,552 word forms. From these word forms, they identified 1,751 lemmas. When the coverage of the MAWL's 1,751 lemmas was compared to the coverage provided by the 819 lemmas on the MAVL, the MAVL was found to provide significantly higher coverage of the MAEC (19.44% compared to 10.52%) as well as the MTEC (20.18% compared to 12.97%) despite containing 53.23% fewer lemmas (see Table 2.29).

Table 2.28

Coverage of MAVL Across General, Academic, and Medical Corpora

	BNC	BNC Academic	MAEC	MTEC
MAVL	3.69%	6.64%	19.44%	20.18%

Note. Adapted from "A new medical academic word list: A corpus-based study with enhanced methodology," by L. Lei and D. Liu, 2016, *Journal of English for Academic Purposes*, *22*, p. 50. (https://doi.org/10.1016/j.jeap.2016.01.008)

Table 2.29

Coverage of MAVL and MAWL in the MAEC and the MTEC

	MAEC	MTEC
MAVL	19.44%	20.18%
MAWL	10.52%	12.97%

Note. Adapted from "A new medical academic word list: A corpus-based study with enhanced methodology," by L. Lei and D. Liu, 2016, *Journal of English for Academic Purposes*, *22*, p. 48. (https://doi.org/10.1016/j.jeap.2016.01.008)

Lei and Liu's study is important in that it shows the benefits of using current techniques such as NLP and lemmatization, along with more sophisticated frequency measures, when developing word lists. The study also shows the importance of

including general high-frequency words that have higher frequencies in the discipline, or discipline-specific meanings, in technical and academic word lists. These techniques are important in the context of this thesis as they have implications for the creation of academic word lists, and have been used in creating recent academic word lists.

Comment

Lei and Liu (2016) study is a worthy addition to the field and represents an improvement in the overall methodology being used to extract words from text. The motivation for the project is clear, and the methodology for choosing words from the corpora are straightforward and clearly described. The corpus contains a wide variety of texts taken from a diverse selection from different medical fields. There are, though, some questions which might be asked about the composition of Lei and Liu's corpus and the presentation and interpretation of their findings.

While a large corpus containing a diversity of texts is desirable, the size and scope Lei and Liu's corpus leads us to question its specificity. Their use of randomly selected journals helped them to obtain their objective of creating a corpus that covered a wide range of specialty areas. However, the use of journals from twenty-one discipline areas may have been too wide-ranging to represent the types of texts that a general medical student, especially a pre-med student for whom this list was also intended, would be likely to encounter in their studies. Ophthalmology, for instance, is a highly specialized subject and most medical students are unlikely to encounter it. It may have been more appropriate for the authors to select journals from specialty areas that medical students are likely to encounter in their general studies, rather than choose the journals randomly.

There are also some potential problems with the way they presented some of the data in the paper. While the authors provide a complete list of all the lemmas in the MAVL, along with the part of speech of each lemma at the end of their paper, this list is presented in alphabetical order rather than by frequency. With a word list of 819 lemmas, it is unlikely that these words all have similar frequencies in the corpus. For example, a study by Coxhead and Hirsh (2007) found that 60 of the total 318 word families provided over half the coverage of their science-specific word list. If the same type of ratio exists in Lei and Liu's list, it would have been beneficial if the final

lemmas were broken down into sub-lists based on frequency. Furthermore, it is doubtful that the lemmas would have had similar frequency across the different specializations, and an argument could be made for the creation of domain specific word lists that focused on the individual specializations. However, this may not have been as useful, or even possible, given that the domains in the MAEC were chosen at random.

To sum up, the value of Lei and Lui's (2016) study is that it demonstrates how frequency-based techniques can extract words from a specialized corpus. It also illustrates the need to examine general words in the context of the domain the word list is being created for, to ensure that high-frequency words with specialized domain specific meanings are included in the final word list. We do however, need to examine Lei and Lui's methodology carefully for ourselves to decide if the use of highly specialized texts chosen at random in their initial MAEC corpus may have led to the erroneous inclusion, or exclusion, of certain words.

2.3.6 Green and Lambert (2018): Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects

Introduction

This paper describes the processes involved in the production of Green and Lambert's (2018) Secondary School Vocabulary Lists (SVL), the development of which is, to my knowledge, the first word list produced specifically for learners enrolled in English as a Medium of Instruction (EMI) classes in the secondary school context. They compiled their list from a collection of eight discipline-specific lemma-based academic vocabulary lists, putting together a corpus of 16,253,350 words from the eight academic disciplines in which secondary school learners are likely to be enrolled. From this corpus, they developed a set of eight discipline-specific lists that contain 4,781 lemmas in total. The study is significant because it uses the techniques for word list development pioneered by Lei and Liu (2016) to create a secondary school academic word list. Furthermore, the list is essential because, unlike previous academic word lists designed for adult learners (e.g., Coxhead, 2000; Gardner & Davies, 2014; Xue & Nation, 1984), they specifically produced this list for secondary school students.

Summary

This article reports on a large corpus project focused on developing eight discipline-specific secondary school academic word lists: the Secondary School Vocabulary Lists (SVL). Green and Lambert's (2018) developed their lists from a corpus of over 16 million words compiled from 206 high school textbooks. They selected the textbooks used for their corpus from the Singaporean Ministry of Education official textbook list and the UK A-Level/O-Level syllabi. These textbooks were supplemented by textbooks that were marked as preparatory texts for A- or O-levels. Eighty-two percent of the textbooks used for the corpus were published within five years of 2018. Following Lei and Liu (2016), Green and Lambert created eight separate discipline-specific vocabulary lists. The SVL covers eight academic disciplines: Biology, Chemistry, Physics, Geography, English, Mathematics, Economics, and History (see Table 2.30 for a breakdown of the disciplines and the number of words in each discipline-specific subcorpus). The lists that Green and Lambert compiled were made up of both the technical and academic vocabulary found to be statistically prominent in each discipline.

Table 2.30

Word Count by Discipline Specific Subcorpora

Biology	Chemistry	Physics	Geography	English
2,011,083	1,908,228	1,911,574	2,221,239	2,110,857
Mathematics	Economics	History	Total:	
1,404,280	2,297,055	2,389,034	16,253,350	

Note. Adapted from "Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects," by C. Green and J. Lambert, 2018, *Journal of English for Academic Purposes*, *35*, p. 109. (https://doi.org/10.1016/j.jeap.2018.07.004)

To compile the corpora used to develop these lists, Green and Lambert (2018) first scanned the textbooks they had selected for inclusion into their corpus and then used the Ominpage 18 software package (https://www.kofax.com/products/omnipage) to OCR the text. OCR, or optical character recognition, is the process of using a computer to automatically convert images of printed or written text into a form that the computer can recognize as words. They then exported the text from these files to text documents and the indexes, references, front matter, and content pages were removed. They then tagged the resulting corpora for parts-of-speech (POS) using CLAWS (https://ucrel.lancs.ac.uk/claws/). CLAWS is a software package developed by the University of Lancaster that adds tags to the words in a text that denotes its part-of-speech. The corpus was then lemmatized using Wordsmith (https://lexically.net/wordsmith/version4/) along with a combination of several lemma conversion lists. Following Lei and Liu (2016), the resulting tagged and lemmatized corpus was then processed to determine what words they would include in the final SVL. They did the processing of the corpora in six stages, as outlined below.

In the first stage, Green and Lambert first identified the high-frequency words within each of the disciplines. Using numbers derived from Lei and Liu (2016) and Coxhead (2000), they identified all words with at least 28.57 occurrences per million words in the discipline-specific corpora. In the second stage, the resulting lists of high-frequency word lists were then narrowed down using the range and dispersion of the words in the lists. Both measures are commonly used in creating word lists to ensure that the words selected are more or less evenly spread across the corpus and do not occur in only one or two texts. To do this, Green and Lambert first looked at the range of the lemmas in each of their lists. Using the procedure utilized by Gardner and Davies (2014), they removed any words that did not occur in at least 50% of the texts within the discipline. In the third stage, they then further narrowed down the words in each of their lists by using the dispersion of these words. To do this, Green and Lambert used the Oakes Dispersion test (Lei & Liu, 2016; Oakes & Farrow, 2007), which divides the corpora into eight equal subcorpora and then evaluates items for the homogeneity of the occurrences of the lemmas across the eight subcorpora. Following Lei and Liu (2016), they set the D value for this test at 0.5 and excluded any lemmas with a dispersion ratio below this.

In the fourth stage, Green and Lambert then used the range ratio (as discussed in the review of Lei and Liu (2016), this is a measure of the frequency of a word in the subcorpora in relation to the frequency of the same word in the whole corpus) to identify and remove lemmas that had a high range across the discipline-specific corpora but were not present, or extremely infrequent, in specific texts. They again followed the procedure used by Gardner and Davies (2014) and Lei and Liu (2016) and removed any lemma that had below 20% of its minimum frequency in less than 50% of the texts. In the fifth stage, Green and Lambert then used the frequency ratio of the lemmas to remove overly discipline-specific lemmas from their list. Accordingly, they removed any words over three times higher in the academic text than in a general corpus, the same ratio Gardner and Davies (2014) used in their AVL. Finally, in the sixth stage, they checked the lemmas for part-of-speech and removed any lemmas that were not nouns, verbs, adjectives, or adverbs, even if these lemmas met the other statistical benchmarks for inclusion into the list.

After the eight discipline-specific word lists had been compiled, the lemmas included in the lists were then checked for problematic tags, scanning noise, and to ensure that there were no proper nouns or adjectives. However, the lemmas included in the final list were not vetted based on the subjective judgment of the researchers. The word lists were then used to create a list of each word's discipline-specific collocations. They checked these collocations to ensure that they had the required range and mutual information score (Xiao & McEnery, 2006) for inclusion into the collocation lists (see Table 2.31 for the criteria collocations had to meet to be included in these lists). Green and Lambert (2018) also used Familiarize Pro (https://www.lextutor.ca/familizer/) to develop a word family list following the steps given by Lei and Liu (2016). Once these word families had been extracted, they checked each member of the word family against the corpora. Those members that did not meet the statistical criteria listed above were removed. The final lists were then arranged by frequency (Table 2.32 gives an example of the five most frequent word families from the chemistry word list).

Table 2.31

Criteria for Inclusion in the Word Association Lists

Statistical threshold: > Mutual Information score of 3.00.
MI is a statistic that computes if two words co-occur significantly more with each other than other words. The threshold of 3 is commonly used as indicating a meaningful relationship (Xiao & McEnery, 2006).
Minimum frequency: > 5 co-occurrences within a five-word span.
MI excludes words that frequently combine with many words (e.g. the), but can overemphasize low-frequency collocations (Liu, 2010, 2013). Therefore, collocates had to occur minimally five times within five words to the left or right of the SVL word.
Range: > approximately 20% of texts.
This metric ensured the collocation was not restricted to only a few texts.

Note. Adapted from "Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects," by C. Green and J. Lambert, 2018, *Journal of English for Academic Purposes*, *35*, p. 112. (https://doi.org/10.1016/j.jeap.2018.07.004)

Table 2.32

Five Most Frequent Word Families in the Chemistry Word List

Headword	Family Freq.	Family Members and Frequency
REACT	27,991	react (2331) reactant (882) reactants (1256) reacted (467) reacting (546) reaction (14114) reactions (3789) reactive (1260) reactivity (878) reactor (37) reactors (18) reacts (2195) unreactive (218)
ACID	15,833	acid (11779) acidic (1189) acidity (205) acids (2660)
ATOM	12,316	atom (4665) atomic (1635) atomise (1) atomised (2) atomises (1) atomise (1) atoms (5932) subatomic (80)
ION	11,944	ion (3611) ionisation (690) ionise (53) ionised (96) ionises (46) ionising (3) ionisation (101) ionise (11) ionised (12) ionises (6) ionising (1) ions (7314)
FORM	11,094	form (5582) formed (3791) forming (607) forms (1114)

Note. Adapted from "Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects," by C. Green and J. Lambert, 2018, *Journal of English for Academic Purposes*, *35*, p. 114. (https://doi.org/10.1016/j.jeap.2018.07.004)

The final SVL consists of eight discipline-specific word lists totalling 4781 lemmas and ranges in size from 253 lemmas to 880 lemmas. From largest to smallest, they report the finalized lists: 880 lemmas for Biology, 519 for Chemistry, 477 for Economics, 686 for English, 702 for Geography, 717 for History, 546 for Physics, and 253 for Mathematics. The words in these lists range from complex technical vocabulary, such as *photosynthesis* and *enzyme* in biology to more general vocabulary. One of the unique characteristics of this list compared to Coxhead's (2000) AWL is the inclusion of high-frequency vocabulary with an academic meaning, like the word *set* in the mathematics list. Green and Lambert considered the inclusion of these words in the list as crucial because their meaning in a specific domain differs from general use. These words would not have been included in the list if words from the GSL were removed.

They then tested the lists against both their domain-specific corpus and the entire corpus to determine the coverage of the lists. They found chemistry and biology to have the greatest coverage, while they found history and English to have the lowest (see Table 2.33 for a summary of the SVL for the different disciplines). One reason for this difference in coverage is that the hard sciences tend to have a more specialized vocabulary, while the humanities often use a richer and more diverse vocabulary (Coxhead et al., 2010), something that is supported by the Type Token Ratio (TTR) of the various corpora. Green and Lambert found that the TTR of their biology and chemistry corpora were 35.14 and 32.14, while the TTR of the English and history corpora were 42.58 and 42.78, respectively. The higher TTR of the English and History corpora indicates a higher degree of lexical diversity for these types of texts. The vocabulary used in these subjects is more varied than it is in the sciences.

Table 2.33

SVL Coverage Per Discipline

% Words Covered	Biology	Chemistry	Physics	Geography
Within discipline	23.00%	25.00%	22.70%	15.90%
Corpus overall	2.00%	2.20%	2.90%	2.40%
	English	Mathematics	Economics	History
Within discipline	13.00%	20.90%	21.80%	14.00%
Corpus overall	2.00%	2.90%	2.20%	1.70%

Note. Adapted from "Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects," by C. Green and J. Lambert, 2018, *Journal of English for Academic Purposes*, *35*, p. 114. (https://doi.org/10.1016/j.jeap.2018.07.004)

Green and Lambert's SVL is an essential contribution to the field of word list development for three important reasons. First, they focus specifically on the academic vocabulary required by secondary school students, unlike most other academic vocabulary lists developed for adult learners. Second, they based the SVL on the frequency count of lemmas rather than word families, ensuring that all the forms of the words included in the lists are ones that secondary school learners are likely to encounter in their texts. Third, and finally, they are domain-specific, both regarding the fact that there are separate lists for each domain and that they include general high-frequency words with different domain-specific meanings. Their focus on secondary school learners, use of lemmas, and focus on domain-specific meanings enable Green and Lambert's (2018) SVL to provide a more complete picture of the types of vocabulary secondary school students are likely to need in the classroom, which is essential for both pedagogical and research purposes.

Comment

Green and Lambert's (2018) SVL represents one of the most extensive word lists for secondary school students, to date. By building on the statistical techniques used by

Gardner and Davies (2014) and Lei and Liu (2016) the authors were able to develop a word list that does not rely on previous word lists. Using this technique allowed them to remove general vocabulary items without having to rely on outdated lists. The authors also made use of lemmas, allowing them to more accurately represent the difficulty some learners would have with understanding the more complicated members of certain word families. However, there are several issues that need to be considered when examining Gardner and Davies' list in the context of EAL learners studying in an international school context. The initial problem is the fact that the list focuses on the textbooks used in schools in Singapore and the UK and may not represent the texts being used in international schools. This is also true of the authors' choice of subcorpora. While these may represent the classes in which students are regularly enrolled in Singapore, they are not representative of the classes taken by students enrolled in the International Baccalaureate program. As such, it is necessary to at least validate this word list with the types of text learners in the international school context would be likely to encounter.

2.4 Discussion

The studies reviewed above follow two related themes which are central to lexical research in the EAL context: identifying the words EAL learners need in the classroom, and understanding how knowledge, or a lack of knowledge, of these words will impact their academic performance. While both areas of research have progressed in recent years, it is still necessary to draw these two threads of research together. In this second section, I review the studies together to acquire a clearer understanding of how they complement each other, providing different parts of the puzzle of what vocabulary to teach in the EAL classroom, and to highlight what still needs to be done. From the literature reviewed above, the following questions have emerged as particularly needing attention:

1. Why is a specialized EAL word list necessary, and what benefits will it have over and above existing word lists?
2. What texts should be used to develop this word list, and how many subcorpora need to be included and how big should each of these corpora be?

3. What techniques should be used to identify the most important words from the texts in this corpus, and what lexical unit should be used?

2.4.1 Why is a specialized EAL word list necessary?

Recently there has been an increasing pressure to provide support for learners who are studying in an EMI context and who speak a language other than English as their first language (Murphy, 2014). In the UK, for example, schools have experienced increases as large as 16.2% annually in the number of students whose first language is not English (Strand et al., 2015). In order to provide the educational support that these English as an Additional Language (EAL) learners require, we first need to better understand their needs (Hawkins, 2005). One area where this support is particularly important is vocabulary knowledge. As highlighted in Coxhead and Boutorwick's (2018) study described above, EAL learners not only enter school with lower levels of vocabulary knowledge than their FLE peers but also take longer to master the vocabulary required to succeed academically.

Reading comprehension is one important area where a lack of vocabulary knowledge can be detrimental to EAL learners' academic success (Murphy & Unthiah, 2015) and research shows that EAL learners often struggle with this important academic skill (Droop & Verhoeven, 2003). A lack of vocabulary knowledge is one important factor as to why these learners often struggle with reading comprehension, learners who are unable to master the vocabulary that is used in their textbooks have been shown to have difficulties in comprehending these texts (Coxhead et al., 2010). While it is as yet unclear the exact number of words EAL learners require to be successful academically, this number is likely dependent on both the grade and the subject (Green & Lambert, 2019; Greene & Coxhead, 2015).

Vocabulary knowledge is important for learners across subjects, and even in technical subjects like math and science vocabulary knowledge has been shown to be a strong predictor of academic success (Trakulphadetkrai et al., 2020). However, the amount of vocabulary EAL learners would need to acquire to succeed in these subjects using existing word lists is extremely large. Coxhead et al.'s (2010) found that with the BNC between 11,000- to 15,000-word families were necessary to reach the 95% coverage necessary to read with assistance in subjects like maths and science. Given

that research shows that it is common for EAL learners to struggle to master even the 2,000 most frequent words (Brooks et al., 2021; Coxhead & Boutorwick, 2018), it is not surprising that they struggle when trying to comprehend the lexically complex textbooks that they are required to read for these subjects. Difficulties with reading comprehension has, in turn, been cited as one of the primary reasons that EAL learners struggle academically (Murphy & Unthiah, 2015). This is evident in the UK where both national test scores (Burgoyne et al., 2009; Strand et al., 2015) and current research (Murphy & Unthiah, 2015) show that English reading comprehension is one of the primary factors behind why EAL learners struggle academically compared to their FLE classmates.

While it is clear that teachers need to provide support for EAL learners to help them improve their vocabulary knowledge, it is less clear what words they should focus on. One of the reasons that word lists are such an important tool for pedagogical purposes is that they help to identify the most important vocabulary for a specific circumstance (P. Nation, 2016). The ability of domain-specific word lists to target the type of texts a learner is likely to encounter in these circumstances is important, because learners are likely to require a different set of words depending on whether they are trying to read a novel, a medical textbook, or a legal document. We know from the research that learners need to encounter a word between six (Rott, 1999) to ten times (Pigada & Schmitt, 2006) in a meaningful context to learn it. We also know that vocabulary follows a predictable pattern, with lower frequency vocabulary items appearing very few times in the text (P. Nation, 2016), so it is unlikely that learners will be able to acquire the vocabulary they need simply by reading their textbooks. However, if we use a general word list such as the BNC, the lower frequency word bands provide little coverage of academic texts, less than 2% for every 1,000 word families for frequency bands below the 4K level (Coxhead et al., 2010). Furthermore, EAL learners often have problems mastering even the high-frequency vocabulary (Coxhead & Boutorwick, 2018), so it would be impractical for these learners to study all the BNC with which they were not familiar. To address these concerns, we need a better understanding of the actual vocabulary these learners need in the classroom; which is what an EAL word list would provide.

2.4.2 What a specialized EAL word list would look like

For a corpus to be useful, it must be representative of the types of texts that one would normally encounter in the domain that it purports to cover (Nation, 2016). This involves consulting with teachers and finding out what texts are being used in the classroom, and what subjects are being taught. To do this, it is also necessary to decide what grades the corpus will cover. For example, in a senior high school corpus a scientific word list would need to be broken up into specialized subjects such as Biology and Chemistry, as they usually teach these separately at the secondary level. However, if one were developing a scientific word list for middle school students, it would probably be preferable to group these subjects into a more general Science word list. It is also necessary to decide if each subcorpus will be confined to a single grade or cover the whole of that subject at the secondary school level. Whilst creating individual grade-level subcorpora may make it easier to focus on the words needed at each grade level, it may make it more difficult to remove technical words from the word lists.

While a secondary school corpus needs to cover a variety of different subjects, each subject must contain sufficient texts in order to meet the optimal size for the corpus. While there are benefits in small corpora, particularly for learner corpora and corpora being used for pedagogical purposes (Nation, 2001), it is common for corpora to contain far more than a million words. For example, Greene and Coxhead's (2015) corpus of middle school textbooks in the USA contained 18 million words, and Green and Lambert's (2018) secondary school corpus contained 16 million words. While more words are better, and it is possible to analyze very large corpora using modern computer programs (for example, the English Web 2020 Corpus has 38 billion running-words), preparing the types of text needed for an EAL corpus would require substantial amounts of work as the textbooks would need to be scanned, made useable using OCR, and then the resulting text files would need to be cleaned.

Another issue is the balance of the corpus, something that is necessary because texts vary in length. For example, the number of running words in the four textbooks in the study by Coxhead et al. (2010) ranged from just over 56,000 for Year 12 up to just over 88,000 for Year 11. In a small corpus (approximately 250,000 running words) of English texts for junior and senior secondary school students (Coxhead, 2012), one

novel contained just over 120,000 running words, which means it was almost half the running words of the whole corpus. In secondary school textbooks, some subjects tend to have more words in the text than other subjects. For example, in Green and Lambert's (2018) secondary school corpus, the Math subcorpus had just under 1.5 million words, while the History subcorpus had nearly 2.5 million words. For the EAL corpus, it would be necessary to find additional texts for some subjects to balance the different subcorpora.

2.4.3 What type of coverage should an EAL word list provide

As we saw in section 2.2 of this dissertation, 95% coverage is widely considered to be the minimum coverage necessary for learners to comprehend a text. Given this coverage, one would, ideally, want a words list to help learners reach at least this level of coverage for the texts that learners are expected to read. However, the coverage provided by the word list needs to be balanced with another factor, the size of the list itself. A word list with too many words, would be unwieldy and difficult for both teachers and students to use. Nation (2016) notes that while longer word lists can be useful for material development, word lists longer than 1,000 words can seem overwhelming for teachers and be difficult to integrate into the classroom. There are two well-known word lists that help to show the ideal length of a word list. One of these is the AWL (Coxhead, 2000) which owes it success in part because of its manageable length. The 570 word families that make up the AWL are subdivided into smaller groups, making it possible for teachers to focus on a manageable number of words in a class or over a semester (P. Nation, 2016). Another is the Essential Word List, (Dang & Webb, 2016) which is exactly 800 words long. The authors chose this length because they found that 800 to be the most practical length for a word list designed for English as a foreign language (EFL) learners. A shorter word list would provide much less coverage, and a longer word list would be impractical to learn over the two-year period most EFL learners spend studying English at a university level. Given the need to keep the EAL word lists to a manageable length, it is worth asking what type of coverage a word list of fewer than 1,000 words could reasonably be expected to provide over a corpus of texts.

The coverage a word list can provide over a corpus of texts is determined by an empirical statistical law called Zipf's law (1935). While the implications of the law itself extends beyond the field of linguistics (Malvern et al., 2004), when it applies to words in a corpus it states that there will be a regular inverse relationship between the rank of the word and the frequency in which the word appears in any corpus. The most frequent word in a corpus will occur significantly more times in the corpus than the less frequent words. Nation (2016) gives the example of an idealized corpus of 10,000 words. In such a corpus, we could expect the rank times the frequency of each word to equal around 700 for all the words in the text. This means that the most common word, *the* with a rank of 1, would occur 700 times in the corpus, while the next most common word, *and* with a rank of 2, would occur approximately 350 times in the corpus, and so on. This law has been consistent over many corpora over including general, academic, technical, written, and spoken corpora (P. Nation, 2016) and has several important implications for this study.

First, most of the coverage of a corpus will comprise a few highly frequent words, the 100 most frequent words in English make up approximately 50% of the running words in any corpus. A high percentage of these 100 words will be function words (Schmitt & Schmitt, 2020). Because it is common practice to remove both function words and general high frequency words from academic and specialized word lists, either through the use of general word lists (e.g., Coxhead, 2000; Lei & Liu, 2016) or through statistical means (e.g., Gardner & Davies, 2014; Green & Lambert, 2018), the total amount of coverage these types of word lists can provide over a corpus is limited. However, as learners are likely to already be familiar with high-frequency words, because they will have encountered them multiple times while reading both their classroom textbooks and more general texts (P. Nation & Waring, 2019), these are words we can reasonably expect EAL learners to know. From a pedagogical standpoint, it is not a problem that these words are excluded from the final word lists, but it will significantly reduce the coverage that these word lists can provide.

The other issue that needs to be considered is that we can expect there to be a high number of very low frequency words in any corpus. According to Zipf's law, almost half of the words in any corpus will occur only one time (Coxhead, 2017; P. Nation, 2016). Given the extremely low frequency of these technical words, they are not

worth including on a word list. Because of these two constraints, the need to exclude both general high frequency vocabulary and very low frequency words, it will be very difficult to construct a word list that provides the 95% coverage EAL learners need to comprehend the texts that they are being asked to read. If reaching this 95% coverage is not possible, we are then left with two questions. First, is it still worth developing a set of word list that will not provide learners with the 95% coverage they need for understanding, and second, what level of coverage would a successful set of EAL word lists be expected to provide?

The answer to the first question is yes. Even word lists that do not provide 95% coverage can be helpful for learners. In an international school classroom, the purpose of these word lists would be to allow teachers to decide what words to focus on when teaching a subject. In this context, teachers should focus on mid-frequency vocabulary that is important for the subjects that they are teaching (Greene & Coxhead, 2015). A set of word lists that helps to identify what this vocabulary is will allow teachers to use what Nation (2016) describes as a field approach to teaching the vocabulary necessary for that subject area. This approach involves the teacher examining the material that they were using in the classroom and dividing the words in these materials into three groups, high-frequency words that they would expect the learners to already know, mid-frequency, academic, and domain-specific words that learners probably do not know but need to learn for that subject, and low frequency and technical words that can be removed or glossed (the teacher can explain the meaning in the learners' L1 or using simple vocabulary to explain the word prior to having the students read, without having the students learn the word). The EAL word lists would be invaluable in identifying this middle group of words, those words that learners would probably not know, but that are important to learn in order to be literate in that subject. The teacher could then gloss or pre-teach the lower frequency, topic specific words, or provide tools that learners can use to better understand these words (see, Chung & Nation, 2004) so that the learners could focus on acquiring the important academic and subject specific words in the lesson. They could also use the word lists to incorporate a number of vocabulary specific activities into the classroom, such as concept maps or vocabulary journals using these words. This approach would allow for learners to focus on the important words through both direct and incidental learning, making it easier for them to acquire those

words. In other words, rather than viewing the word lists as a blueprint that tells teachers exactly what words are necessary to reach 95% coverage, the word lists should instead be seen as providing guidance in how to direct the learner's attention by allowing them to focus on the most important and most salient vocabulary for each subject.

Therefore, if a word list can be successful without providing 95% coverage, what type of coverage should a successful word list provide. The best way to answer this question is by looking to the research. The word lists that I will compile for this dissertation are domain specific academic word lists. By looking at what coverage other domain specific academic word lists provide over the domain that they represent, we can get a better understanding of the type of coverage I am looking to achieve with the words lists that I hope to develop. As I have stated previously, the most famous academic word list is Coxhead's (2000) AWL, which provides between 8% to 12% coverage of most academic texts (Coxhead, 2011). Greene and Coxhead's (2015) domain specific academic Middle School vocabulary lists provide similar coverage (5.83% for Social Studies to 10.17% for Science). Using more modern techniques, Green and Lambert (2018) could achieve from 13% (for English) to 25% (for Chemistry) coverage from their SVL on the domain that the lists were compiled for. If I am able to construct a set of word lists that provide over 8% coverage (the lower end of coverage provided by Coxhead's AWL) and provide better coverage over the textbooks that EAL learners are required to read in their classes than existing word lists, I would consider these word lists to be successful. This is because, if they were to meet these criteria, these word lists would be an effective pedagogical tool that teachers could use to support EAL learners in the classroom and would allow EAL learners to focus on the vocabulary that they need to succeed academically.

2.5 Conclusion

The literature review has examined how our knowledge of what vocabulary EAL learners need has evolved. The examination appears to suggest that, while there have been significant improvements in how we measure vocabulary in this context, there is still more that needs to be done. The experimental chapters build upon these earlier studies by first showing why the existing word lists discussed in the literature review are

not sufficient and then detailing the steps taken to develop and validate an EAL specific word list. A brief summary of each experimental chapter follows.

Chapter Three describes a pilot study that examines the importance of vocabulary for EAL learners' academic success and tries to determine the effectiveness of using existing vocabulary assessment tools with this group. The research questions for Chapter Three are:

1. How effective is an existing vocabulary assessment (McLean & Kramer, 2016) at measuring the vocabulary proficiency of EAL learners?
2. What is the correlation between these learners' vocabulary knowledge and their reading comprehension?
3. How do other factors, such as word decoding skills, reading fluency, and general English proficiency affect their academic proficiency?

Chapter Four is a partial replication of Coxhead and Boutorwick's (2018) study that looks at the vocabulary knowledge of EAL learners in the context of the texts that they are likely to encounter in the classroom. This replication addresses some issues with previous studies as it includes a significantly bigger corpus, covers over three subject areas, and makes use of a newer version of the Vocabulary Levels Test. The chapter has three research questions:

1. What level of vocabulary knowledge do learners studying in an international school setting in Japan have at different grade levels?
2. What are the vocabulary profiles of the textbooks that these learners are using in the classroom?
3. What coverage do other word lists (GSL, AWL, Middle School Vocabulary Lists, SVL) provide over the same set of textbooks compared to the BNC/COCA?

Chapter Five describes the development of an international school corpus and the development of the initial word lists. The research question for Chapter Five is: What coverage can a word list developed specifically for international school students provide for a corpus of secondary school textbooks likely to be used in this setting? This chapter describes how these texts were selected, the steps involved in the digitalization

and cleaning of the texts, and provides a discussion of how these initial word lists were created by building upon the research of Coxhead (2000) and Greene and Coxhead (2015). The chapter will also introduce the initial discipline-specific word lists that will be improved upon in the subsequent chapter.

Chapter Six describes a revision of the word lists compiled in the previous chapter using more modern techniques to see if these techniques can be used to build a more effective set of word lists. I compiled the word lists in this chapter building upon the techniques used by Gardner and Davies (2014), Lei and Liu (2016), and Green and Lambert (2018). In the chapter, I explain the benefits of using these more modern techniques and compare these word lists with the ones I compiled in Chapter Five. I also examine the coverage of these word lists on other corpora, including a non-academic corpus, corpora from other domains, and a domain specific corpus compiled to validate these word lists. To determine the effectiveness of the new word lists, I also compare the coverage of these lists to the coverage provided by existing word lists, including the BNC/COCA, the AWL, the Middle School Vocabulary Lists, and the SVL.

Finally, in my discussion chapter, I aim to discuss the findings from the experimental chapters and examine them considering the literature review. The concluding chapter collects the main strands of the thesis and proposes possible areas for future research.

Chapter Three

The Importance of Vocabulary for EAL Learners' Reading Comprehension

3.1 Introduction

Chapter Two summarized two important strands of research that are necessary to consider when investigating the importance of vocabulary for EAL learners. First, there is a need to better understand what words to use in the assessment process of EAL learners and to develop context-specific ways to assess these words. As Schmitt et al. (2020) point out, there is no one size fits all vocabulary assessment, and it remains necessary to consider the stated purpose of assessment tools before using them. However, at present, there are no vocabulary assessment tools that have been developed specifically for EAL learners. This issue is further compounded by the fact that little research has been conducted on validating the effectiveness of existing general and academic word list assessment tools in the EAL context. Second, to determine the impact the knowledge, or lack of knowledge, of the words in more general word lists will have on EAL learners, these word lists need to be examined in more detail in this context. We know from previous studies (e.g., Ardasheva & Tretter, 2017; Coxhead & Boutorwick, 2018; NALDIC, 2015, October 27) that EAL learners typically enter the educational system with significantly lower levels of vocabulary knowledge than their First Language English (FLE) counterparts. Studies have also shown that in an English as a Medium of Instruction (EMI) context EAL learners often take longer than FLE learners to master the vocabulary necessary for academic success (Coxhead & Boutorwick, 2018). While these studies support the assertion that EAL learners are likely to have less developed vocabularies than their FLE classmates, it is as yet unclear the extent to which this gap in vocabulary knowledge will impact EAL learners' ability to succeed academically.

In this chapter, therefore, I respond to the two issues discussed above, by examining the degree to which an existing vocabulary assessment tool, the new Vocabulary Levels Test (nVLT, McLean & Kramer, 2016), can predict EAL learner academic performance. To establish potential correlations between this assessment and academic performance, the current chapter will focus on one area that has been shown to be extremely important for EAL learner academic success: reading comprehension

(Droop & Verhoeven, 2003). In the UK, both national test scores (Burgoyne et al., 2009; Strand et al., 2015) and current research (Murphy & Unthiah, 2015) show that lower levels of English reading comprehension is one of the primary reasons why EAL learners struggle academically. One issue that EAL learners may struggle with while reading is their lower levels of vocabulary knowledge (Murphy & Unthiah, 2015). Vocabulary is central to the reading process, as I outlined in section 2.2.1, and learners who are unable to master the vocabulary that is used in the texts they are required to read in the classroom often struggle to understand those texts (Coxhead et al., 2010).

Previous studies that have investigated the importance of vocabulary for reading comprehension have mainly focused on adult EFL or ESL learners (Laufer & Ravenhorst-Kalovski, 2010; Qian, 2002). While there have been a few studies that have investigated the importance of vocabulary for young learners, these were conducted in the context of EFL language classes (Henriksen et al., 2004; Stæhr, 2008). While these can provide some insight into the importance of vocabulary for EAL learners, studies that have investigated vocabulary knowledge across learning environments suggest that the relationship between vocabulary and reading comprehension varies according to such factors as the participants' age (Schoonen et al., 1998), their linguistic background (Geva & Farnia, 2012), and the learning context (Miralpeix & Muñoz, 2018). An understanding of the relationship between vocabulary and reading comprehension is further complicated by the fact that a number of additional factors have also been shown to influence reading comprehension (Melby-Lervåg & Lervåg, 2014). These include: i) the ability to decode the orthographic representation of words in English (Droop & Verhoeven, 2003; Melby-Lervåg et al., 2012); ii) reading fluency (C. A. Fraser, 2007; Geva & Zadeh, 2006); and, iii) general language proficiency (Trakulphadetkrai et al., 2020). In the following discussion I will examine the importance of vocabulary and these three additional factors for reading comprehension in more detail in order to illustrate why they are important for the study described in the current chapter.

The first factor to consider when examining reading comprehension is vocabulary knowledge. Research has shown vocabulary knowledge to be a key predictor of reading comprehension in the classroom for EAL learners (Lervåg & Aukrust, 2010; Melby-Lervåg & Lervåg, 2014) as well as for their FLE peers (Ouellette & Beers, 2010; Tunmer & Chapman, 2012). Poor vocabulary skills have been shown to

significantly limit learners' ability to understand written texts (Murphy & Unthiah, 2015). As I discussed in Section 2.2.3 of the previous chapter, studies indicate that vocabulary coverage of 95% to 98% is a necessary condition for comprehension (e.g., Hsueh-Chao & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). Learners who are not able to reach this vocabulary threshold will struggle to comprehend the texts they are required to read for their classes. Additionally, if EAL learners are not able to understand the texts that they are reading as a result of gaps in their vocabulary knowledge, it will be difficult for them to complete tasks and answer questions related to these texts (Burgoyne et al., 2009). This can result in EAL learners getting lower grades and struggling academically.

The second important factor for reading comprehension that I will discuss is word decoding skills. In order to be able to use lexical information to understand the meaning of a text learners first need to be able to effectively decode the words in the text, something referred to as word recognition (Hoover & Gough, 1990). According to Gough and Tunmer (1986) the ability to decode the words of a text is a separate skill from comprehension. This is why it is possible for some learners to have poor decoding skills (such as those learners with dyslexia) but still be able to comprehend what they are reading. The opposite is also true, some learners, who are referred to as poor comprehenders in the literature (Yuill & Oakhill, 1991), can exhibit good word decoding skills but are poor at comprehending what they are reading. Research, however, suggests that such findings are equivocal for EAL learners. Most of the research on EAL learners' reading comprehension shows that even though they often possess good word decoding skills EAL learners struggle with reading comprehension (Hutchinson et al., 2003; Murphy & Unthiah, 2015). Despite the fact that most studies show EAL learners to be good decoders, there are a number of studies (e.g., García & Cain, 2014; K. Nation & Snowling, 2004) that indicate there may be situations where EAL learners struggle with their word decoding skills.

A third factor that has been shown to influence reading comprehension relates to reading fluency. Reading fluency is defined as the ability to quickly and accurately read a text with the proper expression (National Reading Panel, 2000). In other words, a fluent reader is one that can read both fast and accurately (Grabe, 2010). Reading fluency is an important indicator of a learners' reading comprehension because it is

indicative of the amount of cognitive resources they need to allocate towards word decoding and word recognition (Adlof et al., 2006). As it becomes easier for a reader to decode and recognize words, not only do they become more fluent, they are also able to allocate more of their limited cognitive resources towards comprehension (Geva & Zadeh, 2006). This is true for both L1 (Geva & Zadeh, 2006) and L2 (C. A. Fraser, 2007; Jiang et al., 2012) readers, and studies have shown a strong and significant correlation between fluency and reading comprehension for both groups of learners.

The fourth and final factor I examine in this study relates to general language ability. General language ability is important because it allows for individual learner language proficiency to be conceptualized as both unitary and divisible, depending on the level of abstraction and the purpose of the assessment (Harsch, 2014). Including a measure of the participant's general language ability in this study allowed me to investigate both holistic and discrete measures of language proficiency. Accordingly, I was able to explore the extent to which measures of specific components of language ability, such as the participants' vocabulary knowledge, were able to account for variance in reading comprehension over and above measures of the participants' general language ability.

To determine the effect these four different factors have on reading comprehension in the EAL context, I employ a reading assessment that was designed and validated to assess the reading levels of secondary school students. The reading assessment used in this study represents a departure from previous studies, such as Laufer & Ravenhorst-Kalovski (2010), which have used the reading comprehension scores from large-scale high-stakes assessments. The reading sections from institutionalized assessments such as the TOEFL or TOEIC exams are not representative of what EAL learners are required to read as part of their classes. In order to address this issue, in this study I used a reading assessment developed using the Simple View of Reading. This assessment tool is called the York Assessment of Reading Comprehension (YARC, Snowling et al., 2009) and it is a comprehensive test of reading that was developed to assess the reading comprehension level of 11- to 16-year-old learners studying in the UK. While the test was mainly developed for FLE speakers, it has also been used to investigate the reading comprehension ability of EAL learners. The original test development included data from 89 EAL learners studying in

the UK. Previous studies have also used and validated the YARC with L2 English learners' (e.g., Treffers-Daller & Huang, 2020). While the scores of L2 speakers are likely to be lower than those of FLE speakers, it is still possible to use the test to gain insight into how well EAL learners are able to meet the reading comprehension levels required for their grade level (see https://www.gl-assessment.co.uk/support/yarc-support).

The current chapter examines the relationship between reading comprehension and vocabulary in a specific group of learners, young EAL learners. Building on earlier studies that have investigated how different factors can affect reading comprehension, the study that I describe in the current chapter examines the relationship between reading comprehension and the four areas most cited in the literature as impacting learners reading comprehension, namely: (i) word decoding skills; (ii) vocabulary; (iii) reading fluency; and, (iv) general linguistic ability. My three research questions for this chapter are:

1. What level of vocabulary mastery was exhibited by the EAL learners compared to their FLE counterparts?
2. To what degree does vocabulary knowledge correlate with reading comprehension, and how does this relationship compare with the way reading comprehension correlates with other factors such as the learners' word decoding skills, fluency, and general language ability?
3. Does vocabulary knowledge have a significant and independent effect on learners' reading comprehension ability?

3.2 Methodology

3.2.1 Subjects

The study took place at an international school located in western Japan. In the school, outside of a specific Japanese language class, all of the instruction is delivered in English. The school classes follow the International Baccalaureate for Middle School and Diploma programs. There were 31 learners (N=31) who participated in the study (11 males and 20 females). The participants came from a diverse mix of nationalities and language backgrounds including Japanese, Korean, Dutch, and Croatian and

represent the type of heterogeneous population that is commonly seen in the international school context (e.g., Coxhead & Boutorwick, 2018). For this study, I grouped the students into the same three categories that were used by Coxhead and Boutorwick (2018): FLE speakers, EAL speakers, or Non-native Speakers of English as an Additional Language (NNSEAL). Six participants (four male and two female) were classified as L1 (English) speakers, and 25 (seven male and 18 female) participants were identified as EAL learners. Using Coxhead and Boutorwick's (2018) methodology, the 17 EAL learners (five males and 12 females) who required additional language support in the classroom were classified as NNSEAL. All the participants in the study ranged in age from 11 to 15, with the exception of one participant who had just turned 16. All the participants had studied for at least two years in the international school context.

Prior to engaging in the study, I obtained informed consent from the learners and their respective parents or guardians. The consent forms given to the parents are shown in Appendix A. Due to the age of the participants, collecting informed consent was a detailed and collaborative process between all the stakeholders involved, including myself, my university, the principal of the school, the parents, and the students. I developed the initial consent forms using the procedures in place at the university where I was employed at the time and in consultation with researchers who had experience working with minors. I then sent these forms to the principal of the school where the study was conducted and updated the forms to take into any concerns that he had. One important change that I could make to the forms using the feedback from the principal was to include both a "Yes, I agree to be part of the study" and a "No, I don't agree to be part of the study" box. Including two boxes made it easier to follow up with students who had not returned their forms because, without this box, it would not have been possible to have determined whether students had not returned the form because they had either forgotten or because they did not want to participate in the study. After receiving approval for the updated forms from the university, the principal, and the classroom teachers, they sent the finalized forms to the parents. Included with the consent form was a detailed description of what they would ask the participants to do as part of this study in both Japanese and English. This description also explained how I would safeguard participant privacy, assurances that this was not part of the regular

school assessments, and that the students were not required to participate in the study and could withdraw at any time without it affecting their grades. I included both my email address and phone number in the project's description. I received two phone calls from parents with questions about the study. I was able to respond to the concerns that these parents had, and both sets of parents signed the consent forms. The classroom teachers also gave the same description to their students in the classroom, both when they disseminated the forms and on the day of the data collection. They included a further consent checkbox at the bottom of the online survey to give students an additional opportunity to withdraw from the study. I removed the data of any students who did not return the form or who checked the No box on either the consent form or the survey from the analysis. This research has been approved by the Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University (approval number: HR-HUM-000762).

3.2.2 Survey Instruments and Procedure

I collected the data over a two-month period, between December 2017 and January 2018. During the first session, I gave the learners a survey designed to provide insight into their language background (see Appendix B) along with the nVLT and the C-Test. During the second session, I interviewed the learners individually. These interviews involved giving learners a Single Word Reading Test (SWRT) along with Snowling et al.'s (2009) York Assessment of Reading for Comprehension (YARC). It took the learners about an hour to complete the assessments in the first session and the SWRT and the YARC took approximately 45 minutes per learner.

The first session was done during regular class time, and the learners completed the language survey, the C-Test, and the nVLT; a short description of each follows and examples of the assessment tools are provided in Appendix B. The learners' regular classroom teacher distributed these assessment tools and monitored this session. The learners were able to complete the survey and assessments within the hour scheduled for these assessments, most of them took approximately 45 minutes to complete all three. The learners completed the survey first and then completed McLean and Kramer's (2016) New Vocabulary Levels Test (nVLT). The nVLT is a multiple-choice test with 24 questions for each of the first 5,000-word frequency bands along with an additional

section that covers Coxhead's (2000) AWL. Each item on the test comprises the target word, both by itself and in a simple sentence. There are four answer choices for each question, from which examinees must select the word or phrase with the closest meaning to the target word (see Figure 3.1). I chose this test over previous versions of the VLT (e.g., Nation, 1990; Schmitt et al. 2001) for two reasons. First, it covers more frequency bands; previous assessments only measure the second, third, fifth, and 10th frequency bands. Second, earlier VLTs were designed using target words and distracters selected from outdated frequency lists such as West's (1953) General Service List (GSL).

Figure 3.1

An Example of a Question on the nVLT

1. time: They have a lot of **time**.
a. money
b. food
c. hours
d. friends

Note. Adapted from "The Creation of a New Vocabulary Levels Test" S. McLean and B. Kramer, 2015, *Shiken*, *19*(2), p. 4.

Following earlier studies (Coxhead & Boutorwick, 2018; Read, 1988; Schmitt et al., 2001), I set the level of mastery of each frequency band to 86%. This meant that in order to be considered to have mastered a band, learners needed to get at least 21 of the 24 questions correct. For the section that covered the AWL, which included 30 questions, they needed to get 26 or more of the questions correct.

After they had finished the nVLT, the learners completed the C-test, which is a modified version of a cloze test. In this test, learners are given a text where parts of some words have been removed (e.g., 'The miss____ parts are gi___ in bo__.'). They are then required to fill in the missing parts of these words (e.g., 'The miss**ing** parts are gi**ven** in bo**ld**'). The C-Test uses the reduced redundancy principle (RRP) to measure the test takers' general language ability. It does this by introducing interference, the

missing parts of the words, so that the test taker is required to use their other linguistic skills to compensate (Babaii & Ansary, 2001). The C-Test that I used in this study had been validated in several previous studies (Ishihara et al., 2003; Neff, 2015) and I chose it because the texts in the assessment related to situations and contexts that would be easily understood by EAL learners studying in Japan.

In the second session I gave the participants two different assessments. The first of these assessments was a test of their word decoding skills, the SWRT. The SWRT assessment is made up of 70 words that are grouped into bands of ten. The bands become progressively more difficult as test-takers advance through the test. The SWRT serves a dual purpose: the first is to provide a measure of the learner's word decoding skills, and the second is as a diagnostic that allowed me to place the test takers in the correct level of the YARC test. I then used the YARC test (Appendix B) to assess the learners' reading ability (accuracy, comprehension, and fluency).

3.2.3 Scoring the YARC

I used the participant's SWRT scores to determine which of the three levels of the YARC test to give them. Each of the levels is made up of three different sections. For the first section I had the learners read and answer questions about two different passages. The genre of the first passage was fiction and second was non-fiction, learners had to orally answer thirteen comprehension questions about each passage. While answering the questions, the participants were allowed to refer to the passage. For the second part of the YARC, the participants had to give an oral summary of the passage that they just read without looking back at the passage. I awarded participants a point for each of the key details from the passage that they were correctly able to recall, the maximum possible points for each of the summaries was either eight or nine, depending on the passage. The final part of the YARC assessment measured the participants reading fluency. For this section, I asked the participants to read a passage out loud. The passage contained 129 or 137 words, depending on the level. I awarded one point for each word read correctly, and I calculated the reading rate by dividing the number of words read correctly by the time taken to read the passage. I then used the test taker's age, the difficulty of the passage, and the speed at which they were able to correctly read the words in the passage to award them a reading fluency score out of 130.

I used the standardized scores for the YARC test, rather than the raw scores, for analysis. I did this because the YARC consists of different passages for different proficiency levels, which means that not all of the participants read the same passages. The YARC provides a way to convert the raw scores into an ability score, which provides an estimate of the participant's level based on the difficulty of the passage. I then converted these ability scores into a standardized score out of 130 to determine the participant's reading comprehension performance in relation to standardized norms.

3.2.4 Data Collected

Using these assessments, I was able to collect five separate data points for each participant: (1) the C-Test provided a measure of their general language ability in English; (2) the nVLT provided a measure of their vocabulary knowledge; (3) the SWRT provided a measure of their word decoding skills; (4) the fluency passage of the YARC provided a measure of their reading fluency; and, (5) the YARC reading comprehension assessment provided a measure of their reading comprehension. The means and standard deviations for these measures' are given in Table 3.1.

Table 3.1

Means and Standard Deviations of all Variables (N = 31)

	Mean	SD	Minimum	Maximum
Reading Comprehension	108.19	12.74	79	128
nVLT	124.19	22.27	58	147
C-Test	77.23	20.72	14	98
SWRT	58.06	7.97	35	69
Fluency	104.13	14.63	70	130

Note. Reading comprehension = YARC reading comprehension ability scores (maximum 130). nVLT = Raw scores of the new Vocabulary Level's Test (maximum 150). C-Test = Raw scores of the C-Test (maximum 100). SWRT = Raw scores on the Single Word Reading Test (maximum 70). Fluency = YARC reading fluency standardised scores (maximum 130)

3.3 Results

Research Question 1: What level of vocabulary mastery was exhibited by the EAL learners compared to their FLE counterparts?

An overview of the nVLT scores (see Table 3.2) indicates that there was a high number of learners who had not achieved mastery of the mid-frequency word bands or the AWL. Mirroring the results found in previous studies (Coxhead & Boutorwick, 2018), I also found that prior to Grade 9 a high number of the participants still could not master the high-frequency words in the 2,000 and 3,000-word bands, and learners from all grade levels struggled to master the AWL. Given the importance of the AWL for the understanding of school textbooks (Greene & Coxhead, 2015), a lack of mastery of

these vocabulary items would mean that these learners would struggle to read at their grade level.

Table 3.2

Mastery of Vocabulary by Grades (All Learners)

	1000	2000	3000	4000	5000	AWL
Grade 6 to 8 (n = 17)	100%	65%	29%	53%	53%	29%
Grade 9 to 10 (n = 14)	100%	100%	43%	43%	64%	57%

Note. AWL = Academic Word List

Furthermore, EAL participants were much more likely to struggle to master high-frequency vocabulary than FLE learners (Table 3.3). For example, only one of the FLE participants (a Grade 7 student) did not demonstrate mastery of the 5,000 most frequent word families. However, most of the FLE learners at this level did not show mastery of the AWL, showing that vocabulary knowledge may still be an issue even with this group of learners.

Table 3.3

Mastery of Vocabulary by Language Background

	1000	2000	3000	4000	5000	AWL
FLE (*n* = 6)	100%	100%	83%	100%	100%	67%
EAL (*n* = 25)	100%	76%	28%	48%	44%	28%

Note. FLE = First Language English; EAL = English Language Learner; AWL = Academic Word List

Research Question 2: *To what degree does vocabulary knowledge correlate with reading comprehension, and how does this relationship compare with the way reading comprehension correlates with other factors such as the learners' word decoding skills, fluency, and general language ability?*

I conducted a bivariate correlational analysis using the scores on the assessments in relation to the participants' reading comprehension scores. Table 3.4 summarizes all Pearson correlations between the variables for all learners. The correlations between reading comprehension and the four factors assessed reveal that the YARC reading comprehension correlates most strongly with the nVLT (.86***), and the C-test (.83***). The learners' SWRT and reading fluency demonstrated moderate and statistically significant correlations with reading comprehension, with *r* values of .67*** and .70*** respectively. That both the participants' reading rates and their word decoding skills correlated moderately and significantly with their YARC reading comprehension scores means that the test potentially taps into the word decoding dimensions of reading detailed by the Simple View of Reading.

Table 3.4

Correlations Between Variables (N = 31)

	C-Test	SWRT	YARC Fluency	YARC Comprehension
nVLT	.90*** [.79, .95]	.86*** [.73, .93]	.73*** [.50, .86]	.86*** [.73, .93]
C-Test		.85*** [.70, .92]	.71*** [.47, .85]	.83*** [.68, .92]
SWRT			.74** [.52, .87]	.67*** [.42, .83]
YARC Fluency				.70*** [.47, .85]

Note. SWRT = Single Word Reading Test; YARC = York Assessment of Reading Comprehension; nVLT = New Vocabulary Levels Test.

* = correlation significant at $p < .05$; ** correlation significant at $p < .01$; *** correlation significant at $p < .001$ Values in square brackets indicate the 95% confidence interval for each correlation.

I found similar correlations when examining data from only the EAL learners (Table 3.5). Again, for this group of learners, both the nVLT and C-Test correlate most strongly to reading comprehension with r values of .86*** and .87*** respectively.

Table 3.5

Correlations Between Variables for EAL Learners (n = 25)

	C-Test	SWRT	YARC Fluency	YARC Comprehension
nVLT	.90*** [.78, .96]	.89*** [.75, .95]	.71*** [.43, .86]	.86*** [.70, .94]
C-Test		.84*** [.67, .93]	.74*** [.49, .88]	.87*** [.73, .94]
SWRT			.74** [.49, .88]	.70*** [.42, .86]
YARC Fluency				.69*** [.41, .85]

Note. SWRT = Single Word Reading Test; YARC = York Assessment of Reading Comprehension; nVLT = New Vocabulary Levels Test.

* = correlation significant at p < .05; ** correlation significant at p < .01; *** correlation significant at p < .001 Values in square brackets indicate the 95% confidence interval for each correlation.

Research Question 3: *Does vocabulary knowledge have a significant and independent effect on learners' reading comprehension ability?*

I ran a partial correlation test (Table 3.6) to determine the relationship between the participant reading comprehension and their vocabulary ability whilst controlling for general language ability (C-test), reading fluency (fluency), and word decoding skills (SWRT). In order to help compensate for the small sample size, I did the partial correlational analysis using bootstrapping (Field et al., 2012). Bootstrapping provides a more robust method for examining small sample sizes by estimating the properties of the sampling distribution from the sample data (Bruce, 2015). Bootstrapping does this

by treating the sample data as the population and drawing smaller samples from this data, putting back the data before a new case is drawn. The correlational coefficient can then be calculated from each of these samples and the standard deviation of the sampling distribution of the bootstrapped samples can be used to estimate the standard error of the correlational coefficient (see Wright et al., 2011). From this standard error, confidence intervals and significance tests can be computed.

The partial correlational test showed that there was a strong and statistically significant partial correlation between the nVLT scores and reading comprehension (r = .57, $p < .001$) whilst controlling for the other variables. In this model, vocabulary can be said to account for approximately 33% of the variance seen in the participants' reading comprehension scores, showing that vocabulary does indeed have a strong and independent effect on the participants' reading comprehension.

Table 3.6

Partial Correlational Analysis of VLT and Reading Comprehension (N = 31)

Control Variables	Independent Variable			Reading Comprehension
C-Test, SWRT, & YARC Fluency	nVLT	Correlation		.573***
		Significance (2-tailed)		.001
		Bootstrap[a.]	Bias	-.057
			Std. Error	.186
		BCa 95% Confidence Interval Lower		.194
		Upper		.762

Note. SWRT = Single Word Reading Test; YARC = York Assessment of Reading Comprehension; nVLT = New Vocabulary Levels Test.

* = partial correlation significant at $p < .05$; ** partial correlation significant at $p < .01$; *** partial correlation significant at $p < .001$ a = bootstrap results are based on 2000 bootstrap samples

I then used an additional partial correlation analysis to determine if the effects of general language ability, word decoding, and fluency still correlated to reading comprehension after vocabulary knowledge had been accounted for (Table 3.7). In all cases, when I controlled for vocabulary knowledge, the other assessments did not indicate strong or significant correlational relationships with reading comprehension. This lack of a significant correlation seems to indicate that these factors do not explain a significant level of the variance in reading comprehension scores after vocabulary knowledge has been considered, further highlighting the importance of vocabulary for reading comprehension.

Table 3.7

Partial Correlational Analysis of the C-Test, SWRT, Fluency, and Reading Comprehension While Controlling for Vocabulary Knowledge (N = 31)

Control Variables	Independent Variable				Reading Comprehension
nVLT	C-Test	Correlation			.273
		Significance (2-tailed)			.144
		Bootstrap[a]	Bias		-.035
			Std. Error		.149
		BCa 95% Confidence Interval		Lower	-.027
				Upper	.721
nVLT	SWRT	Correlation			-.279
		Significance (2-tailed)			.136
		Bootstrap[a]	Bias		.027
			Std. Error		.198
		BCa 95% Confidence Interval		Lower	-.654
				Upper	.275
nVLT	YARC Fluency	Correlation			.225
		Significance (2-tailed)			.232
		Bootstrap[a]	Bias		-.003
			Std. Error		.139
		BCa 95% Confidence Interval		Lower	-.043
				Upper	.488

Note. SWRT = Single Word Reading Test; YARC = York Assessment of Reading Comprehension; nVLT = New Vocabulary Levels Test.

* = partial correlation significant at $p < .05$; ** partial correlation significant at $p < .01$; *** partial correlation significant at $p < .001$ a = bootstrap results are based on 2000 bootstrap samples

3.4 Discussion

The main purpose of the experiment reported in this chapter was to investigate the impact that vocabulary knowledge has on EAL learner reading comprehension. My findings are consistent with a growing body of EAL research that shows that there is a strong and significant relationship between EAL learners' vocabulary knowledge and reading comprehension (Burgoyne et al., 2009; Melby-Lervåg & Lervåg, 2014; K. Nation & Snowling, 2004). My results also show that the participants' vocabulary knowledge was more strongly correlated with their reading comprehension abilities than other factors, such as fluency or word decoding, which is a similar finding to Burgoyne et al. (2009). Even using the nVLT, which was not specifically designed for EAL learners, the findings highlight the importance of strengthening vocabulary knowledge to improve EAL learners' reading ability. In this discussion, I investigate two issues that arose from the study: the gap between EAL learners' vocabulary knowledge and their FLE counterparts, and the degree to which EAL learners of all age groups struggled with vocabulary knowledge.

The first issue relates to the difference between this study and previous studies regarding the degree to which EAL learners' vocabulary knowledge lagged their FLE counterparts. Looking at young learners studying in the UK, Hutchinson et al. (2003) found that even though EAL learners vocabulary knowledge lagged behind their FLE counterparts, the gap was fairly small. In the current study, my findings suggest a significantly larger gap in the vocabulary knowledge of EAL learners compared to their FLE peers. While this is different from studies such as Hutchinson et al. (2003) which were conducted in the UK, it is consistent with other studies that were conducted in an international school context (Coxhead & Boutorwick, 2018). One potential factor that could explain this difference is the context in which the studies were conducted. Hutchinson et al.'s (2003) participants were EAL learners studying in the UK, while the participants in this study and Coxhead and Boutorwick's (2018) study were studying in an international school context. The limited opportunities EAL learners have to use English outside of the classroom in the international school context may have resulted in the larger differences in vocabulary knowledge seen between the two groups.

The second important issue to consider is the lack of mastery shown by the EAL learners. There were several EAL learners who did not even exhibit mastery of the first 2,000 high-frequency word bands. Given that previous studies (Coxhead, 2012; Coxhead et al., 2010) have shown that learners need to know between 9,000 and 11,000 word families to reach the 95% coverage necessary for understanding, it is not surprising that EAL learners would struggle with these texts. Learners who could not master the 2,000-word frequency band would have coverage of less than 81% of the words in the texts. While this highlights the importance of vocabulary intervention for this group of learners, it also presents us with another issue: the sheer number of vocabulary items needed by these participants before they can even begin to understand these texts is such that even providing lexical support for these learners would be difficult. This is one area where EAL specific word lists would be beneficial for both teachers and learners. By focusing on the words that occur in the EAL context, it would be possible to gain greater coverage with fewer words (Nation, 2016), allowing teachers to focus on the words that were most important for their learners. The creation of the international school academic word lists described in Chapters Five to Seven represents a unique and important step towards answering this question.

One of the strengths of the study described in this chapter is the inclusion of tests for general language ability (C-Test) as well as word decoding skills (SWRT) alongside vocabulary knowledge. Including these additional tests enabled me to examine the potential contributions different aspects of language knowledge can make towards EAL learner reading comprehension. This is an important area of research, and a number of recent studies have begun to investigate how these different aspects of linguistic knowledge, along with general language proficiency, correlate to skills such as listening comprehension (Wang & Treffers-Daller, 2017), reading comprehension (Droop & Verhoeven, 2003), and mathematics (Trakulphadetkrai et al., 2017). The results from the current chapter add to this growing body of research. One interesting finding with regard to the relationship between general language ability and reading comprehension is that the strong correlation seen between these two factors suggest that less successful readers focus on smaller units (word or sentence level) when constructing meaning from a written text, something that has been suggested by recent studies (Trakulphadetkrai et al., 2017). On the other hand, more proficient and

successful learners seem to use a wider range of top-down, global strategies for text comprehension. Because the C-Test requires learners to use a combination of bottom-up, word-based strategies along with top-down, text-based strategies (Babaii & Ansary, 2001) it could prove to be a useful tool for use in EAL classrooms to help identify learners across a range of proficiency levels who are struggling with reading comprehension. However, while the results from this study are promising, my sample size is still too small to draw any major conclusions.

Despite the potential limitations that I discussed above, there are two important implications that we can draw from the current study. First, to improve learners' reading comprehension, EAL teachers should focus first and foremost on improving their students' vocabulary, something that has been advocated by other researchers (e.g., Coxhead et al., 2010; Green & Lambert, 2019). Because I did not look at vocabulary intervention in the current study, I cannot use my results to make any suggestions regarding what specific vocabulary activities teachers should use in the classroom. However, previous studies have shown that the best way to teach vocabulary in the classroom is to follow a principled approach to vocabulary instruction and provide learners with the opportunity to engage in both intentional and incidental vocabulary learning (Graves et al., 2012; Schmitt, 2008). Second, my results from this chapter highlight the benefits of using vocabulary assessment as a means of identifying learners who may require additional language support in the classroom; something that has also been noted by other researchers in the field (e.g., Coxhead & Boutorwick, 2018; Greene & Coxhead, 2015).

3.5 Conclusion

The findings from this study appear to support previous studies (Burgoyne et al., 2009; Melby-Lervåg & Lervåg, 2014; K. Nation & Snowling, 2004) that have shown the importance of vocabulary knowledge for EAL learners' reading comprehension. These findings also indicate that existing assessment tools can assess EAL learners' vocabulary knowledge. However, this study also raised two issues that still need to be considered: the degree to which EAL learners lag their FLE counterparts regarding vocabulary knowledge, and the large gap that exists between the vocabulary that these

learners know and the vocabulary that they require to be able to comprehend their textbooks. Chapter Four examines these two issues in more detail.

Chapter Four

Why Academic Texts May be Difficult for EAL Readers to Understand

4.1 Introduction

Chapter Three highlighted two important issues, that EAL learners have lower levels of vocabulary proficiency than their FLE peers, and that vocabulary is an important predictor of how well learners are able to understand the texts that they are reading. These two discoveries indicate that EAL learners are likely to struggle to read the texts they are required to read in the classroom. This insight sets the foundation for the experiment that I report on in the current chapter. To better understand the impact vocabulary may have on EAL learners' reading comprehension in the classroom, it is necessary to ascertain two crucial variables. First, we need to have a better understanding of what vocabulary EAL learners are likely to know at different grade levels. Second, we need to understand the vocabulary profiles of the academic texts these learners are likely to encounter in the classroom.

To address both questions, this chapter reports on a partial replication of a study by Coxhead and Boutorwick (2018). As I discussed in Chapter 2.2.4, Coxhead and Boutorwick examined the vocabulary profiles of EAL learners studying at an international school in Berlin, Germany, and compared those with the texts that these learners would be likely to encounter in the classroom. The EAL learners came from a diverse array of backgrounds: 43% of the 468 participants had German nationality, and the remaining participants came from over 50 different countries. Coxhead and Boutorwick's study is especially relevant to the present research because it similarly deals with examining the vocabulary knowledge of EAL learners in correlation with the texts that they use in the classroom. Their study used a Vocabulary Levels Test (VLT) to measure the type of vocabulary EAL learners were likely to know at different levels and then correlated that with the vocabulary found in a representative corpus of textbooks that the researchers had compiled for the study.

Replicating this study will facilitate a better understanding of the gaps in vocabulary knowledge that EAL learners may exhibit in the English Medium of Instruction (EMI) context. In addition, replicating Coxhead and Boutorwick's study in the Japanese context will enable me to ascertain the degree to which the vocabulary

knowledge of learners in the Japanese international school context differs from the vocabulary knowledge of EAL learners studying in the German international school context. A better understanding of the similarities and differences between these two groups of learners will facilitate a better understanding of how frequency lists developed for Japanese international school students can be applicable to students at international schools in different countries. If there are overlaps in the vocabulary knowledge and requirements between the two groups, it could significantly expand the audience for the present research.

4.1.1 Coxhead and Boutorwick's findings

Coxhead and Boutorwick (2018) found for all grade levels that EAL learners received significantly lower scores on the VLT compared to both First Language English (FLE) learners and highly proficient L2 learners. One of the key findings was that most EAL learners could not achieve mastery of the 2,000-word level until the start of Grade 9, and then took until Grade 11 to master the 3,000-word level and Academic Word List (AWL) (see Table 4.1). Based on these findings, Coxhead and Boutorwick concluded that most of the EAL learners in their study never achieved mastery of the vocabulary necessary to understand the textbooks that they were reading in the classroom.

Table 4.1

Average VLT Scores with Standard Deviations for NNSEALs

Grade	VLT 2,000	VLT 3,000	VLT 5,000	VLT 10,000	VLT AWL
6	16.0 (7.2)	10.2 (6.2)	6.8 (5.2)	2.1 (2.5)	7.3 (5.7)
7	23.3 (5.1)	17.0 (6.3)	12.5 (5.2)	4.0 (3.4)	12.8 (7.0)
8	25.2 (4.5)	21.4 (6.2)	15.6 (5.0)	6.2 (4.7)	18.0 (7.0)
9	27.4 (2.9)	22.9 (5.7)	19.1 (6.2)	8.7 (5.6)	22.3 (5.3)
10	28.8 (2.2)	25.7 (4.5)	22.9 (3.8)	10.2 (5.2)	25.8 (2.5)

Note. VLT = Vocabulary Levels Test; AWL = Academic Word List. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 598. (https://doi.org/10.1002/tesq.450)

Coxhead and Boutorwick were able to highlight potential gaps in EAL learners' vocabulary knowledge by comparing the scores participants received on a receptive vocabulary test (Schmitt et al., 2001 VLT) with a frequency analysis of a corpus of representative texts using the Range program (Heatley et al., 2004). Accordingly, they used the first 25,000 word families from Nation's (2020) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) lists of word families, which they supplemented with proper nouns, abbreviations, and marginal words. Coxhead and Boutorwick found that the first 8,000 word families provided over 98% coverage of fictional texts, an expected outcome based on the results of previous studies (Coxhead, 2012; Hirsh & Nation, 1992; P. Nation, 2006). The same 8,000 word families provided 92.95% coverage of a Grade 8 Science text and 94.62% coverage of a Grade 11 Maths textbook. Given the expected vocabulary knowledge of EAL learners at these grade levels, the vocabulary knowledge of the learners in Coxhead and Boutorwick's study would fall below the lexical threshold needed to understand these texts (see Table 4.2).

Table 4.2

Coverage of the BNC/COCA High, Mid, and Supplementary Lists Over English, Maths, and Science Textbooks (%)

Frequency Bands	English Grade 6	English Grade 10	Maths Grade 8	Maths Grade 11	Science Grade 8	Science Grade 11
High BNC/COCA 1,000-3,000	92.62	92.28	85.32	73.19	85.75	83.79
Mid BNC/COCA 4,000-8,000	4.39	3.22	5.45	4.58	3.60	7.28
Supplementary Lists	1.08	3.42	7.20	16.85	3.60	4.97
Total	98.09	98.92	97.97	94.62	92.95	96.04

Note. Adapted from "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science," by A. Coxhead and T.J. Boutorwick, 2018, *TESOL Quarterly*, *53*(3), p. 601. (https://doi.org/10.1002/tesq.450)

The results illustrate the difficulties that EAL learners would likely encounter when trying to read these texts.

4.2 The partial replication of Coxhead and Boutorwick (2018)

The replication shares many of the same goals as addressed in Coxhead and Boutorwick's original study. Namely, it investigates the following questions.

1. How does vocabulary knowledge for EAL and FLE learners differ at different grade levels in an EMI context?
2. What are the vocabulary profiles of the textbooks learners are expected to use for the different IB subjects?
3. What coverage does the AWL list provide for representative textbooks in the subjects learners are likely to study in the international school context?

The replication also includes one additional question. I have added this fourth research question for two reasons. First, the vocabulary profiles of the EAL learners in

my study were not uniform; it was paradoxically possible for EAL learners to show mastery of frequency bands that were lower in frequency than some bands that they could not master. For example, it was common for EAL learners who did not demonstrate mastery of the 3,000-word band to display mastery of the 4,000- or 5,000-word bands. Using the coverage of all the bands the learners had shown mastery of in my analysis will allow me to compile a more complete picture of the probable number of unknown words in the texts. Second, calculating the percentage of known words in each frequency band will allow me to better determine the number of words EAL learners would be required to learn, using existing word lists, to understand the texts. This more comprehensive understanding of the gaps that exist in EAL learners' vocabulary knowledge will build the foundations for the subsequent chapters of this dissertation. My additional research question is:

4. What percentage of the words in the textbooks would the EAL learners be likely to understand?

4.2.1 Differences between the replication and the original study

The replication study differs from the original study in several important ways. First, the replication does not track learner data over time, but investigates the vocabulary knowledge of individuals at different grade levels. The replication provides a much more robust analysis of the type of texts learners are likely to encounter in the classroom at specific grades than Coxhead and Boutorwick's study did. They used a relatively small corpus of only 461,554 running words which covered three subjects. While this allowed the authors to highlight the potential problems that EAL learners would likely encounter when reading these textbooks, my use of a larger corpus, compiled using a greater variety of textbooks, will facilitate a much more detailed overview of the vocabulary that this group of learners needs to know.

It is important to better understand the vocabulary in these textbooks because they play such an important part in the EAL classroom. Textbooks can be challenging for EAL learners as they are written in a distinctly academic style. According to Leung (2014), the language that is used in textbooks "tends to be highly structured in organization, specialist in register and new/unfamiliar in meaning" (p. 137). Better

understanding the vocabulary that is used in these textbooks will enable me to identify potential issues that EAL learners may encounter when reading them.

4.3 Methodology

4.3.1 Participants

143 participants (N=143) from two international schools in Japan took the initial VLT tests and answered several survey questions about their linguistic backgrounds. I removed four participants from the data set based on a Rasch analysis of their vocabulary levels test scores, leaving 139 participants (68 males and 71 females). The learners who participated in this study were a mix of nationalities and language backgrounds. The most common nationality was Japanese, with 99 out of the 139 learners identifying as either Japanese or as bilingual Japanese L1 speakers (those learners who reported speaking both Japanese and another language at home). The participants reported speaking 22 different languages, the most common of which were Japanese and English.

Based on their nationality, time spent in English-speaking countries, the languages spoken at home, and their classroom teacher's assessment of their English language proficiency, participants were grouped as FLE, Proficient L2 (PL2), or EAL learners. Using the same methodology as Coxhead and Boutorwick's (2018) study, I identified the English as an Additional Language (EAL) group as those participants who reported not speaking English at home and who also required additional language support in the classroom. I subdivided the remaining participants into two groups. Participants who were identified by their classroom teacher as being proficient in English, and who reported speaking predominantly English at home, or who reported spending a significant amount of time living and studying in a country where English is the primary language of communication, were classified as FLE speakers. Participants who were identified as proficient in English, but who reported primarily speaking a language other than English at home, and had not spent more than a year in a country where English is spoken as a first language were identified as proficient L2 (PL2) learners. The term "PL2" learners was chosen over "NNS," which was used in the original study, to avoid using the potentially controversial deficit model of language learning associated with the terms "native" and "non-native" speaker for this chapter

(see Cook, 1999, for a more detailed explanation of this issue). Table 4.3 shows the number of FLE, PL2, and EAL participants at each grade level.

Table 4.3

Number of FLE, NNS, and EAL Participants by Grade Level

Grade	6th	7th	8th	9th	10th	11th	12th	Total
FLE	0	3	2	1	1	3	1	11
PL2	2	2	4	6	2	6	2	24
EAL	4	13	15	17	21	28	6	104
Total	6	18	21	24	24	37	9	139

As with the study discussed in Chapter Three, collecting informed consent from the participants was a detailed and collaborative process between the stakeholders involved, including myself, the principals of the two schools, the parents, and the students. I have outlined the process for obtaining consent from these groups in more detail in Chapter Three and I include a copy of the consent forms I used in Appendix A. The students' parents were required to sign the consent forms for them to be included in this study. I also obtained consent from both the classroom teachers and the principal at the schools where the study occurred and followed the ethic's guidelines from the university where I was employed. To ensure that participants could withdraw from the study anonymously, I included another consent checkbox in the online survey. I removed the data of any students who did not return the form or who checked the No box on either the consent form or the survey from the analysis. This research has been approved by the Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University (approval number: HR-HUM-000762).

4.3.2 Survey Instruments and Procedure

Two different VLT tests were given to the participants. I gave the first group McLean and Kramer's (2016) New Vocabulary Level Test (described in detail in Chapter Three). This is a multiple-choice test with 24 questions for each of the first 5,000-word

frequency bands with an additional section that covers words from the AWL (Coxhead, 2000). I chose this test over previous versions of the VLT (e.g., Nation, 1990; Schmitt et al. 2001) for two reasons. First, it covers more frequency bands; previous assessments only measure the second, third, fifth, and 10th frequency bands. Second, earlier VLT selected target words and distracters from outdated frequency lists. I conducted this assessment online in January 2018 after receiving informed consent from the participants' parents or guardians. At the same time as taking the nVLT, the participants also filled out a survey about their language backgrounds.

I gave the second group of participants Webb et al.'s (2017) updated Vocabulary Levels Test (uVLT). The uVLT also measures the participants' knowledge of the first five 1,000 frequency bands. As with the nVLT, the items for the uVLT were selected using Nation's (2020) British National Corpus/Corpus of Contemporary American English. However, unlike the nVLT, which uses a multiple-choice format, the uVLT uses a matching format with 10 3-item clusters per level (see Figure 4.1). The items on the test are more representative of the items in the actual frequency bands than those on the nVLT, as the authors of the uVLT included a representative proportion of nouns, verbs, and adjectives (15, 9, and 6 items per level, respectively). As the uVLT does not include questions on Coxhead's (2000) AWL, the AWL questions from the nVLT were included at the end of the assessment. As in Coxhead and Boutorwick's (2018) study, a score of 86% or higher was necessary for mastery of each level. I conducted all assessments online in May 2019 after receiving informed consent from the participants' parents or guardians. As with the cohort who completed the nVLT, the participants also completed a survey about their language backgrounds.

Figure 4.1

An Example of a Question on the uVLT With the Answers Given

	eye	father	year	van	voice	night
body part that sees	✓					
parent who is a man		✓				
part of the day with no sun						✓

Note. Adapted from "The Updated Vocabulary Levels Test" by S. Webb, Y. Sasao, and O. Balance, 2017, *International Journal of Applied Linguistics*, *168*(1), p. 61.

The reason for the difference in VLTs used between the two groups is that the uVLT was not available when the assessments were given to the first group of participants. However, I decided that the construct validity of this test was greater than that of the nVLT due to how items were selected, and the more rigorous validation process used. Initially, this assessment (i.e., the uVLT) was to be given to the participants over an extended period to measure their vocabulary growth. However, because of complications arising from the coronavirus pandemic and subsequent lockdowns, it was not possible to collect data on the participants in 2021 or 2022. As a result, this chapter will look at the two tests in conjunction to measure the participants' vocabulary knowledge.

4.3.3 Textbooks

To represent the textbooks learners would use in the classroom, a sample of books that had been compiled and cleaned for the corpus to be used to create the EAL word lists (Table 4.4). The sample included texts from six different subjects that were being taught at the two international schools. These subjects were English, Biology, Chemistry, Physics, Maths, and Theory of Knowledge (TOK). This corpus of textbooks was scanned into the computer and carefully checked. Following Nation (2016), I cleaned any errors that I identified in the text using BBEdit, a text editor. I then used Excel and R to identify any errors that may have been missed during the initial cleaning process. I discuss the process in more detail in Chapter 5.3. The texts for this corpus were selected because they represent the vocabulary learners would aspire to learn during their time in

the IB program. It was not practical to extend this project to include the analysis of textbooks being used at the middle school grade levels because the teachers at this level reported making a much greater use of handouts and teacher-created materials, as opposed to published textbooks, in the classroom.

Table 4.4

Number of Textbooks and Running Words by Subject

Subject	Number of textbooks	Total running words
Literature	7	1,116,532
Maths	8	1,138,391
Physics	5	1,213,839
Biology	5	966,820
Chemistry	5	1,394,017
Total	30	5,457,786

4.4 Data Analysis

4.4.1 Validating the VLTs

For both the nVLT and uVLT, validity was assessed using a Rasch analysis (see, Beglar, 2010). Overall, most of the nVLT scores displayed a good fit to the Rasch model. However, four learners were shown to have a high Outfit Score (Zstd > 8.76), so I removed their results from the analysis. The residuals indicated no other problems with the fit of the learners' scores.

Following the same method as Coxhead and Boutorwick (2018), I considered participants to have mastered a frequency band if they could score above 86% on the vocabulary test on the target words from that frequency band.

4.4.2 Analyzing the Vocabulary Profiles of the Textbooks

I carried out the frequency analysis using the R programming language (R Core Team, 2022). For the current chapter, I made use of a number of R libraries including the

Tidyverse package (Wickham et al., 2019) and Tidytext (Silge & Robinson, 2016) libraries to help me analyze the text files. The text files for the textbooks were first imported into R and changed into a data frame with one row for each sentence. I cleaned the data frame to remove all the text that came from tables, figures, headers, footers, and the front and back pages of the textbook. I kept the text from the descriptions of the figures and tables, text boxes, and the questions and activities, along with the main body of the textbooks. SpaCy (Honnibal et al., 2020) was then used to identify proper nouns in the text. After that I created a data frame using the first 25,000 word families from Nation's (2020) British National Corpus (BNC)/Corpus of Contemporary American English lists to identify the frequency of the word families in the text. I also created a second data frame using Nation's (2020) lists of proper nouns (e.g., Asia, John), abbreviations (e.g., mm, am, ppm), compounds (e.g., overtime, signpost), and marginal words (e.g., ha, wow, A, Z) to identify these words in the text. I then used a third data frame created using the 2,000 word families from the GSL (West, 1953) combined with the AWL (Coxhead, 2000) to determine the frequencies of the words from these lists in the textbooks. Next, I examined the residuals (which comprised all the off-list words from the textbooks) to make sure that there were no proper nouns, marginal words, or abbreviations missing from the supplementary list, and I added any words from those categories to the appropriate supplementary list. Finally, I ran the analysis again with the new supplementary lists and checked any off-list words against the original PDF to ensure there were no problems with how the text files were cleaned.

4.5 Results

Research Question 1: *What level of vocabulary mastery was exhibited by the EAL learners at different grade levels?*

An overview of the nVLT scores (see Table 4.5) shows that a high number of learners could not achieve mastery of the mid-frequency word bands or the AWL. This finding mirrors the results found by Coxhead and Boutorwick (2018). I also found that, prior to Grade 9, a high number of the participants still could not master the high-frequency words in the 2,000 word band, and under 50% of the two cohorts could master the 3,000-word bands prior to Grade 12. Learners from all grade levels struggled

to master the AWL. Given the importance of these high-frequency words and the AWL for the understanding of school textbooks (Greene & Coxhead, 2015; Nagy & Anderson, 1984), a lack of mastery of these vocabulary items would mean that these learners would struggle to read at their grade level.

Table 4.5

Mastery of Vocabulary by Grade Level

	1000	2000	3000	4000	5000	AWL
Grade 6	100.0%	33.3%	16.7%	16.7%	16.7%	16.7%
Grade 7	88.9%	50.0%	5.6%	27.8%	27.8%	5.6%
Grade 8	95.2%	66.7%	33.3%	42.9%	38.1%	19.0%
Grade 9	95.8%	83.3%	50.0%	45.8%	45.8%	20.8%
Grade 10	91.3%	82.6%	26.1%	52.2%	47.8%	13.0%
Grade 11	94.7%	84.2%	36.8%	47.4%	39.5%	26.3%
Grade 12	100.0%	88.9%	55.6%	66.7%	44.4%	55.6%
Average	94.2%	74.8%	33.1%	44.6%	39.6%	20.9%

Note. AWL = Academic Word List

I then looked at the EAL learners and PL2/FLE learners separately. This analysis showed that EAL participants were much more likely to struggle to master high-frequency vocabulary (Table 4.6) than PL2/FLE learners (Table 4.7). Most of the PL2/FLE participants were able to achieve mastery of the 5,000 most frequent word

families. This is something that most of the EAL learners could not do, and I found that, even at the Grade 12 level, fewer than 50% of the EAL participants could master the frequency lists above the first 2,000 most frequent word families. One potential problem with PL2/FLE learners, though, was that there were still many PL2/FLE learners who could not show mastery of the AWL, indicating that vocabulary knowledge, especially academic vocabulary knowledge, may still be an issue even with this more proficient group of learners.

Table 4.6

EAL Mastery of Vocabulary by Grade Level

	1000	2000	3000	4000	5000	AWL
Grade 6	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Grade 7	84.6%	30.8%	0.0%	7.7%	0.0%	0.0%
Grade 8	93.3%	53.3%	20.0%	26.7%	20.0%	0.0%
Grade 9	94.1%	76.5%	29.4%	29.4%	23.5%	0.0%
Grade 10	84.2%	73.7%	15.8%	36.8%	31.6%	5.3%
Grade 11	96.4%	82.1%	17.9%	35.7%	25.0%	14.3%
Grade 12	100.0%	83.3%	50.0%	50.0%	33.3%	50.0%
Average	92.2%	65.7%	18.6%	29.4%	21.6%	7.8%

Note. AWL = Academic Word List

Table 4.7

FLE/PL2 Mastery of Vocabulary by Grade Level

	1000	2000	3000	4000	5000	AWL
Grade 6	100.0%	100.0%	50.0%	50.0%	50.0%	50.0%
Grade 7	100.0%	100.0%	20.0%	80.0%	100.0%	20.0%
Grade 8	100.0%	100.0%	66.7%	83.3%	83.3%	66.7%
Grade 9	100.0%	100.0%	100.0%	85.7%	100.0%	71.4%
Grade 10	100.0%	100.0%	60.0%	100.0%	100.0%	40.0%
Grade 11	100.0%	100.0%	100.0%	88.9%	88.9%	66.7%
Grade 12	100.0%	100.0%	66.7%	100.0%	66.7%	66.7%
Average	100.0%	100.0%	73.0%	86.5%	89.2%	56.8%

Note. AWL = Academic Word List

Research Question 2: W*hat are the vocabulary profiles of the textbooks learners are expected to use for the different IB subjects?*

I then calculated the vocabulary profiles of the textbooks using the BNC/COCA. Table 4.8 shows the results of this analysis. The textbooks that would be the easiest for the EAL learners to read were the literature ones, as it was possible to achieve the 95% needed for comprehension with the first 5,000 most frequent words. The math textbooks also came close to 95% coverage with the first 5,000-word families. However, the coverage of the first 5,000 words was significantly lower than 95% for the textbooks in all the other subjects.

Table 4.8

BNC/COCA & AWL Coverage Per Discipline (IS-CAT)

Frequency Band	Literature	Math	Physics
2,000	86.84%	83.79%	84.03%
5,000	96.55%	94.28%	91.48%
AWL	6.87%	8.30%	7.36%
Frequency Band	Biology	Chemistry	Average
2,000	77.08%	77.79%	81.91%
5,000	91.45%	92.60%	93.27%
AWL	7.70%	8.60%	7.77%

Note. AWL = Academic Word List

Given that fewer than 35% of the Grade 10, 11, and 12 EAL learners in the current study could master the first 5,000 most frequent words, I would expect them to struggle with these textbooks. I found the biology textbooks to be the most difficult and the physics and chemistry textbooks to be the next most difficult, with the 5,000-word bands providing 91.45%, 91.48% and 92.6% coverage of the texts, respectively. The coverage of the high-frequency first 2,000-word families of these texts was low. These lists cover fewer than 78% of the running words in the text for both biology and chemistry.

Research Question 3: *What coverage does the AWL provide for representative textbooks in the subjects learners would be likely to study in the international school context?*

I found that the chemistry and math textbooks contained the highest proportion of AWL words, with 8.6% coverage for chemistry and 8.3% coverage for math. Biology and physics were next, with 7.7% for biology and 7.36% for physics. The most

surprising finding was the amount of coverage the AWL gave for the literature textbook corpus. Coxhead and Boutorwick's (2018) study found that the AWL only provided 1.5% coverage of the representative novels they investigated in their study, which was similar to the findings of previous studies (Coxhead, 2012). However, when I looked at the literature textbooks learners were likely to use in the IB classroom, the AWL provided 6.87% coverage. The difference between these two results is probably because these books included not only works of literature, such as poems or short stories, but also explanations of how to analyze and write about literature. It is likely that the descriptions of how to analyze literature involve the use of more words from the AWL than a novel would normally include.

Research Question 4: What percentage of the words in the textbooks would the EAL learners be likely to understand?

I used both the learners' mastery levels on the VLTs and the percentage of words from each of the first five frequency bands (Table 4.9) to determine the percentage of the vocabulary in the textbooks that EAL learners at different grade levels would be likely to understand. Learners who had mastered a specific frequency band were determined to have knowledge of the words at that frequency level. I tallied the coverage given by the different levels to determine the overall coverage a learner would have of the vocabulary in the different subjects.

Table 4.9

BNC/COCA Coverage Per Discipline (IS-CAT)

	English	Biology	Chemistry	Physics	Math
K01	74.54%	65.51%	63.39%	70.88%	72.38%
K02	9.56%	10.89%	12.94%	12.37%	9.96%
K03	7.35%	9.23%	9.26%	7.58%	8.17%
K04	1.52%	3.03%	3.78%	3.28%	2.71%
K05	0.84%	2.15%	1.76%	1.28%	1.61%
K6 to 25	2.33%	7.08%	5.96%	3.30%	3.05%
Other	3.86%	2.12%	2.90%	1.30%	2.12%

Figure 4.2 illustrates the different levels of coverage learners would be likely to have of the different corpora. It demonstrates that most of the learners lack the coverage necessary to read these discipline-specific textbooks. For most of the learners, biology and chemistry would be the hardest subjects to understand, whereas math and literature would be the easiest. Looking at just the learners at the IB Diploma level (Figure 4.3), we can see that many participants still lack the knowledge of the vocabulary that they would need to read the textbooks that they are using in the classroom.

Figure 4.2

A Histogram of the Percentage of Words Participants Would be Likely to Understand in Each of the Subject-specific Corpora (All EAL Learners)

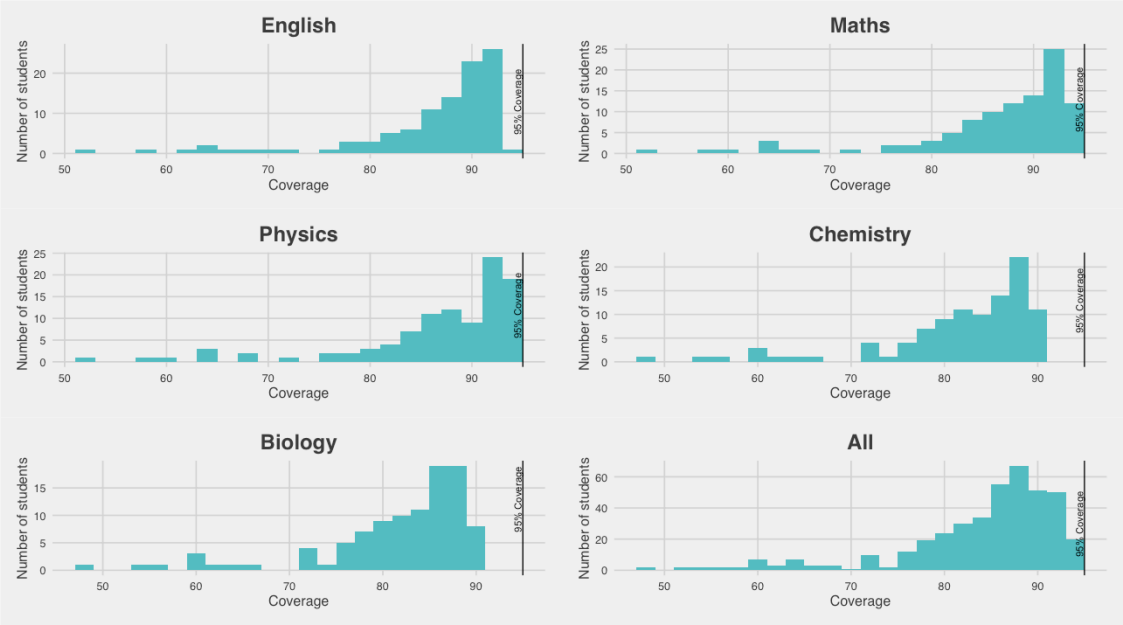
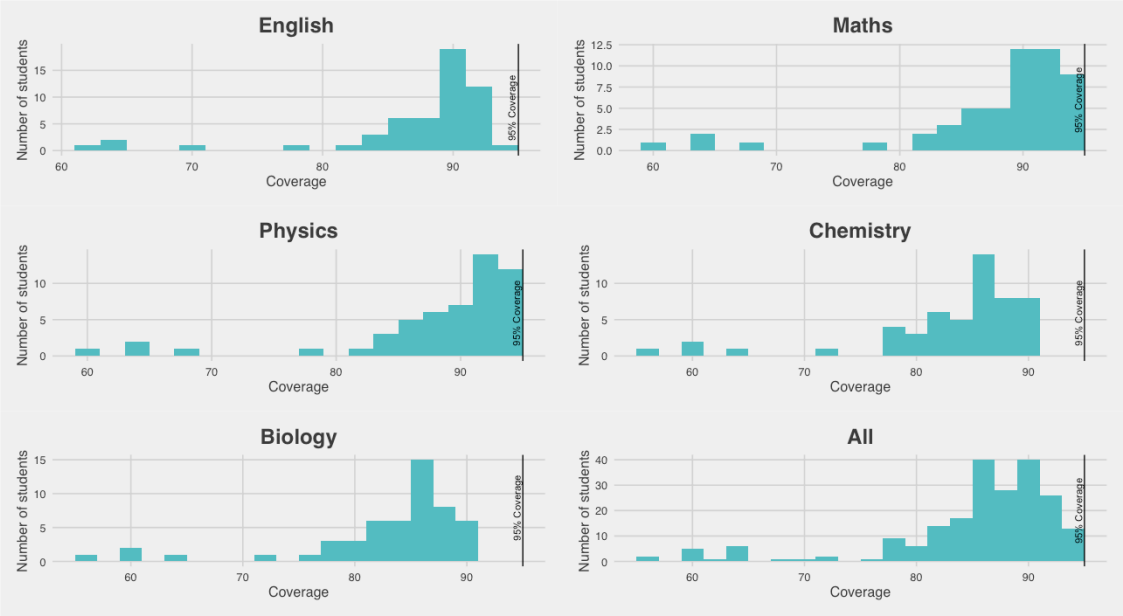


Figure 4.3

A Histogram of the Percentage of Words Participants Would be Likely to Understand in Each of the Subject-specific Corpora (Grade 10, 11, and 12 EAL Learners)



4.6 Limitations with the research

While the current study supports Coxhead and Boutorwick's (2018) finding that most EAL learners studying at an international school would be likely to struggle to read the textbooks that they are required to use in the classroom, it also shows problems with using existing word lists in the EAL context. Two significant problems emerged with using the BNC/COCA in the international school context: the limitations of using existing word lists and the issue of using word families as a unit of counting vocabulary.

4.6.1 Limitations with using existing word lists

The EAL learner vocabulary mastery profiles were not well aligned with these vocabulary lists. Previous studies analyzing the vocabulary profiles of EFL learners have found that learners master the lower and mid-frequency lists more or less in order (P. Nation, 2016), because the number of times that they are likely to encounter a word in the texts that they are usually exposed to decreases as the frequency band increases. Given that learners are more likely to acquire the vocabulary that they encounter most frequently, we would expect these learners to acquire high-frequency vocabulary first. However, this was not the pattern that I observed with the EAL participants in this study. The biggest discrepancy was the 3,000-word band. Although only 18.6% of EAL learners could master the 3,000-word band, a significantly higher number of learners could master the 4,000-word band (29.4%). This paradoxical pattern becomes even more obvious when we look more closely at individual participants (see Table 4.10).

Table 4.10

Individual Participants' Mastery of the Word Frequency Bands

	1K	2K	3K	4K	5K	AWL
S04	**100%**	**97%**	80%	**87%**	**87%**	67%
S09	**100%**	**100%**	67%	77%	**90%**	63%
S12	**100%**	**97%**	70%	**87%**	83%	77%
S21	**93%**	**90%**	53%	23%	43%	30%
S28	**100%**	**100%**	70%	**93%**	**87%**	73%
S32	**100%**	**97%**	67%	**100%**	73%	67%
S51	**100%**	**100%**	**87%**	83%	**87%**	77%

Note. Bold text shows where the participant has shown mastery of that word frequency band.

For example, we can see that participant S04 shows mastery of both the 4,000- and 5,000-word bands, but not the 3,000-word band. A similar situation exists for participant S09, who displays mastery of the 5,000-word band but not the 3,000- or 4,000-word bands. Given that learners would be expected to learn high-frequency vocabulary before mid- and low-frequency vocabulary, this tells us that the frequency bands of the BNC/COCA may not accurately represent the actual frequency of the vocabulary that EAL learners are likely to encounter in their textbooks.

When considering the strong and significant correlation between vocabulary knowledge and reading comprehension outlined above, these lower levels of vocabulary knowledge among international school EAL learners underline the primary importance of vocabulary intervention for this group of learners. Because the international school EAL learners appear to exhibit such different linguistic profiles compared to other L2 learners of English (e.g., McLean & Kramer, 2016; Webb et al., 2017), additional studies are necessary to determine EAL learner vocabulary knowledge and needs in an EMI context. The creation of an EAL academic word list described in Chapters Five and Six represents an important step towards practically addressing this crucial issue.

4.6.2 Limitations of Using Word Families

Using word families may not be the most appropriate unit when identifying and counting words in a text. As with Coxhead and Boutorwick's (2018) study, the study described in this chapter makes use of word lists that are based on word families. However, there are several acknowledged issues with using word families in the construction of word lists. First, unique items within the same word family may not have the same meaning (Gardner & Davies, 2014; Nagy & Townsend, 2012). This is especially true for words that are being used in a specific academic context. For example, Gardner and Davies (2014) give the example of the word family with *react* as the headword. The headword itself means to respond. However, this differs from some of the other words in the same word family. *Reactionary*, which occurs frequently in the domains of history and politics, means strongly opposed to social or political change; *reaction* and *reactor*, which occur frequently in the domains of physics and chemistry, mean a chemical process and a device or apparatus, respectively. This problem is compounded further because word families do not consider the grammatical parts of speech either (e.g., nouns, verbs, adjectives, adverbs).

In this study, another major concern with counting word families instead of lemmas is that many lower-proficiency EAL learners do not have complete knowledge of derivational word relationships, which has been shown to come much later than knowledge of inflectional word relationships (Gardner, 2007). The ability to engage in morphological analysis of words depends on learners' existing vocabulary (Nagy, 2007), which would make using lists based on word families difficult for EAL learners. Because lower-proficiency learners could not be expected to be able to correctly identify different members of the word family, each word in the family would have to be taught separately to ensure that the learners can understand them. This would make these lists difficult to use in the EAL context, as the number of words that learners would have to study would be much greater than just the number of headwords. Fortunately, recent studies (Gardner & Davies, 2014; Green & Lambert, 2019) have shown that word lists based on lemmas can provide similar coverage of a corpus with many fewer items.

4.7 Conclusion

The current replication study makes manifest both the importance of vocabulary for EAL learners and the fact that the vocabulary knowledge of EAL learners in the international school context in Japan may actually be lower than that shown in Coxhead and Boutorwick's (2018) study. An analysis of the vocabulary coverage that the participants' existing vocabulary knowledge would be likely to give them shows that existing word lists may not be useful for teachers and learners in choosing what vocabulary to focus on in the classroom. The findings help to clarify some of the specific issues that exist when teaching vocabulary to this group of learners.

I conclude the current chapter with a summary of the most important findings from the replication study, followed by a discussion of their implications for the overarching research goals of the thesis.

4.7.1 Summary

Using two recent VLTs, I obtained similar results to Coxhead and Boutorwick (2018) for the number of words EAL learners are likely to know compared to their PL2/FLE learner counterparts. Likewise, my findings for the coverage of a corpus of representative textbooks corresponded closely with Coxhead and Boutorwick's findings in their study. However, the EAL learners in my context could master fewer frequency bands, and most EAL learners still had not mastered up to the 5,000-word band by the time they entered Grade 12. This may reflect a difference between the learners at these two international school contexts, Japan and Germany, such as the differences in the learners' respective L1 backgrounds.

Importantly, I could corroborate Coxhead and Boutorwick's finding of the importance of providing EAL learners with sufficient support for their vocabulary acquisition. In both my own and Coxhead and Boutorwick's study, EAL learners could not demonstrate mastery of the word families required for them to comfortably read the textbooks assigned for their classes without support. I also found that the number of words that EAL learners would be required to learn to know enough vocabulary to read these texts would be difficult to achieve using existing word lists.

4.7.2 Implications of this research

An important aim of this dissertation is to provide support for EAL learners studying in an international school context. The findings of the replication represent a first step towards this important goal. We now know more about the vocabulary that EAL learners are likely to know, and we also have a better understanding of the vocabulary that they would need to understand the textbooks that they are likely to use in the classroom. From this analysis, we have seen that most EAL learners do not have the vocabulary knowledge necessary to understand these textbooks. To further complicate matters, we have also discovered that existing word lists, such as the BNC/COCA, are unlikely to prove useful for this group of learners as the amount of vocabulary that they would need to learn to reach the coverage necessary to understand these textbooks is greater than they could realistically acquire in such a short time.

To bridge this gap, it is necessary to compile several domain-specific word lists that will significantly reduce the number of words necessary to achieve coverage of the textbooks that EAL learners are required to read in the classroom. Although there are several middle and secondary school word lists available that provide greater coverage than the mid-frequency bands of the BNC/COCA or the AWL for certain textbooks, the amount of coverage these word lists will provide for our IB corpus is unclear. These are the questions that I address in Chapters Five and Six of this dissertation.

Chapter Five

Creating an International School Word List

5.1 Introduction

The replication of Coxhead and Boutorwick's (2018) study I reported in Chapter Four provides two significant implications for this dissertation. First, it shows that gaps exist in the vocabulary profiles of learners studying in international schools in Japan and that these gaps are significant enough to cause these learners considerable problems when reading textbooks. Second, it demonstrates that the existing word frequency lists are inadequate for dealing with these gaps because they are not well aligned with the vocabulary in the textbooks.

Thus, in the current chapter, I detail an initial attempt to address these issues by building a corpus of international school textbooks and creating a more appropriate word list from this corpus. I begin the current chapter by presenting a brief review of the previous two chapters and the problems identified therein. I then outline the steps that I took to create a corpus of international school textbooks and create a word list to support EAL learners studying in this context. The methodology I used to create this initial academic word list is based on the techniques used by Greene and Coxhead (2015) and Coxhead (2000). This methodology was discussed in more detail in Chapter 2 sections 2.3.2 and 2.3.4. I end the current chapter by examining the effectiveness of this newly created word list with reference to existing academic word lists (Gardner & Davies, 2014; Green & Lambert, 2018; Greene & Coxhead, 2015) and discuss its strengths, weaknesses, and a prospective way forward.

5.1.1 The replication study: Findings and implications

The replication study in Chapter Four enabled me to make some important preliminary discoveries about the vocabulary used in international school textbooks and the implications for EAL learners; I summarize these as follows:

1. The frequency of the vocabulary used in these textbooks differs from that of the vocabulary used in the English texts that researchers have used to compile general word lists, such as the BNC/COCA.

2. The vocabulary profiles of EAL learners do not match the profiles of learners that have been studying English using general texts in that EAL learners often show better mastery of lower-frequency word bands.
3. EAL learners are likely to struggle with vocabulary outside the first 2,000 most frequent words, making it difficult for them to reach the coverage necessary for understanding by using existing word lists.

In Chapter Four, we saw that, on average, the first two 1,000 word bands of the BNC/COCA (P. Nation, 2020) only provide 81.91% coverage of the running words found in the IB textbooks international school students are expected to use in the classroom. Although the coverage is higher for some subjects, such as Literature, over which it provides 86.84% coverage, it is much lower for other subjects, such as Biology and Chemistry, over which it provides only 77.08% and 77.79% coverage, respectively. For all the IB subjects, except Literature, even if learners were to know the first 5,000 most frequent word families in the BNC/COCA (P. Nation, 2020), this would not be sufficient for the 95% coverage necessary for understanding. Despite its academic nature, the AWL only provides an additional 7.77% coverage over the IB textbook corpus, much lower than the 10% coverage Coxhead (2000) found that it provided for academic articles.

The importance of these results becomes more apparent when considering the insights gained from the pilot study reported on in Chapter Three. We can see from this study that vocabulary knowledge is a strong and significant predictor of reading comprehension for EAL learners (Brooks et al., 2021). In other words, EAL learners need to be familiar with a certain threshold percentage of the words in a text to be able to understand it. As I discussed in 2.2.1 and 2.2.3 of the literature review, Nation (2006) and Laufer and Ravenhorst-Kalovski (2010) found that knowledge of at least 95% of the words in a text is necessary for comprehension. Indeed, subsequent studies (e.g., Schmitt et al., 2011) have suggested that even greater amounts of coverage may be necessary for learners to read independently. Given these gaps that exist between EAL learners' existing vocabulary knowledge and the knowledge that they would need to be able to understand these textbooks, the existing general and academic word lists are insufficient.

The replication study reported in Chapter Four was helpful because it highlights the gaps that exist between the vocabulary needs of international school students and existing word lists. The replication study was also beneficial for another important reason: it allowed me to compile a corpus of academic textbooks used in the IB context. The corpus of textbooks used in Chapter Four forms the basis of this new International School Corpus of Academic Texts (IS-CAT) that I used to create the word lists discussed in this chapter. Although the corpus from Chapter Four does not include all the subjects covered in an IB education, it provided me with a starting point from which to develop the final IS-CAT. It also allowed me to test and refine the techniques necessary for compiling such a corpus using IB textbooks.

5.2 Methodology

5.2.1 Building a corpus of International School Textbooks

The first issue when building this word list was how to develop a corpus that would allow me to identify the most useful words in the international school context. Hunston (2002) discusses four criteria researchers must consider when compiling a corpus. These are size, content balance and representativeness, and permanence.

Thus, the first criterion I had to consider when developing the IS-CAT was the corpus size. Although, as a general rule, the larger the corpus, the better, compiling this corpus involves several trade-offs. The first factor to consider regarding the size of the corpus was the availability of materials. The second factor was time. Converting the textbooks into a form that the computer could read is necessary to compile the word lists but is a time-consuming process. It was necessary to compile a final corpus that balanced the amount of data necessary for analysis, the availability of data, and the time available to the researcher.

The next criterion I had to consider when creating the corpus was that of content. For it to be effective, we must compile a corpus from the types of texts learners in this context are likely to encounter (P. Nation & Sorell, 2016). As Greene and Coxhead (2015) note, one would not want to create a word list for high school students using automotive repair manuals. For the proposed International School corpus, I first had to identify what subjects IB learners are likely to study and what textbooks are used to teach those subjects. For this, I focused on the textbooks that students would be using

when they were studying for their IB Diploma. In the schools that I was working with, the students entered the diploma program in Grade 10 and completed the program during their Grade 12 year, therefore all of the textbooks would be ones that EAL learners would use during their secondary school education. The reason for this is that the IB Diploma is the focus of the IB program, and the middle school years are designed to prepare learners to successfully complete their diploma. I worked with the principals and teachers at the IB schools where I did the replication study to identify texts to include in the corpus. I used their advice, along with the four years of experience I have teaching at an IB school in South America, to identify eight different prominent IB subjects that could be used to compile the corpus.

According to the IB website, the IB Diploma program comprises three cores and six subject areas. The core areas are: Theory of Knowledge, the Extended Essay, and a Creativity, Activity, and Service (CAS) component, which focuses on volunteering. The six subject areas are: Studies in Language and Literature, Language Acquisition, Individuals and Societies, Science, Mathematics, and The Arts. Figure 5.1 provides a visual representation of the IB curriculum.

Figure 5.1

A Visual Representation of the Different Areas of The IB Curriculum

Note. From *Logos and Programme Models*, by the International Baccalaureate Organization, 2022, (https://www.ibo.org/digital-toolkit/logos-and-programme-models/). Reprinted with permission.

From the six subject areas, I identified eight diversely representative and prominent courses taught at the majority of international schools in Japan. These are Literature, Economics, Social Studies, Biology, Chemistry, Physics, Mathematics, and Theory of Knowledge. The Language Acquisition, Extended Essay, CAS, and Visual Arts classes were not included in the corpus as there were not enough textbooks available in the subject to make a big enough corpus, or the language used in the course was already similar to that covered by existing word lists. After I had identified the subjects that I would include in the corpus, I then had to identify what textbooks to use to create the corpus.

There are a limited number of textbooks for each subject taught in the IB classroom. To identify the most appropriate textbooks for the corpus, I worked with the teachers at the IB schools where I did the replication study. Working together, we could

identify several standard, popular textbooks that the teachers were using in their own classrooms. I then expanded the list of books to include in the corpus by identifying similar textbooks available through other publishers. For most subjects, there are between five and eight core textbooks written specifically for the IB classroom.

The next areas that researchers must consider when compiling a corpus are balance and representativeness. To ensure that the corpus was balanced and represented all the subjects identified for inclusion in the corpus, I first had to determine how many texts to include in each subcorpus. To do this, I first compiled the biology subcorpus to glean a better idea of how many tokens each textbook would contain and the process necessary for compiling the various corpora. I used the five textbooks that I had identified as textbooks commonly used in IB Biology classes. I then scanned and OCRed these books to produce a biology subcorpus of around 1 million running words. I then analyzed this corpus to ensure that it was large enough to identify the words used in these textbooks. Previous research (Brysbaert & New, 2009) has shown that, although a corpus of over 30 million tokens is necessary for identifying low-frequency words, a one-million token corpus is sufficiently large to identify the most common high-frequency words in a single subject area. To ensure that the corpus was balanced, I tried to keep the size of all subcorpora to around 1 million tokens. For most subjects, this meant scanning and processing the five most commonly used textbooks. However, some subjects, such as mathematics, required more texts to reach this number of tokens. Table 5.1 provides a list of the textbooks that I used to create a corpus for all the subjects in, along with the running words of each book (see Appendix C for the complete bibliographical information for each of the textbooks).

Table 5.1

A List of All the Textbooks in the IS-CAT Along with the Running Words for Each Book

Subject	Book	Tokens
Literature	*Hodder English Language for the IB Diploma*	210201
Literature	*Oxford English A: Literature*	88108
Literature	*Pearson English A Literature 2nd Edition*	142517
Literature	*Cambridge English A: Language and Literature*	137218
Literature	*Hodder English Language & Literature for the IB Diploma*	305946
Literature	*Oxford English A: Language and Literature 2nd edition*	170472
Literature	*Hodder English Language & Literature Skills for Success*	62070
Economics	*Cambridge Business Management for the IB Diploma Second Edition*	257909
Economics	*IBID Business Management Fourth Edition*	321112
Economics	*Oxford Business Management 2014 Edition*	165371
Economics	*Cambridge Economics for the IB Diploma*	329982
Economics	*IBID Economics in Terms of the Good, the Bad and the Economist 3rd Edition*	146668
Economics	*Oxford Economics Course Companion*	237767
Social Studies	*Oxford Global Politics Course Companion*	91024
Social Studies	*Pearson Global Politics*	63741
Social Studies	*Cambridge History for the IB Diploma: Causes, Practices and Effects of wars*	114037
Social Studies	*Cambridge History for the IB Diploma: Nationalist and Independence Movements*	116987
Social Studies	*Hodder History for the IB Diploma: Causes, practices and effects of wars*	145246
Social Studies	*Hodder History for the IB Diploma: Independence movements*	89176
Social Studies	*Oxford History for the IB Diploma: Causes and Effects of 20th-Century Wars*	115522
Biology	*Cambridge Biology for the IB Diploma 2nd Edition*	135972
Biology	*Hodder Biology 2nd Edition*	171471
Biology	*IBID International Baccalaureate Biology*	170676
Biology	*Oxford Biology Course Companion 2014 Edition*	197332
Biology	*Pearson Higher Level Biology 2nd Edition*	291369

Subject	Book	Tokens
Chemistry	*Cambridge Chemistry for IB Diploma 2nd Edition*	206284
Chemistry	*Hodder Chemistry for the IB Diploma 2nd Edition*	345811
Chemistry	*IBID Chemistry 3rd Edition*	261630
Chemistry	*Oxford Chemistry Course Companion 2014 Edition*	179985
Chemistry	*Pearson Higher Level Chemistry 2nd Edition*	400307
Physics	*Cambridge Physics for IB Diploma 6th Edition*	189092
Physics	*Hodder Physics for the IB Diploma 2nd Edition*	250016
Physics	*IBID Physics 3rd Edition*	273773
Physics	*Oxford Physics Course Companion 2014 Edition*	283715
Physics	*Pearson Higher Level Physics 2nd Edition*	217243
Mathematics	*Cambridge Mathematics Higher Level*	185504
Mathematics	*Haese Mathemalics Analysis and approaches HL 2*	194100
Mathematics	*Haese Mathematics Core topics HL 1*	41142
Mathematics	*Hodder Mathematics Analysis and Approaches HL*	117763
Mathematics	*IBID Mathematics Common Core*	96053
Mathematics	*Oxford Mathematics: Analysis and Approaches Higher Level*	118225
Mathematics	*Oxford Mathematics for the IB Diploma*	144461
Mathematics	*Pearson Mathematics: Analysis and Approaches Higher Level*	241143
TOK	*Cambridge Theory of Knowledge 2nd Edition 2021*	86258
TOK	*Cambridge Theory of Knowledge 3rd Edition 2020*	85593
TOK	*Hodder Theory of Knowledge Skills for Success 2nd Edition*	60305
TOK	*Hodder Theory of Knowledge 4th Edition*	251510
TOK	*Oxford Theory of Knowledge 2020 Edition*	206415
TOK	*Pearson Theory of Knowledge Essentials*	97651
TOK	*Pearson Theory of Knowledge for the IB Diploma 3rd Edition*	164968

Note. TOK = Theory of Knowledge

The final area for consideration is permanence. There are two factors to consider when assessing permanence: the purpose of the corpus and how the texts that constitute

the corpus change over time. However, as the purpose of the corpus is to identify the words that EAL learners need to know and not to examine how the language in these books changes over time, I selected textbooks that were published at more or less at the same time. Because the topics taught in the IB Diploma also change over time, I selected the most recently available textbooks for each of the subject areas. An example of why that is important becomes clear if one looks at how the IB Biology course has changed over the last 15 years. Recent iterations of the course have included much more of a focus on topics such as the environment and global warming, topics that were not covered in much detail when I was teaching in the IB program. By selecting the newest version of each of the textbooks, I could capture words that are current and specific to these topics in the corpus.

The final International School Corpus of Academic Texts (IS-CAT) comprises just under 9 million running words. It is made up of eight subject-specific subcorpora ranging in size from just over 800,000 words to just over 1,400,000 words. The total number of tokens for the corpus and each of the subcorpora is given in Table 5.2.

Table 5.2

The Total Number of Running Words for the IS-CAT and Each of the Subcorpora

Subject	Total Words
Literature	1,116,532
Social Studies	735,733
Economics	1,458,809
Biology	966,820
Chemistry	1,394,017
Physics	1,213,839
Mathematics	1,138,391
Theory of Knowledge	952,700
Total	8,976,841

5.2.2 Creating and cleaning the corpus

Because the textbooks were only available in printed form, creating the corpus was quite labour intensive. First, I removed the bindings of the textbooks and scanned the pages into the computer as images. I then converted the scanned image files into text files using optical character recognition (OCR). The OCR process uses the computer to recognize images of printed words and convert them into text that a computer can understand. Because scanning and OCRing a document is not faultless, the resulting text files had to be cleaned manually before the computer processed them.

I scanned the textbooks using a Fujitsu Scansnap and OCRed using Abbyy Finereader 15 (https://pdf.abbyy.com). After I converted the PDFs to text, the text was exported as a plain text file. I then divided these files into chapters with the pages and paragraphs clearly marked using an R script written for this purpose (Appendix D contains the general R script used for this purpose. However, the script had to be revised to consider differences in layouts between textbooks).

A few issues I encountered when preparing the corpus are worth noting here. Secondary school textbooks use many text boxes (e.g., visual boxes highlighting key ideas, activities, or links to other subjects), figures, and tables. This means that they have less running linear text than academic texts, such as journal articles. As a result, there was more noise in the corpus than there would have been in a corpus compiled using web pages, PDF files, or typed materials such as student essays. Some of this noise involved missing punctuation and scanning errors. For example, alphanumeric characters that looked similar, such as the letter “I” and the number “1,” were occasionally mixed up, and sometimes spaces were not recognized, resulting in non-existent compound words. Thus, the files themselves were extensively cleaned before they were processed.

After the files had been cleaned by hand, the words in the file were checked against the 25,000 words of the BNC/COCA and against a more comprehensive list of English words (https://ftp.gnu.org/gnu/aspell/dict/en/). Any words not included in these two lists were checked manually, and any remaining errors were fixed. Despite this extensive data cleaning and preprocessing, it is impossible to eliminate all noise in such

a large corpus. However, the amount of noise left in the corpus only represents a small fraction of the total running words of the corpus. Previous studies have shown that the effects of this noise on the resulting word lists are minimal (Green & Lambert, 2018).

After the words in the corpus had been checked, there were several other factors that had to be considered when cleaning the corpus. I will discuss these next.

5.2.3 Hyphenated compounds and contractions

I expanded both contractions and hyphenated compounds into their constituent words. Hyphenated compounds were treated as two separate words. This was done because it made it possible to count both the hyphenated compounds and their nonhyphenated variants as the same word. This is important because hyphenated words tend to be semantically transparent, which means that if a learner understands the two words that make up the hyphenated word they are usually able to determine the meaning of the whole word. To do this, I globally replaced hyphens with spaces. I also removed any hyphens that were used as line-break hyphens and recombined the resulting word segments into their original word.

I also expanded contractions into their full constituent words. This meant that contractions such as *don't* or *aren't* were replaced with *do not* or *are not*. I used a list of contractions I created myself to expand these words into the non-contracted forms. I took the initial list from an online source (https://github.com/john-james-ai/NLPLists/blob/master/data-raw/contractions.csv), and I revised this list to remove any informal contractions.

5.2.4 Letter-digit combinations

Following the procedures set out by other researchers (e.g., P. Nation, 2016) I removed all numbers from the corpus. To ensure that I had completely removed biological and chemical formulas from the corpus, I also removed any combination of letters and numbers. Although perhaps it may have been appropriate to retain some number-letter combinations, such as numbers with cardinal endings like *-st*, *-th*, *-nd*, as they contain a semantic meaning, it was just not possible to do so efficiently. Given the prevalent use of such endings, their meaning is probably familiar to most, if not all, IB students by the time they enter the IB Diploma program. Although other researchers have tried to keep these number-letter combinations, they often found that it had minimal effect on the

resulting word lists and slowed the processing of the corpora (e.g. Sorell, 2013). The only exception to this process was regarding measurements, such as *km*, *ml*, or *kg*. Given the importance of these abbreviations for the maths and sciences, they were kept in the final corpus.

5.2.5 Proper nouns and marginal words

The final two categories of words that needed to be identified in the corpus were proper nouns and marginal words. I did this using a three-step process. First, I downloaded the lists of proper nouns and marginal words that are distributed as part of the Range Program (Heatley et al., 2004). I then used the natural language processing program spaCy (Honnibal et al., 2020) to identify any proper nouns in the text. I checked these newly identified proper nouns manually to ensure there were no mistakes and then added all the correctly identified proper nouns to the list of proper nouns created in the first step. Finally, during the cleaning phase described above, I added any off-list words that were proper nouns or marginal words to the appropriate list. I left both proper nouns and marginal words in the cleaned corpus, so they both count towards the final running word count for each subcorpus. However, I used the final CSV file of proper nouns and marginal words to identify and remove all the proper nouns and marginal words in the corpus while creating the final word lists.

5.2.6 Tagging and annotating the corpus

After I cleaned the text files, I added xml code to the text files to help identify the different parts of the text. I then converted the cleaned and tagged text files into a clean data frame using another R script (see Appendix D for an example of this code). Because of the differences between the layouts of different textbooks, it was necessary to adjust the resulting data frames manually. Using the conventions for creating a tidy corpus in R (Silge & Robinson, 2017), each row of the data frame included a single sentence from the text file, along with some metadata that would allow me to better process the text. The metadata included the textbook code, the chapter number, the page number, the paragraph number, and the text type. The text type metadata was used to identify the features of the sentence, such as if it was part of the main body of the textbook, a figure or table, part of a textbox, or part of an assignment, that could then be used to remove unnecessary text when creating the word list.

5.3 Data Analysis: Creating the General International School Word List

After constructing a corpus that was large enough, representative, and balanced, I had to extract the word lists from the corpus. As with other corpus projects (e.g., S. Fraser, 2007; Konstantakis, 2007), I wanted to come up with a list of words which, with the high-frequency words that EAL learners are already likely to know, would take coverage of the IB textbooks as close as possible to the target of 95% of the running words (the suggested minimum coverage for comprehension). Researchers have found (e.g., P. Nation & Waring, 2019) that a high-frequency word list such as the GSL only gives an average of 82% coverage of most written texts.

To create this first iteration of the International School Academic Vocabulary Lists (IS-AVL), I followed the procedure set out by Coxhead (2000) and Greene and Coxhead (2015). As I noted in Chapter Two section 2.3.2, Coxhead (2000) used three criteria when creating the AWL. First, she excluded the high-frequency words that learners would already be likely to know. Second, she wanted to focus on words with an appropriate range in the corpus. She therefore excluded any word that did not occur at least 10 times in each of the four discipline areas of her corpus (arts, commerce, law, and science). Third and finally, she looked at the frequency of the words. To do this, she excluded any words that did not occur at least 100 times in her corpus. This process allowed her to compile an academic word list of 570 word families that provides around 10% coverage of an academic corpus (see Table 5.3).

Table 5.3

Coverage of Coxhead's Academic Corpus by the GSL and AWL

Word list	Coverage of corpus
Most frequent 1,000 words	71.4%
2nd 1,000 most frequent words	4.7%
Academic Word List	10.0%
Total	**86.1%**

Note. Adapted from "A new academic word list," by A. Coxhead, 2000, *TESOL Quarterly*, *34*(2), p. 224. (https://doi.org/10.2307/3587951)

In their middle-school corpus, Greene and Coxhead (2015) followed a slightly different methodology. They compiled their corpus from four groups of words. They selected the first two groups of words using the AWL as a starting point. The first group included any AWL family members that had a minimum frequency of 28.5 tokens per million within the middle school corpus and a minimum frequency of 11.4 tokens per million in each of the subcorpora. To capture words that were frequent in only one subject, they then compiled a second group of words from words that had a frequency of 28.5 tokens or higher per million for the whole corpus and a frequency of 11.4 tokens per million in one of the subcorpora. The third group was made up of non-AWL words that had a minimum frequency of 28.5 tokens per million within the whole corpus and 11.4 tokens per million in each of the subcorpora. The final group comprised any non-AWL words that occurred with a minimum frequency of 100 times per million in a specific subcorpus. They added the words from groups one and two to all the subject-specific word lists. They only added the words from groups three and four to the word list in which they occurred at a frequency higher than the minimum Greene and Coxhead set for inclusion in those groups. As discussed in Chapter 2 section 2.3.4, this resulted in a collection of five word lists that ranged in size from 321 word families to 435 word families (see Table 5.4).

Table 5.4

Word Types and Word Families in the Different Content Area Middle School Vocabulary Lists

Content area list	Number of types	Number of word families
English grammar and writing	722	374
Heath	802	406
Mathematics	616	321
Science	859	435
Social studies and history	809	394

Note. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 147. Paul H. Brookes Publishing.

When combined with the GSL, Greene and Coxhead's word lists provide coverage of 88.49% to 92.17% of the various subcorpora of the middle school corpus (see Table 5.5).

Table 5.5

Coverage of the parallel subcorpora by the Middle School Vocabulary Lists and the GSL

	Subcorpora of the parallel corpus				
	English grammar and writing	Health	Math	Science	Social studies and history
GSL (%)	82.41	84.00	79.45	79.36	78.53
MSVL (%)	6.08	8.17	9.41	9.48	5.95
Total	88.49	92.17	88.86	88.84	84.48

Note. GSL = West's (1953) General Service List, MSVL = Middle School Vocabulary Lists. Adapted from *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*, by J.W. Greene and A. Coxhead, 2015, p. 150. Paul H. Brookes Publishing.

Although both Coxhead's AWL and Greene and Coxhead's Middle School Vocabulary Lists provide less than the recommended 95% coverage when combined with the GSL, they do still provide a significant amount of coverage with few word families. For the initial attempt at compiling a set of International School Academic Vocabulary Lists, I followed a modified version of the procedure outlined by Greene and Coxhead.

5.3.1 Identification of International School Academic Vocabulary Words

In this chapter, I followed Greene and Coxhead's (2015) methodology to create the IS-AVLs. I carried all of the calculations for compiling the word lists in R using the corpus data frames that were compiled and cleaned in the method outlined in section 5.2.2. The creation of these word lists involved identifying three different groups of words.

The first group of words comprises words which are common across multiple IB subjects. I identified these words based on two criteria, that the words: 1) occurred with a minimum frequency of 28.5 times per million words across the whole IS-CAT corpus; and, 2) had a minimum frequency of 11.4 times per million words in each of the eight

subject-area subcorpora. These are the same frequencies and ranges used by Coxhead (2000) and Greene and Coxhead (2015). I tried adjusting these numbers while creating the corpus, but these seemed to give the best balance of the number of words included in each list and the coverage provided. I included the 210 words that I identified as being part of this first group in all eight of the subject area word lists.

The second group of words were academic words that occurred frequently in a single subcorpus. I identified members of this group of words using three criteria: 1) the word was included in Coxhead's (2000) AWL; 2) the word occurred at least 28.57 times in the whole IS-CAT corpus; and, 3) the words occurred at least 11.4 times per million words in a particular subcorpus. I added these words to the subject in which they appeared over 11.4 times per million. In this stage, I was able to identify between 201 and 367 additional words that provided between 2.38% and 5.04% coverage (see Table 5.6). These numbers may seem low, but it is important to remember that the 1,000 word families of the BNC/COCA (P. Nation, 2020) 3,000 frequency band only provided around 2% coverage for most of the subject-specific corpora.

Table 5.6

Number of Words Identified in Stage Two and the Coverage of Those Words

Subject	Words	Coverage
Literature	284	2.85%
Biology	297	2.68%
Chemistry	284	3.74%
Physics	238	2.38%
Math	201	3.63%
Economics	367	5.04%
Social Studies	312	3.07%
TOK	363	3.19%

The final group consisted of non-academic words frequently occurring in a specific subcorpus. I created this group of words by identifying words that occurred with a minimum frequency of at least 100 words per million in one subject. I added these words to the subject list of any subject they occurred in, with a frequency of over 100 words per million. The number of words that I identified at this stage varied from 52 to 305 for the different subjects, and these words provided an additional 1.18% to 9.48% coverage (see Table 5.7).

Table 5.7

Number of Words Identified in Stage Three and the Coverage of Those Words

Subject	Words	Coverage
Literature	52	1.18%
Biology	305	9.33%
Chemistry	258	9.48%
Physics	204	5.9%
Math	125	4.51%
Economics	102	2.11%
Social Studies	133	2.89%
TOK	63	1.37%

The final word lists that I identified using the three steps above ranged from 530 words to 812 words (see Table 5.8). All the subject-specific lists contain fewer word families than the 570 word families that make up the AWL, showing that they should be a more manageable size for most international school students. The coverage these lists provide of the different subjects in the IS-CAT is between 6.72% to 17.77%. While this level of coverage is slightly higher than that calculated by Greene and Coxhead (2015), this is only to be expected given that the corpora described in this chapter are more focused than the ones in Greene and Coxhead's study. For example, in my corpus, each

of the sciences is divided into a separate subcorpus, whereas Greene and Coxhead combined the different sciences into a single subcorpus.

Table 5.8

Number of Words and Coverage of the Final IS-AVL

Subject	Words	Families	Coverage
Literature	546	357	6.72%
Biology	812	554	15.56%
Chemistry	752	483	17.77%
Physics	652	430	12.15%
Math	530	348	11.77%
Economics	679	438	10.76%
Social Studies	655	436	8.56%
TOK	636	409	7.77%

5.3.2 Coverage of the International School Academic Vocabulary Lists

I created the International School Academic Vocabulary Lists (IS-AVL) using frequency and range criteria, following the procedures laid out by other researchers in the field. That it provides coverage of around 6.72%–17.77% across the eight subject-specific subcorpora indicates its general utility and validity. The high coverage means that it has the potential to provide the EAL learner with an extremely useful set of words. Furthermore, the size of each of the subject-specific lists is manageable while still providing coverage of a significant percentage of the vocabulary learners are likely to encounter in the classroom.

When we include the GSL lists in our analysis, we can see that the coverage of general vocabulary, along with the subject-specific word lists, gives us around 83% to 90% coverage, depending on the subject (see Table 5.9). We need to remember that the 95% coverage required for understanding comes from studies investigating the reading

of general texts and reading for pleasure. The required amount for reading academic texts in the classroom may be different. Furthermore, researchers such as Schmitt et al. (2011) have suggested that there is not, in fact, a set point at either 95% or 98% and the actual amount required may be both text- and learner-dependent. Nevertheless, it seems indisputable that a range of 9% to almost 17% of unknown words found in the different subcorpora is considerable. Such a large number of unknown words would be equivalent to one or two unknown words in every line.

Table 5.9

Combined Coverage of the GSL and the IS-AVL

Subject	GSL	IS-AVLs	Combined
Literature	81.99%	6.72%	88.71%
Biology	71.46%	15.56%	87.02%
Chemistry	69.1%	17.77%	86.87%
Physics	76.51%	12.15%	88.66%
Math	71.15%	11.77%	82.92%
Economics	79.77%	10.76%	90.53%
Social Studies	74.88%	8.56%	83.44%
TOK	81.02%	7.77%	88.79%

5.4 Limitations and Potential Criticisms of the IS-AVL

While the production of the word lists outlined in the current chapter are an important step towards creating a vocabulary list for international school students, there are still several potential criticisms of the lists in their current form. The main problems relate to the issue discussed in the introduction: the IS-AVL is built upon two outdated word lists, the GSL and the AWL. Although this was necessary as a way of ensuring that the corpora created for this endeavour are capable of producing word lists that are representative of the textbooks they are taken from, it should be possible to create word

lists built upon more up-to-date foundations. The New General Service Lists, developed by Brezina and Gablasova (2015), are one possible option for a list of general high-frequency vocabulary that could be used when compiling the IS-AVL. There is also the option to use the BNC/COCA (P. Nation, 2020), or to not rely on a general frequency list at all (see, Gardner & Davies, 2014).

This brings me to an additional problem with the current study. Despite the usefulness of the IS-AVLs, by focusing exclusively on academic and mid-frequency vocabulary, they ignore a large number of academic words and lay-technical words which are found in the most frequent 2,000 word lists. As explained in Chapter 2 section 2.3.6, other studies have shown that a large number of general words found in academic textbooks are actually forms of words found on the high-frequency word lists. The danger is that, even if the learners are familiar with all the words in the GSL, they will not necessarily be aware that many of them have an additional academic meaning. Some examples of GSL words with academic meanings found in the IS-CAT are given in Table 5.10.

Table 5.10

Words with Academic Meanings Found in the GSL

First 1,000 words	action, activity, agent, control, current, expression, local, population, rate, trial
Second 1,000 words	blind, block, messenger, model, resistance, risk, sample, treatment

The final issue worth noting when considering the word lists compiled in the current chapter is how words were selected and the use of word families. Modern methods of textual analysis have made it much easier to lemmatize words and tag them as particular parts of speech (POS tagging). This enables researchers to differentiate between different forms of the word, for example, the difference between “to estimate” or “an estimate”, where the same word can be used as a verb or a noun. By first

lemmatizing the corpus and tagging for parts of speech, we can get a much better idea of the actual lemmas present in the corpus.

5.4.1 Words not in the list

It is also useful to investigate the words that were not included in any of the lists to determine what kind of words they are and how important they might be. This can help to glean a better understanding of how difficult it would be for learners to understand these words in context. A better understanding of what words I did not include can also help inform us if it is better to decrease the frequency of the cut-off points for inclusion into the final word lists. Although doing so would increase the number of words in the final lists, it could allow us to include words that occur more infrequently but convey crucial information.

Initially, many of the words not included in the word lists are technical and subject specific. Perhaps not including such words in the list is acceptable as words closely related to learning the subject present conceptual rather than linguistic difficulties (S. Fraser, 2010). For example, in chemistry, many of the terms not on the list are the names of specific elements or chemical compounds. The names of these compounds could, perhaps, be treated similarly to proper names. If they were treated in such a way, they could then be excluded from the list of vocabulary that learners need to comprehend a text, just as it is unnecessary to teach students all the names of the characters in a novel before asking them to read it. Also, many technical words, such as chemical compounds, contain regular affixes, such as *hyper-* or *hydro-*, which make them easily identifiable and easier to learn than other low-frequency words.

5.5 Conclusion

The main objectives of the study presented in the current chapter were to: 1) build a large corpus that represents the textbooks being used in the IB classroom; and, 2) use the frequency and range criteria formulated in Greene and Coxhead (2015) to create an initial version of the IS-AVL that can be profitably used by EAL learners. These lists, in conjunction with the most frequent words, were constructed to enable EAL learners to reach the threshold necessary to understand the textbooks that they are using in the classroom. In this concluding section, I will discuss the areas in which the study was

successful and then show the limitations which need addressing. Finally, based on these findings, I will outline a plan for how to proceed from here.

5.5.1 Achievements of the study

Although it was not possible to reach coverage of 95%, the study has been successful in creating a list which, when combined with the GSL, achieved 82.92% to 90.53% coverage of a corpus of subject-specific textbooks. The 6.72% to 17.77% coverage provided by the IS-AVL is greater than the coverage the larger Academic Word List provides over the corpus from which it was compiled. The current preliminary study has also confirmed the usefulness of the IS-CAT from which I compiled these lists. The subcorpora that make up the larger corpus are large enough to identify the words learners would need to know to understand these textbooks. In addition, I have identified several issues with the processes used to compile these initial lists that are deserving of further investigation. If I can address these issues appropriately, it will allow me to improve the IS-AVL described in this chapter.

5.5.2 Limitations

There are three limitations that I identified when compiling these word lists that need to be addressed going forward. First is the use of the AWL and GSL. Although using the AWL allowed me to better identify academic words in the corpus, it was not without its issues. The words identified using the AWL provide less coverage for more tokens than the words identified in the other two steps. Using the AWL and GSL led me to exclude certain high-frequency words with academic meanings from the final word lists. Second, counting the words in the corpus, rather than first lemmatizing them and tagging them for part of speech, resulted in certain words essentially being counted twice (i.e., as separate words). This is clear when looking at the word list derived from the literature corpus, where *text* and *texts* were the two most frequent words on the list. By lemmatizing the texts, these two words would be more accurately counted as the same entry. As we know that even lower-proficiency learners can understand the different forms of lemmatized words (Pinchbeck et al., 2022; Webb, 2021), this would provide a more effective way for counting the tokens in the text. Finally, the procedures used by Coxhead and Greene, while still effective, are now outdated. As I explained in Chapter 2, more modern techniques for compiling word lists can produce better results.

Other than the issues described above, it would also be good to experiment with different range and frequency criteria, as those used in the present study may not be the most appropriate for all our purposes. This may help us address one of the biggest problems with the study at this stage: the IS-AVLs, despite their benefits, are still not sufficient to provide learners with all the words that they need to know to understand the textbooks they have to read. Adjusting the frequency cut-offs to include more words, allowing for the inclusion of high-frequency words, and using lemmas rather than words as the unit of counting will help to address these issues.

5.5.3 A way forward

I have identified several areas meriting further attention. However, the next step should be the creation of an improved word list: ideally, one which provides better coverage and also considers the fact that general words list such as the GSL often contain academic words. One way of getting around the problem would be to adjust the methodology used to create the word lists. The adoption of a methodology based on frequency and range, which does not distinguish between general purpose and academic vocabulary, that uses lemmas as the unit of counting, and that allows me to experiment with different frequency and range cut-off points would enable me to create a better set of word lists; I discuss these in the following chapter, Chapter Six.

Chapter Six

Integrating Modern Techniques and Validating the Word Lists

6.1 Introduction

In Chapter Five, I detailed how I created the IS-CAT corpus and outlined my initial attempts to create a set of subject-specific word lists from that corpus. I begin this chapter with an overview of the achievements of that study, together with the problems that I identified with the word lists I created in that chapter and the processes used to create them. I then explain the steps that I have taken to address these issues and introduce the final versions of the International School Academic Vocabulary Lists (IS-AVL). Finally, I look at how the coverage the word lists that I have created provide over several different corpora and compare them to other available word lists. I hope that this enables me to demonstrate how the revised and more modern processes for creating word lists described in this chapter have allowed me to create a set of word lists that consider current research on the importance of lemmas, do not arbitrarily separate academic from general vocabulary, and do not rely upon outdated general word lists.

6.1.1 Creating the International School Academic Vocabulary Lists: Achievements

In Chapter Five of this dissertation, I could achieve the following:

1. I compiled and cleaned a 9.25-million-word corpus that covered eight subject areas that are regularly taught in the IB programs at international schools. The subcorpora were all around 1 million words long, making them large enough for me to identify the important high- and mid-frequency words in each subject.
2. I constructed a set of preliminary word lists using the techniques set out by Greene and Coxhead (2015) when they created an academic corpus for middle-school students. Each of the subject-specific word lists ranged in size from 348 to 547 word families. All the lists are smaller than the AWL and are thus of a more manageable size for both students and teachers.
3. I showed that these word lists provide a significant amount of coverage over the eight academic subjects (6.72% to 17.77%). This is much greater than the coverage provided by the larger AWL over the same set of subcorpora. This is

important because it means that these word lists would be more effective in supporting EAL learners in the classroom.

6.1.2 Creating the International School Academic Vocabulary Lists: problems

However, despite these initial achievements, the lists that I created in Chapter Five are not without their flaws. Some concerns that still need to be addressed are:

1. The current version of the IS-AVL relies on the GSL and the AWL: therefore, they inherit some issues associated with both lists. As I discussed in the previous chapter, the GSL is considered to be outdated (see, e.g., Brezina & Gablasova, 2015; Gardner & Davies, 2014), and the AWL's use of this list and its reliance on word families are problematic (see, e.g., Green & Lambert, 2018; Hyland & Tse, 2007)
2. Because the GSL was used to remove high-frequency words, the IS-AVL I created in Chapter Five excludes high-frequency words with an academic meaning.
3. The word lists in Chapter Five use words and not lemmas as their unit of counting. This leads to some very similar words, such as "text" and "texts", being included separately in the list.

Despite these three problems, creating the IS-AVL in the previous chapter was worthwhile for three important reasons. First, when compiling these lists, I was able to determine how effective the corpora I used were at representing IB textbooks. Second, it allowed me to identify and correct any problems that existed with the corpora. Third, it allowed me to determine that the corpora were large enough to allow me to identify the words that EAL learners would need to be able to understand the texts in each of the discipline-specific subcorpora. Although the lists themselves will need to be updated in this chapter, the techniques and tools that I acquired through creating these initial word lists have enabled me to progress to the next stage of this project.

6.1.3 Addressing the issues

In the current chapter, my goal is to revise the IS-AVL that I created in Chapter 5 to consider the newer techniques that more powerful computers and more adaptive programming languages (such as Python and R) have made possible. I therefore hope to

address the three issues I raise above. Before discussing the methodology that I use to address these problems, it is worth discussing in more detail why these are, in fact, problems that need to be fixed.

The most important issue that needs to be addressed is the use of the AWL and the GSL to remove high-frequency words from the corpus. Using this type of general word list is an important step in the creation of many academic, or technical, word lists as it allows the researcher to remove general high-frequency words from the resulting word lists; researchers do this for two reasons. First, we assume that most language learners who need an academic word list are probably already able to master most, if not all, of the common high-frequency words. Nation (2005) classified words into high-frequency, academic, technical, and low-frequency words. Schmitt and Schmitt (2014) expanded this classification slightly by proposing a set of what they called mid-frequency words to cover the vocabulary in between the high-frequency and low-frequency groupings. The idea is that learners should aim to acquire the high-frequency words first, as those are the ones that provide the "biggest-bang-for-your-buck". This is because high-frequency words make up such a large part of any given corpus. For example, the same top 10 words, words such as *the*, *and*, and *a*, make up about 20% of any given corpus (Schmitt & Schmitt, 2020). Because of this, it is generally assumed that by the time learners reach more academic or technical classes they will already have enough exposure to English to have had already learned these words (see Schmitt & Schmitt, 2020, p. 14 for a comparison of the expected vocabulary size for EFL learners from different countries).

Second, many of the words included in these high-frequency lists are function words. Function words make up around 40% of most corpora, and the percentage of function words in a corpus, unlike content words, does not vary across corpus type (general, academic, specialized) or discipline (Schmitt, 2010). Function words also have what Green and Lambert (2018, p. 110) referred to as "low teachability" and are not related to the discipline of the corpus they appear in.

Function words can be removed from a word list in several ways. Coxhead (2000), for example, used West's (1953) General Service List (GSL) to exclude these high-frequency words from her academic corpus. Gardner and Davies (2014) compared

the frequency of words in their academic corpus to a non-academic corpus and removed words with a similar frequency in both. Because the frequency of function words does not change much with the type of corpus being analyzed this would have allowed them to remove most, if not all, function words from their academic corpus. Green and Lambert (2018), explicitly removed these words by filtering out any words that were not nouns, verbs, adverbs, or adjectives from their final list.

In Chapter Five, I followed Coxhead's technique and used the GSL to remove all high-frequency words from my final word lists. However, as I discussed previously, this use of high-frequency lists to remove words from a corpus during the creation of a word list is controversial (e.g., Gardner & Davies, 2014; Lei & Liu, 2016; Neufeld et al., 2011). There are two issues with this approach. First, the high-frequency word list most researchers use for this purpose is the GSL. West's GSL was compiled in 1953 and is now very out of date. For example, the GSL includes words such as *sow, barber*, and *shilling* as frequent words, but it does not include more modern words such as *computer* or *television*. Second, excluding words from the GSL may result in the exclusion of high-frequency words with academic meaning (Gardner & Davies, 2014). Unfortunately, in Chapter Five, it was necessary for me to use the GSL because many of the steps I took to compile these lists involved using Coxhead's (2000) AWL, which itself is based upon the GSL.

The second issue that I need to address from the previous chapter is the exclusion of high-frequency words from our academic word lists. Even setting aside the issues with GSL discussed above, excluding high-frequency words from an academic word list is problematic. Just because a word appears in a high-frequency list does not mean that it is not academic. Neufeld et al. (2011) found that many of the items from the AWL, are actually among the most frequent words in the British National Corpus (BNC) and Corpus of Contemporary American English (COCA). Gardner and Davies (2014) also that there are a number of general high-frequency words that occur more frequency in academic English than they do in general English. In fact, many researchers (e.g., Gardner & Davies, 2014; Lei & Liu, 2016; Neufeld et al., 2011) have argued that it is not possible to make a clear-cut division between high-frequency words and academic words based on frequency alone. Using other techniques, such as frequency ratio (Green & Lambert, 2018) and manual inspection of the excluded high-

frequency words (Lei & Liu, 2016), are important steps that a researcher should take to ensure they include these important high-frequency academic words in their final lists.

The third and final problem that needs to be addressed concerning the word lists I compiled in Chapter Five is the use of words, and word families, as a unit of counting. Gardner and Davies (2014) and subsequent researchers who have compiled word lists (e.g. Brezina & Gablasova, 2015; Browne, 2014; Lei & Liu, 2016) have adopted the lemma, instead of the word family, as the primary unit of counting and reporting for their word lists. There are several reasons why using the lemma is better suited as a unit of counting than either words or word families. First, the lemma form is much more informative and user-friendly than the word family form for both pedagogical and research purposes (e.g., Brown et al., 2020; Myint Maw et al., 2022). It can be extremely difficult for foreign language learners to understand the inflectional relationship between the head word and the other word-family members (Gardner, 2007). For example, most learners would not recognise that *prerecord*, *recorder*, and *unrecorded*, along with their inflections (such as *prerecorded*, *recordings*, *recorders*), belong to the same word family (McLean, 2021). If researchers use word families to create a pedagogical word list, this will further disadvantage those most in need of help, as a learner's ability to understand word morphology depends on their existing vocabulary knowledge (Nagy, 2007).

Second, using lemmas is preferable as this allows the researcher to focus on smaller units as well as take into account grammatical parts of speech (e.g., nouns, verbs, adjectives, adverbs). Using lemmas is important because without knowing the part of speech many words that are written the same can have very different meanings. Gardner and Davies (2014) use the example of the word *proceeds* to exemplify this: "without grammatical identification, the verb *proceeds* (meaning continues, and pronounced with stress on the second syllable) and the noun *proceeds* (meaning profits, and pronounced with stress on the first syllable) would be counted as being in the same word family" (p. 308).

By using more modern techniques to develop my word lists, I can overcome many of the problems discussed here. However, the reason that I started with the method described in Chapter Five is that these modern techniques are more challenging

than the ones Coxhead (2000) and Greene and Coxhead (2015) used to create their word lists. Furthermore, these updated techniques require the use of more flexible computer programming languages such as Python and R, making the barrier to entry much higher. Fortunately, there are several researchers whom I can look to in helping me in this endeavour. These include Gardner and Davies (2014) who used a new method that considers all the issues noted above to develop an effective and representative general academic vocabulary list. Lei and Lui (2016), further refined this approach to compile a new word list for Medical English. Green and Lambert (2018), used a similar set of techniques to compile domain-specific academic vocabulary lists for secondary school students studying in Singapore.

Thus, in this chapter, I build upon the work by these researchers to develop a new set of domain-specific academic word lists for EAL. In doing so, I address the following two questions:

1. Do these new methods do a better job of identifying academic vocabulary that would be useful for EAL learners than the techniques used by Greene and Coxhead (2015) that were described in the Chapter Five?
2. How does the coverage provided by the academic vocabulary lists developed using these techniques compare to the coverage provided by existing word lists over a corpus of international school textbooks?

6.2 The study

6.2.1 Methodology

In the current chapter, I followed a revised version of the methods described by Gardner and Davies (2014), Lei and Liu (2016), and Green and Lambert (2018) to revise the IS-AVL that I created in Chapter Five. As before, R (R Core Team, 2022) was used for most of the data analysis. The one exception was that the Python Natural Language Processing (NLP) library spaCy (Honnibal et al., 2020) was used to lemmatize and tag the texts for part of speech. This was because NLP libraries rely on a trained model to effectively lemmatize and tag the text. While it is possible to train your own model, this would have been beyond the scope of this dissertation, so I used one of the pre-trained models that spaCy makes available for download. To balance the time taken to process the data with the accuracy of the part of speech tagging, I used the en_core_web_md

library (https://spacy.io/models/en), which gives above 97% accuracy when tagging parts of speech. While I could have achieved greater accuracy by using a larger model, even the largest available model only had 98% accuracy, and it would have taken significantly longer to run on a corpus of this size.

As with the previous corpus, selecting words for inclusion on the IS-AVL was an iterative process that involved several stages. The next section discusses how and why I performed each of these steps. I show all the R scripts I used to complete this process in Appendix D.

6.2.2 Compiling the word lists

Compiling the final word lists entailed six steps. The first two stages involved selecting the words with the highest frequency and removing word classes that would not be included in the final word lists.

Step 1: Following the approach used by Lei and Liu (2016) and Green and Lambert (2018), I first removed any words that did not occur with a minimum frequency of 28.57 times per million words. However, this time, rather than select words for their coverage across the whole corpus, I instead used the coverage that the words gave within each discipline.

Step 2: The second stage of the process was to eliminate the function words from the lists. As discussed above, most academic word lists do not include function words for three reasons: 1) they tend not to vary much across academic and general corpora, so do not classify as academic words; 2) they are usually extremely frequent across both general and academic corpora, so most learners will already be familiar with them; and, 3) they have a low level of teachability, meaning that they are difficult for learners to pick up even through explicit instruction. Although Green and Lambert (2015) waited until the last stage of their analysis to remove these words from their word lists, I felt it was best to remove these words at the start of the process. Removing them early, made it easier to see the type of coverage I was able to achieve at each stage of the process. The word lists I compiled at this stage consisted of between 1,621 lemmas and 2,523 lemmas and provided between 38.47% and 49.53% coverage of the different corpora (see Table 6.1).

Table 6.1

Number of Words Identified in Stages One and Two and the Coverage of Those Words

Subject	Step 1		Step 2	
	Lemmas	Coverage	Lemmas	Coverage
Literature	2874	89.47%	2280	38.47%
Economics	2630	93.30%	2231	49.53%
Social Studies	3153	91.71%	2390	41.04%
Biology	2852	92.77%	2523	47.33%
Chemistry	2370	94.52%	2020	48.55%
Physics	2327	94.88%	1972	45.70%
Maths	2008	94.90%	1621	43.17%
TOK	2615	91.39%	2193	41.84%

Step 3: The next two steps look at the dispersion of the lemmas across the texts that make up the IS-CAT. First, I investigated how many books each of the lemmas occurred in, to ensure that the lemmas were not only clustered within a small number of texts. I therefore checked to make sure that the lemmas in each of the word lists appeared in at least half of the textbooks that make up the relevant subcorpora. I removed any words that had a range of below 50% from the subject-specific word list. Because the IS-CAT is made up of a few longer textbooks, only between 16 and 124 lemmas were removed from each of the lists. The coverage of the lemmas that I removed at this stage was never above 1%.

Step 4: Next, I calculated the dispersion of each lemma across each of the subcorpora. Before looking at the results that I attained from this analysis, and the resulting word lists, we first need to establish exactly what dispersion is and why calculating it is an important part of compiling a word list. While it may be obvious why frequency is important (a word that only occurs once for every 1 million or 2 million words is probably not worth learning), the importance of dispersion can be harder to understand. However, nearly all modern word lists are compiled using both frequency

and dispersion measures (Biber et al., 2016). Dispersion allows learners to focus on words that are both frequent and occur widely throughout the texts and to ignore words that, while frequent, only occur in one or two texts, or one or two places in the text.

There are several measures of dispersion that can measure the distribution of a word across the parts of a corpus. The most common measure of dispersion is Juilland's D (see, Burch et al., 2017; Juilland & Chang-Rodriguez, 1964). Calculating Juilland's D returns a measure between one and zero. Zero indicates that the word only occurs in one part of the corpus, while one shows that the word is uniformly spread across the parts of the corpus. In previous studies, researchers have used values of between .8 (Gardner & Davies, 2014) and .5 (Green & Lambert, 2018; Lei & Liu, 2016). However, Juilland's D is not without its problems (for a discussion of these issues, see, Biber et al., 2016; Gries, 2022). Because of this, I adopted the use of an alternative and more modern measure proposed by Gries (2008) called the Deviation of Proportions (DP). DP is both more effective and easier to compute than Juilland's D.

I calculated the DP for each of the lemmas by splitting each of the subcorpora into ten equal parts. I then compared the observed to the expected frequency of each lemma in each of the parts of the subcorpus. Because DP produces a value that is the opposite of Juilland's D (the lower the number, the more uniform the dispersion of the word), I used the technique suggested by Biber et al. (2016) and subtracted the calculated DP from one. Words with a DP lower than .7 in their respective subcorpora were removed from the word lists. The resulting word lists now contain between 842 and 1451 words and give between 30.38% and 40.28% coverage of their respective subcorpora (see Table 6.2).

Table 6.2

Number of words identified in stages one and two and the coverage of those words

Subject	Words	Coverage
Literature	1451	32.26%
Economics	1015	30.38%
Social Studies	1040	26.38%
Biology	1412	39.25%
Chemistry	1174	40.28%
Physics	1133	37.71%
Maths	842	34.10%
TOK	1377	34.94%

Step 5: In the next two steps, I examined the lemmas on the word list based on their range in one part of the corpus compared to another part of the corpus. Lei and Liu (2016) referred to this as a range ratio. Unlike the range calculation from Step 2, the range ratio ensures not just that the lemmas occur in most of the texts in the corpus, but that they occur with similar frequency throughout these texts. To do this, I calculated how many times each lemma occurred in each text in the subject-specific subcorpus. I then removed any lemmas that did not occur at a frequency of at least 20% of the total frequency at which they occurred in the whole subcorpus in at least 50% of the text. I did not remove any lemmas at this stage, probably because I would have removed lemmas that were clustered together in a few texts in steps one and four.

Step 6: I had assumed that I would be able to stop at this point in the process. The lists I have been able to compile up to this point in the process provide much better coverage per token of the IS-CAT than existing word lists, including the ones that I compiled earlier as reported in Chapter Five. The number of lemmas and coverage they provide is also similar to what Green and Lambert (2018) achieved in their study. However, there were two issues that I had with the word lists at the end of Step 5. First, the number of lemmas on each of the lists is much too great: it would be difficult for

EAL learners to acquire this many lemmas in the time that they have. Second, when I inspected the lemmas on the lists, I found that there were too many general high-frequency lemmas with no additional academic meaning on the list. An inspection of the resulting word lists shows (see Table 6.3 for a list of the top 15 lemmas on each list at this stage in the process) that the lists contain words such as *have*, *use*, *make*, and *write*. To remove any non-academic high-frequency words, I followed Lei and Lui' (2016) procedure and used a general word list to remove the high-frequency words; I then checked these words manually to see if any could be considered to be academic and re-added those to the final lists. While Lei and Lui used the new General Service List (Brezina & Gablasova, 2015), I opted to use the first 1,000-word band from the BNC/COCA (P. Nation, 2020). This is because research has shown that EAL learners in an international school context will likely have acquired the vocabulary on this list well before entering the IB diploma program (see, Brooks et al., 2021; Coxhead & Boutorwick, 2018). This removed between 263 and 534 lemmas from each of the domain-specific lists. Of the total 1,010 lemmas that I initially removed from the lists at this stage, I added 89 lemmas back to the lists after checking to see if the removed lemmas were academic.

Table 6.3

An Example of the Top 10 Most Frequent Words for Five Subjects Before Using the BNC/COCA to Remove High-Frequency Words

Rank	Literature	Chemistry	Biology	Economics	Social
1	text	reaction	cell	market	war
2	**have**	electron	**have**	**have**	**have**
3	**use**	atom	**use**	cost	government
4	**work**	ion	plant	**use**	power
5	**language**	**use**	molecule	firm	force
6	**make**	**have**	water	make	country
7	reader	energy	gene	example	political
8	**time**	acid	blood	**such**	**other**
9	**way**	bond	protein	increase	**make**
10	**write**	molecule	produce	**high**	**also**

Note. Bold type denotes high-frequency words.

6.3 Results and discussion

The International School Academic Vocabulary Lists (IS-AVL) that I compiled in the current chapter are an improvement on the ones I created in Chapter Five. The resulting lists achieve greater coverage of the different disciplines with fewer words. They also maintain useful academic words that were removed from the lists in the previous chapter because of the reliance on West's (1953) GSL to remove high-frequency words in that chapter. The word lists that I was able to compile using the steps outlined above provide coverage of between 11% to 23% of the relevant subcorpora across the eight subjects (see Table 6.4). The lists range from 379 to 845 lemmas. While slightly longer than the lists I compiled in Chapter Five, they are still of a size that should be manageable for EAL learners. It should be noted, at this stage, that by increasing the dispersion measure to one more in line with Gardner and Davies (2014), I was able to

achieve word lists that were much smaller than those I compiled in the last chapter but that nevertheless provide greater coverage (between 296 to 580 lemmas while still providing 7.5% to 19.1% coverage). However, the size of Gardner and Davies' corpus (over 120 million running words) was significantly larger than the one in this study, so a more conservative number seemed both more appropriate, and the resulting lists provided improved coverage.

Table 6.4

The Number of Words and Coverage Provided by the IS-AVLs

Subject	Words	Coverage
Literature	610	9.96%
Economics	481	11.44%
Social Studies	526	9.43%
Biology	845	20.70%
Chemistry	681	22.55%
Physics	578	17.45%
Maths	379	13.54%
TOK	685	12.81%

One disadvantage of not using a general word list in the process of compiling the IS-AVL is that it is difficult to look at the coverage of the resulting word lists in conjunction with a high-frequency word list. This is important if we want learners to be able to reach the 95% coverage of the different texts research (e.g., Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010) shows is the minimum coverage necessary for understanding. However, given what we know about the vocabulary profiles of EAL learners (Brooks et al., 2021), it is reasonable to assume that they probably cannot reach 95% coverage of the IS-CAT, even with the assistance of the IS-AVLs. While this is not ideal, the coverage provided by these lists is greater than that of the other available lists, making them a good starting point for teachers in the international school context.

The first 30 words from each of the subject-specific lists are shown in Table 6.5. Looking over these lists, we can see that these are words one would intuitively consider to represent the types of vocabulary a learner would need to know for that subject. I give the complete word lists in Appendix E. The final IS-AVL contains 845 lemmas for Biology, 685 for TOK, 681 for Chemistry, 610 for Literature, 578 for Physics, 526 for Social Studies, 481 for Economics, and 379 for Maths. Together the lists contain 4,785 total lemmas and 2,136 unique lemmas. An examination of the words on each of the lists shows they do a good job of capturing the vocabulary EAL learners need to understand both underlying concepts, such as *government* and *revolution* in History and *supply and demand* in Economics, and technical terms, such as *photosynthesis* and *enzym*e in Biology and *molecule* and *hydrogen* in Chemistry. They will also help teachers to support their students with words that have different meanings in different domains. For example, the word *reaction* is on both the Social Studies and the Chemistry list, but it carries a different meaning in each of those domains (to act in response to something in Social Studies, and what happens to chemicals when they come together in Chemistry).

Table 6.5

The Top 30 Most Frequency Lemmas, With Frequency, for Each Subject

Rank	Literature Lemma	Freq	Biology Lemma	Freq	Chem Lemma	Freq	Physics Lemma	Freq
1	text	5290	cell	8469	reaction	7901	energy	7422
2	language	2489	molecule	2786	electron	5317	mass	2938
3	poem	1969	gene	2642	atom	4437	speed	2554
4	literary	1795	protein	2501	ion	4423	particle	2440
5	example	1281	produce	2454	energy	4290	electron	2191
6	character	1165	organism	2383	acid	4251	example	2145
7	culture	1067	example	2111	bond	3965	temperature	1996
8	literature	1043	dna	2062	molecule	3748	distance	1865
9	image	984	species	1964	solution	3235	calculate	1714
10	novel	963	chromosome	1962	example	3127	equation	1707
11	create	932	acid	1907	hydrogen	2651	direction	1659
12	explore	913	structure	1868	concentration	2473	graph	1618
13	author	911	enzyme	1861	carbon	2422	object	1600
14	include	903	energy	1811	temperature	2305	velocity	1597
15	issue	871	carbon	1807	equation	2301	value	1564
16	feature	837	membrane	1778	value	2266	constant	1546
17	context	832	occur	1631	structure	2195	surface	1516
18	chapter	827	reaction	1564	metal	2144	current	1447
19	audience	798	process	1455	compound	1968	potential	1416
20	non	778	contain	1259	mass	1942	frequency	1359
21	perspective	721	result	1195	element	1932	measure	1330
22	poetry	717	concentration	1159	produce	1636	electric	1326
23	structure	713	population	1134	cell	1605	motion	1274
24	fiction	664	oxygen	1109	product	1511	produce	1251
25	effect	626	muscle	1074	oxygen	1486	source	1199
26	global	623	sequence	1070	contain	1455	magnetic	1194
27	describe	621	allele	1057	calculate	1415	diagram	1174
28	response	604	area	1034	involve	1322	radiation	1167
29	narrator	595	include	1032	increase	1308	angle	1111
30	extract	579	tissue	994	determine	1272	resistance	1097

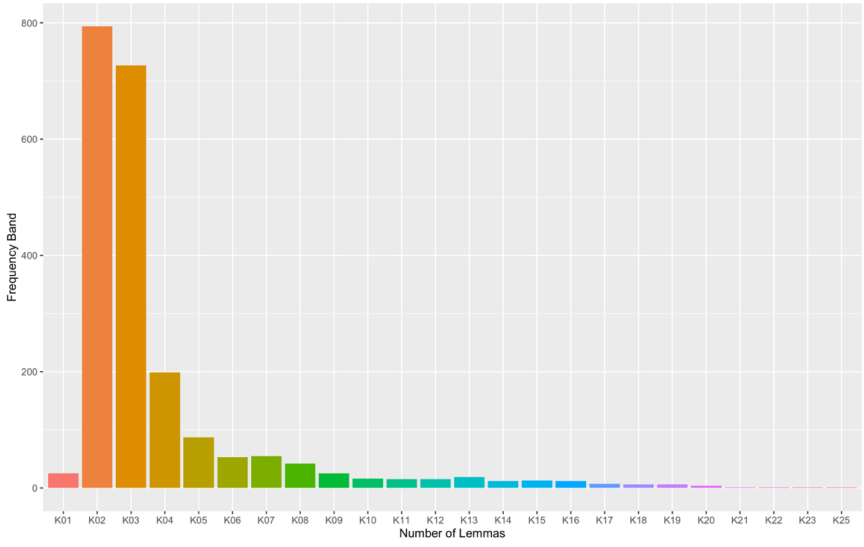
Rank	Maths	Freq	Economics	Freq	Social	Freq	TOK	Freq
	Lemma		Lemma		Lemma		Lemma	
1	function	6268	firm	3443	political	1957	knowledge	8943
2	equation	5521	example	2480	economic	1802	example	2821
3	value	5274	increase	2324	military	1678	language	2614
4	graph	3815	production	2201	army	1395	perspective	1499
5	example	2576	consumer	2161	troop	1223	claim	1366
6	solution	2296	profit	2148	include	1126	belief	1266
7	area	2107	output	1679	social	1012	area	1183
8	axis	1663	value	1587	attack	916	method	1166
9	hence	1618	produce	1414	policy	834	ethical	1162
10	solve	1599	resource	1399	economy	770	object	1162
11	curve	1439	include	1341	area	750	theory	1087
12	diagram	1391	revenue	1309	major	708	evidence	1079
13	variable	1355	reduce	1118	population	679	culture	1061
14	triangle	1329	factor	1109	increase	673	community	1049
15	angle	1292	unit	1087	establish	662	develop	1009
16	series	1223	decision	1041	develop	653	explore	1008
17	result	1198	provide	1026	region	647	include	892
18	length	1162	develop	981	role	635	argument	887
19	sequence	1158	social	980	create	633	value	864
20	sum	1116	capital	979	century	632	social	863
21	random	1104	benefit	959	result	622	concept	843
22	coordinate	1081	labour	930	event	598	chapter	842
23	operation	1077	investment	861	issue	598	extent	774
24	distance	1075	industry	857	impact	593	create	752
25	prove	1019	method	855	achieve	566	describe	750
26	chapter	986	financial	820	battle	565	role	741
27	calculate	978	objective	809	provide	549	process	726
28	standard	961	advantage	782	effect	540	society	712
29	unit	952	calculate	771	victory	528	individual	707
30	complex	938	loss	752	territory	510	event	699

6.3.1 What type of words are included in the IS-AVL

We can see that the coverage provided by the IS-AVL is quite comprehensive. The lists provide between 10% to 20% coverage across all the different subcorpora and can do so using very few words. Examining the words that comprise the word list might help us to develop a better understanding of these words. Doing so will help us to get a better understanding of how useful these words are likely to be for EAL learners. We first need to consider how difficult the words on the IS-AVL are likely to be for EAL learners to acquire, by looking at what type of words (high-, mid-, or low-frequency, or technical words) are on the different lists. Previous studies (see, P. Nation & Waring, 2019) have shown that learners acquire high-frequency vocabulary more easily than low-frequency or technical vocabulary. We also know that many EAL learners are likely to have some knowledge of at least the first 2,000 high-frequency words (Brooks et al., 2021; Coxhead & Boutorwick, 2018). An analysis of the vocabulary in the IS-AVL (see Figure 6.1) shows that most of the words come from the first 2,000 to 4,000-word bands of the BNC/COCA, with a few more items at the 5,000 to 8,000 level and very few after that. Such figures indicate that, given what we know about the vocabulary profiles of EAL learners, these are words they should be able to acquire.

Figure 6.1

The Relative Frequency of the Different Levels of the BNC/COCA in the IS-AVL



6.3.2 Comparison with other word lists

We can see from the discussion above that the IS-AVL can provide a large amount of coverage over the respective corpora. However, to determine whether these lists are indeed useful for supporting EAL learners, we must look at how they compare in relation to existing word lists. I selected the word lists from the general and academic word lists I looked at in Chapter Two, focusing on word lists that would commonly be used in an international school context. I checked each of these word lists against the various subcorpora of the IS-CAT. Where there were domain-specific word lists, I checked those lists against the domain they were compiled for. For the Middle School Vocabulary Lists, where all the sciences are grouped into a single list, I ran the science list against the biology, chemistry, and physics subcorpora. An analysis of the coverage provided by these different lists over the various subdomains of the IS-CAT shows that, mostly, the IS-AVL provides greater coverage than the lists currently being used by teachers in the classroom (see Table 6.6). What makes the IS-AVL even more effective than these existing lists is that it can provide better, or similar, coverage with significantly fewer words. With one exception (which I discuss below) the few times that another word list provided similar, or better, coverage, it required significantly more words to do so. For example, the BNC/COCA 3,000 to 5,000 frequency bands provide 10.69% coverage of the economics corpus, compared to 11.44% for the IS-AVL, which constitutes a fairly negligible difference. However, it requires 15,039 lemmas to achieve this coverage, compared to 481 for the IS-AVL.

Table 6.6

Coverage Provided by the IS-AVL Compared to the Most Common Word Lists Being Used in EMI Classrooms

Subject	IS - AVL		IS-AVL from Chapter 5		AWL		BNC/COCA 2K	
	Tokens	Cov	Tokens	Cov	Tokens	Cov.	Tokens	Cov.
Literature	610	9.96%	546	6.72%	6370	9.49%	3082	6.88%
Economics	481	11.44%	679	10.76%	6370	**15.36%**	3082	10.69%
Social Studies	526	9.43%	655	8.56%	6370	**12.06%**	3082	7.80%
Biology	845	20.70%	812	15.56%	6370	10.68%	3082	7.62%
Chemistry	681	22.55%	752	17.77%	6370	12.14%	3082	9.46%
Physics	578	17.45%	652	12.15%	6370	11.50%	3082	6.99%
Math	379	13.54%	530	11.77%	6370	8.71%	3082	8.12%
TOK	685	12.81%	636	7.77%	6370	11.38%	3082	8.38%

Subject	IS - AVL		BNC/COCA 3K – 5K		SVL		MSVL	
	Tokens	Cov	Tokens	Cov	Tokens	Cov	Tokens	Cov
Literature	610	9.96%	3082	6.88%	686	11.71%	722	6.03%
Economics	481	11.44%	3082	10.69%	477	**15.04%**	NA	NA
Social Studies	526	9.43%	3082	7.80%	717	**10.53%**	NA	NA
Biology	845	20.70%	3082	7.62%	880	14.69%	858	11.53%
Chemistry	681	22.55%	3082	9.46%	519	14.98%	858	12.97%
Physics	578	17.45%	3082	6.99%	545	14.00%	858	9.51%
Math	379	13.54%	3082	8.12%	253	12.17%	616	9.46%
TOK	685	12.81%	3082	8.38%	NA	NA	NA	NA

Note. Bold text indicates where that word list provides better coverage than the IS-AVL. Abbreviations used in the table are: Cov = Coverage, BNC/COCA 2K = the 2,000-frequency band of the BNC/COCA, BNC/COCA 3K – 5K = the 3,000 to 5,000 frequency bands of the BNC/COCA

It is important to acknowledge that the IS-AVL provides lower coverage over English, Social Studies, and Economics. There are two reasons this may have occurred. First, this reflects the differences in the type of vocabulary between the humanities and the sciences. While the humanities have a richer vocabulary, they tend to use a larger number of words from high-frequency vocabulary lists. The sciences tend to have a more technical and specialized vocabulary. Second, the textbooks used to compile the corpora for the sciences tended to be uniform. That is, most of the textbooks in those domain-specific corpora covered very similar topics and had similarly named chapters. On the other hand, the Economics and Social Studies textbooks were much more diverse. For example, the Social Studies corpus was made up of both history and political science textbooks. A more uniform selection of texts may have improved the coverage for these subjects. However, there are fewer IB textbooks for these subjects than there are for the sciences, which would make it harder to compile a corpus of the required size if we were limited to using books that had the same chapters as each other.

There is also one area where the SVL does provide better coverage of the corpus. This is for the Economics corpus. Again, this is probably due to how this corpus was compiled. The Economics section of the IB program contains both practical and theoretical texts. The practical section is made up of Business Management textbooks, and the theoretical section is made up of Economics textbooks. On the other hand, the SVL is made up exclusively of words from Economics textbooks. This means fewer lemmas in the Economics section would have been removed using the range and dispersion measures. It may have been possible to mitigate this issue by dividing the corpora into two sections and compiling a word list for both sections. However, as discussed above, because of the number of IB textbooks available for this subject, this would have been problematic. Even with this limitation, the coverage provided by this is over 10% for fewer than 500 lemmas. Therefore, despite the shortcomings discussed in this paragraph, I feel that this word list would still be useful for IB teachers and learners who are engaging with this subject.

6.3.3 Coverage over other corpora

The final thing that we need to do before ending this section is to look at what coverage the IS-AVL will give over other corpora. This involves checking the IS-AVL against

three different corpora: a non-academic corpus, the corpora from different academic domains, and a corpus from the same domain that was not used in the process of compiling the word lists.

First, I checked the word lists against a non-academic corpus. As I detailed in Chapter Two section 2.3.2, one of the tests that Coxhead (2000) performed to confirm that the AWL was academic was to look at the coverage that this list provided over a non-academic corpus. For this purpose, I compiled a corpus of 15 different novels from project Gutenberg (http://www.gutenberg.net). The corpus contained just over 3 million running words. I lemmatized this corpus and then ran it against the different domain-specific word lists from the IS-AVL. The results showed that the coverage ranged from 1.35% to 2.68% of the corpus, depending on the word list used. As expected, the Literature corpus provided the best coverage, and the maths corpus provided the worst coverage. Some words, such as *species* and *enzyme,* did not occur at all in the corpus of novels, which is probably because of a combination of the academic nature of the words and the age of the fictional texts. The coverage provided by all the lists was markedly lower than the coverage each list provided over its own domain, suggesting that all the lists are academic.

Table 6.7

Coverage Provided by the IS-AVL over a Corpus of Novels

	Lit	Eco	SS	Bio	Chem	Phys	Maths	TOK
Coverage	2.68%	1.47%	1.84%	2.18%	1.81%	1.85%	1.35%	2.42%

Note. Lit = Literature, Eco = Economics, SS = Social Studies, Bio = Biology, Chem = Chemistry, Phys = Physics, TOK = Theory of Knowledge

Second, I wanted to determine if the word lists were domain specific. Accordingly, I looked at the coverage that each list provided over the other subcorpora compared to its own subcorpora. This involved running each of the word lists against each of the subcorpora and comparing the results. Again, the results were promising, as the word lists performed significantly better on the corpus that they had been compiled

against (see Table 6.8). The one exception was the fact that some of the word lists from the humanities had very high coverage of the Theory of Knowledge (TOK) corpus. This is not surprising as TOK is a course that looks at how knowledge is expressed in the subjects studied across the IB curriculum. This means that we can expect a lot of overlap between the words used in TOK and those used in the other subjects. Together with the previous analysis of the coverage the lists provided over a corpus of fictional tests, the suggestion is that all the lists are indeed academic and domain specific.

Table 6.8

Coverage Provided by the IS-AVL over the Corpora from the Other Domains

Subject	Lit	Eco	SS	Bio	Chem	Phys	Maths	TOK
Literature	**9.96%**	4.84%	5.01%	5.63%	5.20%	4.46%	3.92%	8.71%
Economics	8.60%	**11.44%**	10.10%	9.05%	8.98%	9.10%	7.35%	10.19%
SS	8.12%	6.97%	**9.43%**	6.90%	5.94%	5.02%	3.86%	8.28%
Biology	7.04%	7.37%	6.17%	**20.70%**	12.09%	9.84%	7.48%	10.66%
Chemistry	7.63%	8.89%	6.51%	18.30%	**22.55%**	16.11%	9.14%	11.15%
Physics	6.58%	7.50%	5.38%	13.40%	14.59%	**17.45%**	9.23%	10.05%
Maths	7.02%	6.85%	5.26%	10.31%	11.41%	12.15%	**13.54%**	9.01%
TOK	10.38%	7.25%	6.71%	9.02%	8.08%	6.45%	5.76%	**12.81%**

Note. Lit = Literature, Eco = Economics, SS = Social Studies, Bio = Biology, Chem = Chemistry, Phys = Physics, TOK = Theory of Knowledge. The bold numbers indicate the coverage the list provides over the subject it was compiled from.

Third and finally, I wanted to see if the levels of coverage the word lists provide against the corpus from which they were compiled can be replicated with another corpus in the same domain. The ideal way to do this would be to compile a separate corpus for each domain and then use these corpora to determine what type of coverage the word lists provide. Unfortunately, corpus construction is time-consuming, especially

when there are eight different domains, and not all the domains had additional textbooks that I could use for this process. This issue is not exclusive to this study since other researchers, such as Green and Lambert (2018), could not perform this step either. However, to provide some insight into the type of coverage we would expect to see, I looked at the levels of coverage provided over a different corpus in a single domain: biology. The reason for choosing biology was that it is easier to clean the biology texts than the texts from mathematics and physics. There are also more texts available in this subject than in some of the other subjects such as social studies and economics.

To conduct this analysis, I compiled a parallel corpus that was 87,850 tokens long using an IB Biology textbook that was not included in the original corpus (Primrose, 2019, see Appendix C for the full bibliographical information about this book). An analysis of the coverage the biology word list from the IS-AVL gave over this corpus compared to the AWL showed that the coverage provided by the IS-AVL was much higher than that of the AWL over the parallel corpus, and very close to the coverage provided over the whole corpus (see Table 6.9).

Table 6.9

The Coverage Provided by the Biology IS-AVL over a Parallel Corpus of Biology Textbooks

Corpus	IS-AVL	AWL
Biology Corpus	20.70%	10.68%
Parallel Corpus	19.85%	7.47%

6.4 Conclusion

I can now return to the two questions that I posed at the start of the current chapter:

1. Do these new methods do a better job of identifying academic vocabulary that would be useful for EAL learners than the techniques used by Greene and Coxhead (2015) that were described in the previous chapter?

2. How does the coverage provided by the academic vocabulary lists developed using these techniques compare to the coverage provided by existing word lists over a corpus of international school textbooks?

With the first question, we can clearly see that the lists from the current chapter are superior to the ones reported in Chapter Five. Not only do they provide greater coverage of the corpora, but they also ease some issues that we encountered with the lists in Chapter Five. By using lemmas as the unit of counting, I could consider the part of speech of the words in the list, which will make it easier for teachers to teach these words in the classroom. Also, when compiling these lists, I did not rely on the GSL to eliminate high-frequency words, so important subject-specific words such as *model*, *risk*, and *sample* were not arbitrarily removed from the list.

With the second question, an analysis of the coverage provided by the IS-AVL compared to word lists that are currently being used in the international school and EAL context, I could show that the IS-AVL consistently outperformed these lists. With one exception, the IS-AVL provided better coverage with fewer words over all the IS-CAT subcorpora. Because of this, I believe these lists to be a valuable resource for EAL learners and teachers.

6.4.1 Findings of the study

In the current chapter, I was able to refine the techniques that I used to compile the IS-AVLs. By integrating modern techniques into the process that I used to compile these lists, I could compile a set of word lists that provided better coverage and included a more pedagogically useful method of counting words than my previous attempt at creating a set of word lists. I then looked at the coverage these lists provided compared to other word lists. This showed that the IS-AVL was more representative of the types of words learners need to know for each of the subjects in that it provided better coverage of the domain-specific corpora with fewer lemmas. The only exception to this was with the economics word list, which, while not providing as good a coverage of the economics subcorpora as the SVL, still provided excellent coverage with few lemmas. I subsequently checked the IS-AVL against a non-academic corpus and the subcorpora from the other domains, which showed that the lists were both academic and domain-

specific. Finally, an analysis of the biology word list against a representative corpus that was not used in compiling the word lists showed the lists were valid.

The current chapter brings to a close the second thread of this dissertation, which focuses on whether it is possible to provide a set of word lists that can address the vocabulary gaps that exist in EAL learners' vocabulary knowledge. There are, of course, some issues that need to be addressed before these lists can be used as pedagogical tools in the classroom, including how to go about teaching and assessing the words on these lists. In the conclusion of this dissertation, I address these issues and bring together the two threads we have followed throughout the course of the dissertation, what words EAL learners know and what words they need to know. In this final chapter, I hope to show that I have been effective in identifying the gaps that exist in EAL learners' vocabulary knowledge and have been able to provide a pedagogical tool that can help teachers and learners to address these gaps.

Chapter Seven
Review and Discussion

7.1 Introduction

Over the last four experimental chapters, we have moved from investigating why existing word lists may not be suitable for EAL learners to developing a set of lists to address this important gap. In this final chapter, I review what we have covered in the dissertation and summarize what I have achieved in the dissertation to this point. I then examine some issues that have arisen during the creation of these lists and briefly discuss what still needs to be done. This includes a look at future research that could both improve these lists and develop the pedagogical tools necessary for the lists to be used in the classroom.

7.2 Review

In Chapter One, I began by looking at what we mean when we refer to EAL learners and the challenges these learners may face. I began by looking at how changes in classroom demographics has made research into EAL learners an important educational priority. I also explained why it made sense to focus on vocabulary to support these learners and looked at how EAL learners' struggles with vocabulary knowledge can lead to them struggling academically. Part of this discussion involved emphasizing the fact that, to date, there are few tools available to teachers that would allow them to provide EAL learners with the vocabulary support that they need in the classroom. Although there are several general, academic, and specific word lists available, these do not provide the coverage that EAL learners require from the texts that they are being asked to read in the classroom. I argued that, because of this, to better support EAL learners, it is necessary to develop a set of word lists that are specifically designed for this group of learners. However, to do so, it is first necessary to determine what words EAL learners are likely to know as well as what words they need to know to be successful in the classroom.

Chapter Two introduced a selection of the most important studies that have been carried out over the past few decades in two different areas. First, I looked at studies that focused on how much vocabulary L2 English speakers are likely to need to understand a written text. Second, I looked at how the process for compiling vocabulary

lists has changed over time and highlighted some key word lists that would be useful in the context of this dissertation. In doing so, I discussed how the process of creating a word list from a corpus has evolved from the use of simple measures of frequency and range (e.g., Coxhead, 2000; Greene & Coxhead, 2015) to the use of more sophisticated techniques from the field of corpus linguistics (e.g., Gardner & Davies, 2014; Green & Lambert, 2018; Lei & Liu, 2016). I also discussed why a set of word lists designed specifically for EAL learners is necessary, and what those word lists would look like.

In Chapter Three, I discussed a pilot project that looked at just how important vocabulary is for EAL learners' academic success. Chapter Three showed us that vocabulary is one of the most important predictors of EAL learners' academic success. It also allowed us to look at the types of words EAL learners would be likely to know and those that they would likely struggle with. Finally, we saw that an assessment tool that was developed using the BNC/COCA (McLean & Kramer, 2016) was able to effectively identify EAL learners who would be likely to struggle to read at a level that is appropriate for their age. While this is a promising start, at this point in the dissertation, I did not know what type of coverage the BNC/COCA (P. Nation, 2020) provided over the texts that EAL learners are likely to encounter in the classroom.

In Chapter Four, I investigated this question with a replication of Coxhead and Boutorwick's (2018) study. Coxhead and Boutorwick and my replication examined the vocabulary knowledge of EAL learners and compared this to the vocabulary profiles of the textbooks they are required to read in the classroom. This investigation led me to two important discoveries. First, the vocabulary knowledge of EAL learners studying at international schools in Japan would not be sufficient for them to understand the textbooks that they are required to read in the classroom. This is because, even at the higher grade levels, these learners do not have the vocabulary knowledge needed to achieve the 95% coverage necessary for understanding (e.g., Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010). Second, the BNC/COCA is not able to provide these learners with the support that they need in the classroom. This is because using the BNC/COCA, EAL learners would have to learn too many words to bridge the gap between their existing knowledge and where they need to be. In Chapter Four, I also started compiling the IS-CAT corpus, which I used to develop a set of word list to better support EAL learners in the classroom.

In Chapter Five, I described how I compiled the IS-CAT and discussed the initial attempt I made to construct a set of EAL specific word lists. The method in this chapter drew upon those Coxhead (2000) and Greene and Coxhead (2015) used to construct their word lists. Despite several limitations with these techniques, I could construct a set of eight word lists of between 357 and 554 word families that provide 6.72% to 17.77% coverage of their respective corpora, which is significantly better than the coverage provided by other academic word lists, such as Coxhead's (2000) AWL. However, there are several issues with these lists, including their reliance on word families and the use of the GSL (West, 1953) to remove high-frequency words.

Chapter Six addressed concerns from the earlier experimental chapters by using more modern techniques to identify relevant words in the corpus. Using these more recent techniques, I was able to revise the word lists I created in Chapter Five to be more representative of the different corpora and to provide better coverage with fewer tokens. Most importantly, in Chapter Six, I introduced the final IS-AVL and validated these lists by comparing them to existing word lists and looking at their coverage over different corpora.

7.3 Summary of achievements

The work that I carried out in this thesis has allowed me to create eight domain-specific word lists that can be used independently or together to support EAL learners in the classroom. As I noted above, these lists have the potential to be extremely useful pedagogical tools in the EMI classroom. The most obvious application of these lists is that they will help teachers to identify what words to teach in the classroom. Because they target the academic vocabulary EAL learners are likely to encounter in their textbooks and need to use in their spoken and written responses, they allow teachers to focus on the words that matter for their students. Like Coxhead's (2000) AWL, they also provide a road map for younger learners and can help them start mastering the vocabulary they will need before entering the IB Diploma program.

However, they have the potential to do more than just provide guidance for teachers in selecting what words to teach explicitly. If these lists were to be integrated into services like Lextutor (lextutor.ca) they could help identify which text would be difficult for learners to read. In a similar manner, they could be used to help identify

which words are not worth teaching but would be better to gloss. Teachers could identify highly technical terms, or language that is only used in a certain unit or text and provide their students the support they need to understand these words in context, without explicitly teaching them. With this information, the teacher will be in a much better position to help learners master the vocabulary that really matters.

Finally, with additional research, it would be possible to use these lists to help identify learners who would be likely to struggle in the classroom. Recently, researchers (Schmitt et al., 2020) have called for the creation of more context specific vocabulary assessment tools. Rather than using a one size fits all approach, we should develop vocabulary assessments to measure the vocabulary that is needed in a specific context. For EAL learners, this means measuring their knowledge of the vocabulary that they need to succeed in the classroom. The IS-AVL provides us with a starting point from which we can develop the tools necessary to do just that.

7.4 Limitations of the experimental chapters

While I have been successful in reaching my goals of providing a better understanding of what gaps exist in EAL learners' vocabulary knowledge and creating a set of word lists to help bridge these gaps, there are several potential limitations with the study that require our attention in future studies. These include the composition of the corpora, the generalizability of my findings, and the amount of coverage these lists provide when combined with EAL learners' existing vocabulary knowledge.

7.4.1 Composition and size of the corpus

Some of the most important questions to ask with a study that is based upon the analysis of a corpus are whether that corpus is large enough and representative enough for the findings to be valid. As I outlined in Chapter Five, I did my best to ensure that the corpus was representative of the textbooks that learners are likely to be required to read in the classroom. The corpus was also quite large; the total corpus includes almost 9 million running words and each subcorpus is over, or very close to, 1 million running words. While a larger corpus would, of course, be better, this corpus is more than sufficient in size to create a representative list of mid-frequency and specialized vocabulary (P. Nation, 2016). The one area that I would have liked to have addressed are the concerns regarding the variety of texts in the social studies and economics

subcorpora. Both corpora combine textbooks from different topics within the same subject area. Economics contains both practical, business management, and theoretical, economics, texts and social studies contains both texts that cover history and those that cover politics. It would have been preferable to divide each of these corpora into two separate subcorpora. This would have allowed me to create a set of word lists that would have provided better coverage of each of these subcorpora. However, this was not possible because there were not enough textbooks available for each of these topics. In the future, it may be useful to work with the IB schools who participated in this study to determine whether they can identify other textbooks that are representative of the topics covered in these classes to expand these subcorpora.

7.4.2 Generalizability of the findings

To compile the corpora used in this study, I worked closely with IB schools here in Japan to identify books for inclusion in the corpus. While this allows me to say, with a fair amount of certainty, that the word lists developed in this study will be useful for the Japanese international school context, I would like to see if these lists are also relevant for international schools outside of Japan. I would therefore need to work with schools in several countries to look at what textbooks they are using and to see if they are the same as the textbooks being used in Japan. If they are not, then I could add the textbooks that are in use at those schools to my corpus to make it more representative of international schools around the world, and not just in Japan. Given the uniformity that exists within the IB program, it is likely that the books would be the same as (or at least very similar to) the ones I use in my corpus. However, I would need to investigate further before I could say definitively how well the word lists I developed for this thesis can be generalized to international schools in other countries.

7.4.3 The coverage provided by the IS-AVLs

In this study, I could compare the IS-AVL to existing word lists and look at their coverage over other corpora to validate them. However, one limitation of the more modern techniques for compiling word lists (e.g., Gardner & Davies, 2014; Green & Lambert, 2018) is that there is no easy way to determine what type of coverage these lists would produce when combined with learners' existing vocabulary knowledge. Researchers who used a general word list, such as the GSL (West, 1953), to remove

high-frequency words, can add the coverage provided by their word list to the coverage provided by the general word list. However, as I did not use a high-frequency word list in my study, this is not possible for the IS-AVLs. To determine the coverage these lists provide above the knowledge that learners already have, it would be necessary to test these lists in the classroom, which is something that I hope to do in the future but is beyond the scope of this dissertation. However, given what we know about EAL learners' vocabulary knowledge, it is unlikely that these lists alone would allow them to reach the 95% threshold identified as necessary for understanding. To achieve such coverage, we would need a much larger list, which would be impractical for pedagogical purposes because it would contain too many items. One suggestion to help us address this problem is that it may be best to focus less on the 95% target and focus more on providing as much support as we can. Other researchers have noted (S. Fraser, 2010) that the 95% target is intended as guidelines for learners who are reading for pleasure and that academic reading may be different. It is not unreasonable to expect learners to make use of glossaries and dictionaries when reading academic texts. Also, EAL teachers often provide additional language support in the classroom including pre-teaching unknown words, explaining the ideas in the texts in simpler language, and providing learners with handouts that include simplified explanations of the ideas in the texts (Coxhead & Boutorwick, 2018). Given this additional support, even if we can only obtain a lower level of coverage with these lists, we can still see them as successful as they highlight the most important words that these learners need to know.

7.5 Further research

In the next section I explore some areas which need further examination to improve our understanding of the vocabulary that EAL learners need and improve the usefulness of the lists in the classroom. The potential future research should cover two areas: improving the corpus and the word lists; developing pedagogical tools that ensure teachers can use these lists in the classroom.

7.5.1 Revising the corpus

As discussed above, one of the potential strands of future research would be to expand the IS-CAT; I would do this as a way of achieving three important aims. First, I would like to expand the corpus in a way that allowed me to create more specific subcorpora,

discussed in more detail in section 7.4.1, allowing me to create a set of word lists that would provide better coverage of the Social Studies and Economics subcorpora. Second, I would expand the corpus to consider textbooks used at international schools in other countries, because the focus of the dissertation was international schools in Japan. To expand the number of learners who could benefit from these lists, it would be good if future versions of the word lists were more international. Finally, I will need to revise the corpus in the future as the IB program revises its curriculum. For example, a revised version of the IB science curriculum will be implemented at the end of 2023 and completed by 2025 (*IB curriculum updates and subject briefs*, n.d.). The shift in focus in some subjects to include a greater emphasis on developing the technical skills used in the sciences may result in some changes in the language being used in the textbooks. To keep the corpus current, it will be necessary to revise it in the future as the textbooks that learners are being required to read change.

While there are several changes and revisions that need to be made to the IS-AVL in the future, the tools that I have developed when creating the current word lists greatly simplify this process. The R and Python scripts used for analyzing the IS-CAT can easily be adapted to include new textbooks, as those become available. This makes updating the corpus much less time-consuming than developing the original corpus was.

7.5.2 Developing tools to help teachers use the corpus in the classroom

One of the main goals I had behind developing the IS-AVL was to ensure that teachers can use these lists in the classroom to support their students. While investigating the activities that would be best used to teach this vocabulary to EAL learners in the classroom is beyond the scope of this dissertation, there are a number of resources that exist that can help teachers to do this. Two books mentioned extensively throughout this dissertation (Greene & Coxhead, 2015; P. Nation, 2016) provide a great starting point for developing activities based around word lists. While I will not go into detail about all the activities that these authors recommend, it is worth considering how the IS-AVL can help make these activities more useful for the students. One traditional means of having learners acquire new vocabulary is to keep a learner word book. We expect learners to write the meaning of the words along with important information about those words in a book. They then study the words that they have written in the book and are

assessed on their knowledge of those words. The IS-AVL can improve this process by allowing teachers to focus on discipline-specific words in the subject in which those words are used. A chemistry teacher can not only focus on the words that are important for chemistry, they will also be able to focus on the meaning of the word that is most relevant to that subject (for example, highlighting the use of *solution* as a mixture of two or more substances rather than as a way of solving a problem). Concept maps are also a popular tool for teaching vocabulary (Green & Lambert, 2018), and this process is much easier if learners know what words to focus on for a specific subject, something that using the IS-AVL will enable.

Another important goal that we need to consider when investigating the pedagogical uses of the IS-AVL is the development of assessment tools for EAL learners. Given the importance of vocabulary for understanding and academic success, it makes it an invaluable tool in identifying learners who are likely to struggle in the classroom. One way to do this is to use vocabulary levels tests designed to measure the vocabulary knowledge of EAL learners, similar to what I detail in Chapter Three section 3.2.2. However, currently there are no vocabulary levels tests designed specifically for EAL learners. One reason for this is that there were no previous word lists identified for the vocabulary that EAL learners need to know to be successful. Now that I have been able to identify this vocabulary, it should be possible to develop a set of assessment tools that can measure EAL learners' knowledge of relevant subject-specific vocabulary. The ability to measure their mastery of these words would help to identify which learners need support in the classroom and answer recent calls for the development of such tools (Schmitt et al., 2020).

7.6 Conclusion

Throughout this thesis, I have pursued two different, yet connected, goals: to identify the gaps that exist in EAL learners' vocabulary knowledge, and to provide teachers with the tools that they need to bridge those gaps. Despite the limitations with the IS-AVL that I have outlined above, I feel that I have been able to meet these goals and have started along the path of being able to provide teachers with the support that they require. I have laid the groundwork for the creation of assessment tools and other

pedagogical materials that can support EAL learners in the classroom. I have also developed a methodology and a set of tools that can expand these lists in the future.

One of the most important things that I was able to detail in this dissertation is the importance of vocabulary knowledge for EAL learners. Considering this importance and considering the gaps that I have identified in EAL learners' vocabulary knowledge, I feel it is essential for teachers to provide vocabulary support to EAL learners in the classroom. Teaching EAL learners the vocabulary they need to succeed academically is something that can be done both inside the classroom and through pullout programs. In this context, pullout programs are programs that take individual students, or groups of students, out of their regular classes to give them additional language support. While helpful for some learners, these programs are controversial because they introduce learners to the language in an isolated way and can cause learners to fall behind in their regular classes. EAL and FLE learners alike should be encouraged to expand the depth and breadth of their vocabulary knowledge, as both have been shown to be important factors for reading comprehension in both this and other studies (e.g., Treffers-Daller & Huang, 2020). This greater focus on vocabulary in the classroom is essential if EAL learners are to comprehend the texts that they are being asked to read for class. It is important to stress that vocabulary must be looked at in the context of the discipline that it is being used in. It would be prudent for teachers to incorporate tasks into the class that help students to improve their vocabulary knowledge, rather than relying on pullout programs. However, for either pullout programs, or integrating vocabulary instruction into the classroom, to be effective, teachers need to know what words to focus on.

In this dissertation, I have responded to this need by developing the International School Academic Vocabulary Lists (IS-AVL). These lists include discipline-specific lists of lemmas for eight subjects: Literature, Social Studies, Economics, Biology, Chemistry, Mathematics, and TOK. Teachers can use these word lists to supplement existing word lists, such as the new General Service List (Brezina & Gablasova, 2015), the AWL (Coxhead, 2000), and the AVL (Gardner & Davies, 2014). Together, this should allow teachers to identify the vocabulary that EAL learners need to succeed in the classroom. These lists also work with existing middle school (Greene & Coxhead, 2015) and secondary school (Green & Lambert, 2018) word lists to cover most of the learning environments EAL learners can expect to find themselves in. I hope that as the

tools available to teachers to support their students in the classroom increase, those students will find it easier to succeed academically. I also hope that the findings and tools discussed in this dissertation can contribute to this progress.

References

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing*, *19*(9), 933–958. https://doi.org/10.1007/s11145-006-9024-z

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Afitska, O., & Heaton, T. J. (2019). Mitigating the effect of language in the assessment of science: A study of English‐language learners in primary classrooms in the United Kingdom. *Science Education*, *103*(6), 1396–1422. https://doi.org/10.1002/sce.21545

Ardasheva, Y., & Tretter, T. R. (2017). Developing science-specific, technical vocabulary of high school newcomer English learners. *International Journal of Bilingual Education and Bilingualism*, *20*(3), 252–271. https://doi.org/10.1080/13670050.2015.1042356

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice: A Publication of the Division for Learning Disabilities, Council for Exceptional Children*, *20*(1), 50–57. https://doi.org/10.1111/j.15405826.2005.00120.x

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle. *System*, *29*(2), 209–219. https://doi.org/10.1016/s0346251x(01)00012-4

Barber, C. L. (1962). Some measurable characteristics of modern scientific prose. *Contributions to English Syntax And*.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland's *D* to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, *21*(4), 439–464. https://doi.org/10.1075/ijcl.21.4.01bib

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

Brooks, G., Clenton, J., & Fraser, S. (2021). Exploring the importance of vocabulary for English as an additional language learners' reading comprehension. *Studies in Second Language Learning*, *11*(3), 351–376. https://doi.org/10.14746/ssllt.2021.11.3.3

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*. https://doi.org/10.1093/applin/amaa061

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, *3*(2), 1–10. http://www.vli-journal.org/issues/03.2/issue03.2.full.pdf#page=5

Browne, C., Culligan, B., & Phillips, J. (2013). The new academic word list. *Retrieved Fromwww. Newacademicwordlist. Org*.

Bruce, P. C. (2015). *Introductory statistics and analytics*. John Wiley & Sons.

Brysbaert, M., & New. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Burch, B., Egbert, J., & Biber, D. (2017). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(2), 189–216. https://doi.org/10.1558/jrds.33066

Burgoyne, K., Kelly, J. M., Whiteley, H. E., & Spooner, A. (2009). The comprehension skills of children learning English as an additional language. *The British Journal of Educational Psychology*, *79*(4), 735–747. https://doi.org/10.1348/000709909X422530

Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list*. New Zealand Council for Educational Research.

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

Clegg, J., & Afitska, O. (2011). Teaching and learning in two languages in African classrooms. *Comparative Education Review*, *47*(1), 61–77. https://doi.org/10.1080/03050068.2011.541677

Clenton, J., de Jong, N. H., Clingwall, D., & Fraser, S. (2020). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency for a small group of pre-intermediate Japanese L1 users of English (L2). In C. J. B. Paul (Ed.), *Vocabulary and the four skills* (pp. 126–145). Routledge. https://doi.org/10.4324/9780429285400-15

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, *33*(2), 185. https://doi.org/10.2307/3587717

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

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, *45*(2), 355–362. https://doi.org/10.5054/tq.2011.254528

Coxhead, A. (2012). Researching vocabulary in secondary school English texts: 'The hunger games' and more. *English in Aotearoa*, *78*, 34–41. https://doi.org/10.3316/informit.984965735419059

Coxhead, A. (2017). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. Routledge. https://doi.org/10.4324/9781315146478

Coxhead, A. (2019). Analysis of corpora. In *The Routledge handbook of research methods in applied linguistics* (pp. 464–473). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780367824471-39/analysis-corpora-averil-coxhead

Coxhead, A., & Boutorwick, T. J. (2018). Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science. *TESOL Quarterly*, *52*(3), 588–610. https://doi.org/10.1002/tesq.450

Coxhead, A., & Hirsch, D. (2007). A pilot science-specific word list. *Revue française de linguistique appliquée*, *XII*(2), 65. https://doi.org/10.3917/rfla.122.0065

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*(1), 121–135. https://doi.org/10.1007/s40841-015-0002-3

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read. *New Zealand Studies in Applied Linguistics*, *16*(2), 37–52.

Coxhead, A., & White, R. (2012). Building a corpus of secondary school texts: First you have to catch the rabbit. *New Zealand Studies in Applied Linguistics*, *18*(2), 67–73. https://search.informit.org/doi/abs/10.3316/informit.108520980622633

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy* (Vol. 6). Multilingual Matters Clevedon.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. *Encyclopedia of Language and Education*, *2*(2), 71–83.

Cummins, J., & Yee-Fun, E. M. (2007). Academic language. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 797–810). Springer US. https://doi.org/10.1007/978-0-387-46301-8_53

Daller, M., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University Press.

Daller, M., & Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, *4*(1), 173–193. https://doi.org/10.1515/applirev-2013-0008

Dang, T. N. Y. (2021). Selecting lexical units in wordlists for EFL learners. *Studies in Second Language Acquisition*, *43*(5), 954–957. https://doi.org/10.1017/S0272263121000681

Dang, T. N. Y., Coxhead, A., & Webb, S. A. (2017). The Academic Spoken Word List. *Language Learning*, *67*(4), 959–997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y., & Webb, S. A. (2016). Making an essential word list for beginners. In P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167). John Benjamins Publishing Company.

Davies, M. (2002). *The corpus of contemporary American English*. Brigham Young University.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/s0142716413000210

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first‐ and second‐language learners. *Reading Research Quarterly*, *38*(1), 78–103. https://doi.org/10.1598/rrq.38.1.4

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes*, *43*, 49–61. https://doi.org/10.1016/j.esp.2016.01.004

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.

Francis, W. N., Kučera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin.

Fraser, C. A. (2007). Reading rate in L1 mandarin Chinese and L2 English across five reading tasks. *Modern Language Journal*, *91*(3), 372–394. https://doi.org/10.1111/j.1540-4781.2007.00587.x

Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education*, *10*, 127–143. https://cir.nii.ac.jp/crid/1390009224855229824

Fraser, S. (2010). *The lexis of pharmacology texts: A corpus linguistic analysis* [Doctoral dissertation]. Swansea University.

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the

strength of the relationship in English. *Review of Educational Research*, *84*(1), 74–111. https://doi.org/10.3102/0034654313499616

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241–265. https://doi.org/10.1093/applin/amm010

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. https://doi.org/10.1093/applin/amt015

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, *25*(8), 1819–1845. https://doi.org/10.1007/s11145-011-9333-8

Geva, E., & Zadeh, Z. Y. (2006). Reading efficiency in native English-speaking and English-as-a-second-language children: The role of oral proficiency and underlying cognitive-linguistic processes. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, *10*(1), 31–57. https://doi.org/10.1207/s1532799xssr1001_3

Ghadessy, P. (1979). Frequency counts, word lists, and materials preparation: A new approach. *English Teaching Forum*, *17*, 24–27.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341–363. https://doi.org/10.1093/applin/11.4.341

Grabe, W. (2010). Fluency in reading—Thirty-five years later. *Reading in a Foreign Language*, *22*(1), 71–83.

Graves, M. F., August, D., & Mancilla-Martinez, J. (2012). *Teaching vocabulary to English language learners*. Teachers College Press.

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families

for eight secondary subjects. *Journal of English for Academic Purposes*, *35*, 105–115. https://doi.org/10.1016/j.jeap.2018.07.004

Green, C., & Lambert, J. (2019). Position vectors, homologous chromosomes and gamma rays: Promoting disciplinary literacy through Secondary Phrase Lists. *English for Specific Purposes*, *53*, 1–12. https://doi.org/10.1016/j.esp.2018.08.004

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

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. T. (2020). Analyzing dispersion. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 99–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_5

Gries, S. T. (2022). Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics*, *1*(1), 1–19. https://doi.org/10.1016/j.rmal.2021.100002

Harmon, J. M., Hedrick, W. B., & Fox, E. A. (2000). A Content Analysis of Vocabulary Instruction in Social Studies Textbooks for Grades 4-8. *The Elementary School Journal*, *100*(3), 253–271. https://doi.org/10.1086/499642

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*(2), 152–169. https://doi.org/10.1080/15434303.2014.902059

Hawkins, E. (2005). Out of this nettle, drop-out, we pluck this flower, opportunity: re-thinking the school foreign language apprenticeship. *The Language Learning Journal*, *32*(1), 4–17. https://doi.org/10.1080/09571730585200141

Heatley, A., Nation, P., & Coxhead, A. (2004). *The Range Program*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

Henriksen, B., Albrechtsen, D., & Haastrup, K. (2004). The relationship between vocabulary size and reading comprehension in the L2. *Angles on the English-Speaking World*, *4*(1), 129–140.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–696. https://openaccess.wgtn.ac.nz/articles/journal_contribution/What_vocabulary_size_is_needed_toread_unsimplified_texts_for_pleasure_/12560417/files/23412206.pdf

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zendo. https://doi.org/10.5281/zenodo.1212303

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127–160. https://doi.org/10.1007/BF00401799

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

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430. https://doi.org/10.26686/wgtn.12560354.v1

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press. https://doi.org/10.1017/cbo9781139524773

Hutchinson, J. M., Whiteley, H. E., Smith, C. D., & Connors, L. (2003). The developmental progression of comprehension‐related skills in children learning EAL. *Journal of Research in Reading*, *26*(1), 19‐32. https://doi.org/10.1111/1467-9817.261003

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

IB curriculum updates and subject briefs. (n.d.). International Baccalaureate. Retrieved 19 November 2022, from https://www.ibo.org/university-admission/latest-curriculum-updates/ (Original work published 2022)

Ishihara, K., Hiser, E., & Okada, T. (2003). Modifying C-Test for practical purposes. *Doshisha Studies in Language and Culture*, *5*(4), 539–568. https://doi.org/10.14988/pa.2017.0000004420

International Baccalaureate Facts and Figures. (2022, November). International Baccalaureate. https://www.ibo.org/about-the-ib/facts-and-figures/

Jiang, X., Sawaki, Y., & Sabatini, J. (2012). Word reading efficiency, text reading fluency, and reading comprehension among Chinese learners of English. *Reading Psychology*, *33*(4), 323–349. https://doi.org/10.1080/02702711.2010.526051

Juilland, A., Brodin, D., & Davidovitch, C. (1970). *Frequency dictionary of french words*. Walter de Gruyter.

Juilland, A., & Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words*. Mouton & Co.

Konstantakis, N. (2007). Creating a business word list for teaching business English. *Elia*, *7*, 79–102. https://idus.us.es/handle/11441/34157

Kurnia, N. (2003). *Retention of multi-word strings and meaning derivation from L2 reading* [Doctoral dissertation]. Victoria University of Wellington.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. In C. Lauren & M. Nordman (Eds.), *From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B. (1992). Reading in a foreign language: how does L2 lexical knowledge interact with the reader's general academic ability. *Journal of Research in Reading*, *15*(2), 95–103. https://doi.org/10.1111/j.1467-9817.1992.tb00025.x

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

Leech, G., Rayson, P., & Wilson, A. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42–53. https://doi.org/10.1016/j.jeap.2016.01.008

Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *51*(5), 612–620. https://doi.org/10.1111/j.1469-7610.2009.02185.x

Leung, C. (2014). Researching language and communication in schooling. *Linguistics and Education*, *26*, 136–144. https://doi.org/10.1016/j.linged.2014.01.005

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, *39*, 1–11. https://doi.org/10.1016/j.esp.2015.03.001

Long, M. H., & Richards, J. C. (2007). Series editors' preface. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. xii–xiii). Cambridge University Press.

Lynn, R. W. (1973). Preparing word-lists: A suggested method. *RELC Journal*, *4*(1), 25–28. https://doi.org/10.1177/003368827300400103

Malvern, D. D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan. https://doi.org/10.1057/9780230511804

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

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183–198. https://doi.org/10.1016/j.esp.2009.04.003

McLean, S., & Kramer, B. (2016). The creation of a new vocabulary levels test. *Shiken*, *19*(1), 1–9. http://teval.jalt.org/node/33

Meara, P. (1996). The dimensions of lexical competence. *Performance and Competence in Second Language Acquisition*, *35*, 33–55.

Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*(2), 409–433. https://doi.org/10.1037/a0033890

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychological Bulletin*, *138*(2), 322–352. https://doi.org/10.1037/a0026744

Merriam-Webster's medical English dictionary, New Edition. (2006). Merriam-Webster, Inc.

Miller, J. (2009). Teaching refugee learners with interrupted education in science: Vocabulary, literacy and pedagogy. *International Journal of Science Education*, *31*(4), 571–592. https://doi.org/10.1080/09500690701744611

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4*(1), 151–172. https://doi.org/10.1515/applirev-2013-0007

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org/10.1515/iral-2017-0016

Movick, J. (1977, May 30). Latos: The proficient obstacle. *Salient*, *40*(12). http://nzetc.victoria.ac.nz/tm/scholarly/tei-Salient40121977-t1-body-d14-d1.html

Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford University Press.

Murphy, V. A., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL)*. Educational Endowment Foundation. http://www.naldic.org.uk/Resources/NALDIC/Research%20and%20Information/Documents/eal-systematic-review-prof-v-murphy.pdf

Myint Maw, T. M., Clenton, J., & Higginbotham, G. (2022). Investigating whether a flemma count is a more distinctive measurement of lexical diversity. *Assessing Writing*, *53*, 100640. https://doi.org/10.1016/j.asw.2022.100640

Nagy, W. E. (2007). Metalinguistic awareness and the vocabulary-comprehension connection. In R. K. Wagner, A. Muse, & K. Tannenbaum (Eds.), *Vocabulary acquisition and its implications for reading comprehension* (pp. 52–77). Guildford.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English. *Reading Research Quarterly*, *19*(3), 304–330. https://doi.org/10.2307/747823

Nagy, W. E., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108. https://doi.org/10.1002/RRQ.011

NALDIC. (2015, October 27). *EAL Achievement: The latest information on how well EAL learners do in standardised assessments compared to all students*. National Association for Language Development in the Curriculum. https://www.naldic.org.uk/research-and-information/eal-statistics/ealachievement/

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: broader language skills contribute to the development of reading. *Journal of Research in Reading*, *27*(4), 342–356. https://doi.org/10.1111/j.1467-9817.2004.00238.x

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, P. (2005). *Teaching and learning vocabulary*. Taylor Francis. https://doi.org/10.4324/9781410612700-44/teaching-learning-vocabulary-nation

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

Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208

Nation, P. (2020). *The BNC/COCA word family lists* [PDF file]. https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf

Nation, P. (2021). Thoughts on word families. *Studies in Second Language Acquisition*, *43*(5), 969–972. https://doi.org/10.1017/S027226312100067X

Nation, P., & Sorell, J. (2016). Corpus selection and design. In P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 95–105). John Benjamins Publishing Company.

Nation, P., & Waring, R. (2019). *Teaching extensive reading in another language*. Routledge. https://doi.org/10.4324/9780367809256

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

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* [NIH Pub. No. 00-4769]. National Institutes of Health, National Institute of Child Health and Human Development. https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

Neufeld, S., & Billuroğlu, A. (2005). *In search of the critical lexical mass: How 'general' is the GSL? How 'academic' is the AWL?* https://www.researchgate.net/publication/317661877_In_search_of_the_critical_lexical_mass_How_'general'in_the_GSL_How_'academic'is_the_AWL

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System*, *39*(4), 533–538. https://doi.org/10.1016/j.system.2011.10.010

Neff, P. (2015). Peer review use in the EFL writing classroom [Doctoral dissertation, Temple University, Japan Campus]. Temple University Electronic Theses and Dissertations. https://digital.library.temple.edu/digital/collection/p245801coll10/id/329588/rec/1

Oakes, M. P., & Farrow, M. (2007). Use of the Chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, *22*(1), 85–99. https://doi.org/10.1093/llc/fql044

Ogle, D., Blachowicz, C., Fisher, P., & Lang, L. (2015). *Academic vocabulary in middle and high school: Effective practices across the disciplines*. Guilford Publications.

Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*(2), 189–208. https://doi.org/10.1007/s11145-008-9159-1

Pinchbeck, G. G., Brown, D., McLean, S., & Kramer, B. (2022). Validating word lists that represent learner knowledge in EFL contexts: The impact of the definition of word and the choice of source corpora. *System*, *106*, 102771. https://doi.org/10.1016/j.system.2022.102771

Praninskas, J. (1972). *American university word list*. Longman.

Pressley, M., & Allington, R. L. (2014). *Reading instruction that works: The case for balanced teaching* (4th ed.). Guilford Publications.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

Qian, D. D., & Lin, L. H. F. (2019). The relationship between vocabulary knowledge and language proficiency. In S. A. Webb (Ed.), *The Routledge handbook of vocabulary*. taylorfrancis.com. https://doi.org/10.4324/9780429291586-5/relationship-vocabulary-knowledge-language-proficiency-david-qian-linda-lin

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* [Doctoral dissertation, Lancaster University]. https://ucrel.lancs.ac.uk/people/paul/publications/phd2003.pdf

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12–25. https://doi.org/10.1177/003368828801900202

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589–619. https://doi.org/10.1017/S0272263199004039

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

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

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

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

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press. https://doi.org/10.1017/9781108569057

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1191/026553201668475857

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145. https://doi.org/10.2307/3588328

Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language-specific knowledge in native and foreign language reading comprehension: An empirical study among dutch students in grades 6, 8 and 10. *Language Learning*, *48*(1), 71–106. https://doi.org/10.1111/1467-9922.00033

Sharples, R. (2021). *Teaching EAL: Evidence-based strategies for the classroom and school*. Multilingual Matters. https://doi.org/10.21832/9781788924443

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software*, *1*(3), 37. https://doi.org/10.21105/joss.00037

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (1st ed.). O'Reilly Media.

Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., Nation, K., & Hulme, C. (2009). *YARC York Assessment of Reading for Comprehension Passage Reading*. GL Publishers.

Sorell, C. J. (2013). *A study of issues and techniques for creating core vocabulary lists for English as an international language* [Doctoral dissertation, Victoria University of Wellington]. http://researcharchive.vuw.ac.nz/handle/10063/3016

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Strand, S., Malmberg, L., & Hall, J. (2015). *English as an Additional Language (EAL) and educational achievement in England: An analysis of the National Pupil Database*. Educational Endowment Foundation. https://doi.org/10871/23323

Taber's Cyclopedic medical dictionary (22nd ed.). (2013). F.A. Davis Company.

Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. Teachers College Press.

Trakulphadetkrai, N. V., Courtney, L., Clenton, J., Treffers-Daller, J., & Tsakalaki, A. (2020). The contribution of general language ability, reading comprehension and working memory to mathematics achievement among children with English as

additional language (EAL): an exploratory study. *International Journal of Bilingual Education and Bilingualism*, *23*(4), 473–487. https://doi.org/10.1080/13670050.2017.1373742

Treffers-Daller, J., & Huang, J. (2020). Measuring reading and vocabulary with the Test for English Majors Band 4: A concurrent validity study. In J. Clenton & P. Booth (Eds.), *Vocabulary and the four skills: Current issues future concerns*. Taylor & Francis. https://doi.org/10.4324/9780429285400-11

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities*, *45*(5), 453–466. https://doi.org/10.1177/0022219411432685

US Department of Education, National Center for Education Statistics. (2022, May). *English learners in public schools*. https://nces.ed.gov/programs/coe/indicator/cgf/english-learners

Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*(1), 33–41. https://doi.org/10.1016/j.esp.2008.08.003

Wang, J., Liang, S.-L., & Ge, G.-C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442–458. https://doi.org/10.1016/j.esp.2008.05.003

Warrell, D. A., Cox, T. M., & Firth, J. D. (Eds.). (2010). *Oxford Textbook of medicine* (5th ed.). Oxford University Press. https://doi.org/10.1093/med/9780199204854.001.1

Webb, S. A. (2021). Word families and lemmas, not a real dilemma: Investigating lexical units. *Studies in Second Language Acquisition*, *43*(5), 973–984. https://doi.org/10.1017/S0272263121000760

Webb, S. A., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL - International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *The Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wright, D. B., London, K., & Field, A. P. (2011). Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology*, *2*(2), 252–270. https://doi.org/10.5127/jep.013611

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics*, *27*(1), 103–129. https://doi.org/10.1093/applin/ami045

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Yuill, N., & Oakhill, J. (1991). *Children's problems in text comprehension: An experimental investigation*. Cambridge University Press.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior*, *27*(2), 201–212. https://doi.org/10.1080/10862969509547878

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin.

Appendix A: Student Consent Forms

Information and Consent Form

Name of Project: Bridging the vocabulary gap for English as an Additional Language learners: Establishing an EAL word list

Your child has been invited to join a research study to look at the relationship between vocabulary and general academic performance. The purpose of the study is look at how students' vocabulary knowledge affects their linguistic proficiency and their general academic performance and how their vocabulary grows over time. This area of research is very important because recent studies have shown that a lack of vocabulary proficiency is often one of the main factors leading to the lower levels of academic achievement. This proposed research aims to address this issue by investigating these two research questions.

3. What is the vocabulary knowledge of English language learners and what growth can be seen in this knowledge?
4. How does vocabulary, in conjunction with other factors, influence English language learners' general academic abilities?

The study is being conducted by Gavin Brooks, an Assistant Professor at Doshisha University's faculty of Global Communications, and Jon Clenton, an Associate Professor at Hirsohima University's School of Integrated Arts and Sciences. If you have any questions about the research you can contact us at: gavinbrooks@gmail.com or at (81) 080 3855 2381. This research is partially supported by JSPS Grants-in-Aid for Scientific Research (C), Project Number 17K03035.

If you child participates in this research they will be asked to participate in three separate assessments. The first is a vocabulary levels test which is a multiple-choice test designed to measure students' vocabulary knowledge. The second is the York Assessment of Reading for Comprehension (YARC), which is a test designed to measure students' reading comprehension level. The final assessment is the C-test, which is a passage with certain words removed that is designed to measure general academic proficiency. We think that these three assessments will take him/her a total of about one hour to complete and the assessments will be administered over the course of two days.

Any information or personal details gathered in the course of the study will be kept confidential. No individual will be identified in any publication of the results. Only the researchers will have access to the data and when the data is entered into the computer it will not include your child's

name or any information that could be used to identify your child. A summary of the results of the data can be made available to you on request. Please feel free to email the researchers at gavinbrooks@gmail.com at any time and we can send you a copy of the research data or meet with you to give you a copy.

Participation in this study is entirely voluntary: your child is not obliged to participate and if you decide to allow your child to participate, your child is free to withdraw at any time without having to give a reason and without consequence. You are also free to withdraw you child from the study at any time without having to give a reason and without consequence. If your child stops participating in the study he/she will not lose any benefits.

The investigators may stop the study or take your child out of the study at any time they judge it is in your child's best interest. They may also remove your child from the study for various other reasons. They can do this without your consent.

同意文書

プロジェクト名: Bridging the vocabulary gap for English as an Additional Language learners: Establishing an EAL word list

あなたのお子様に、語彙知識と全般的な学力の関係を探る調査にご参加いただきたくご案内申し上げます。この研究は、JSPS 基盤研究(C)17K03035 の助成を受けたものです。目的は、語彙知識の変化が、全般的な学力および言語能力に及ぼす影響の程度を測定することです。この研究の動機は、語彙知識が学業成績に関係していると示す最近のいくつかの研究に端を発しています。下記 2 つが、我々の研究課題です。

5. 英語学習者の語彙知識はどの程度なのか、そしてこの知識にはどのような成長が見られるのか?
6. 語彙は、他の要素と共に、英語学習者の全般的な学力にどのように影響するのか？

この研究は、同志社大学グローバル・コミュニケーション学部助教のギャビン・ブルックスと広島大学総合科学部准教授ジョン・クレントン博士によって行われています。ご不明な点がございましたら、 電子メール gavinbrooks@gmail.com、 または電話 (+81) 080-3855-2381 までご連絡ください。

この調査にご参加いただくお子様方には、3 つの独立した評価課題に答えていただきます。1 つ目は、多肢選択式の語彙知識テストです。2 つ目は、読解力のヨークアセスメント(YARC)です。3 つ目は C テストで、全般的な学習能力を測定するために、内容の一部が削除された文章を用います。この 3 つの課題は全て、他の研究でも広く使われています。これらの課題は完了するまでに約 1 時間かかり、2 日間にわたって行う予定です。お子様には、語彙の伸びを測るために、学年中に 2 回に分けて、これらの評価活動にご参加いただくようお願いする場合があります。

この研究で集められた情報や個人情報は機密情報として管理いたします。結果の公表では、いかなる場合も個人は特定されません。本研究に関わる研究者だけがデータへのアクセス権限を有し、データがコンピューターに入力される際には、個人情報は一切含まれません。ご要望いただければ、データの結果をまとめた物はお渡し可能で

す。いつでもお気軽に、研究者(gavinbrooks@gmail.com)に電子メールをお送りください。

この調査への参加は完全に自由意思によるものです。お子様の参加を強制するものではありません。また、保護者の方がお子様の参加を認めた場合でも、お子様は何ら理由をお伝えいただく必要なく、また何の影響もなく参加をいつでも撤回できます。また、保護者の方も、何ら理由をお伝えいただく必要なく、また何の影響もなく、いつでも、お子様のこの調査への参加を取りやめる事ができます。お子様がこの調査への参加を取りやめても、なんら不利益を被る事はありません。

調査者は、お子様の最善の利益になると判断した場合、いかなる時でも、本調査を中止したり、他の様々な理由でお子様に調査から出ていただく場合もあります。調査者は、保護者の同意なしにこの事を行う事ができるものとします。

Information and Consent Form

We will be assessing the vocabulary levels of the students at Hiroshima International School (HIS) on December 15th, 2017. We would like your permission to use the data from your child's vocabulary level assessment in our study. This study will be conducted in conjunction with HIS and is intended to help improve the English language materials available to students at HIS as well other schools in Japan, with a focus on helping those for whom English is a second language. Participation in this study is anonymous and voluntary. Only the data from the assessment will be used, there will be no information in the study that could be used to identify your child. Either you or your child can request to withdraw from the study at any time, without penalty. This research is supported by JSPS Grants-in-Aid for Scientific Research (C), Project Number 17K03035.

私共は、2017 年 12 月 15 日に広島インターナショナルスクール（HIS）におきまして、生徒の語彙レベル評価の実施を予定しております。つきましては、私共の研究における、あなた様のお子様の語彙レベル評価から得られるデータの使用許可を頂きたいと存じます。本研究は HIS と連携して実施され、HIS や日本の他校の生徒が利用できる英語資料を改善することを目的とし、英語を第二言語にしている方々を支援することに焦点を当てています。本研究への参加は、匿名かつ任意です。本評価から得られるデータのみが使用され、あなた様のお子様の特定することに使用可能な研究情報は含まれません。あなた様又はあなた様のお子様は、ペナルティなしで、いつでも研究からの離脱を要求することができます。本調査は、科学研究費助成事業（KAKEN）認可番号 17K03035 によって支援されております。

Please indicate below if you are willing to have the data from your child's vocabulary assessment used in this research.

本研究におけるお子様の語彙評価のデータ使用の許可の有無を以下にご表示下さい。

許可する / Yes ☐　　拒否する / No ☐

保護者の氏名 ／ Parent or Guardian's Name:_____
保護者の署名 ／ Parent or Guardian's Signature:_____

日付 ／ Date: _____

研究者の名前 ／ Investigator's Name:_____

研究者の署名 ／ Investigator's Signature: _____

日付 ／ Date: _____

If you have any questions please email: gavinbrooks@gmail.com; or telephone: (81) 080 3855 2381.

Appendix B: Assessment Tools

Appendix B.1: Student Questionnaire

If you agree to participate in research, please check the "I agree to participate" box at the bottom of this form. This consent can be changed at any time. (I agree to participate / I don't agree to participate)

研究に参加することに同意する場合、この書面の最下部にある「参加に同意します」のボックスにチェックしてください。この同意はいつでも変更することができます。（参加に同意します。/ 参加に同意しません。）

Name (first name, family name): e.g. Ichiro Suzuki

Gender: (Male/ Female / I would rather not say)

Date of Birth: (Day, Month, Year) e.g. 12th January, 1994

Place of Birth: (City, Country) e.g. Hiroshima, Japan or London, England

Current school grade: e.g. Grade 8

Number of years, months resident in Japan: e.g. 4 years 2 months:

Number of years, months resident outside of Japan, when, and where: e.g. 3 years in the USA when I was 5 and 2 years 1 month in the UK when I was 12 etc

Languages spoken:

Languages spoken at home:

Languages spoken at home: e.g. 50% Japanese, 50% English

Language proficiency self-evaluation:

Please list the languages you can speak below. For each of the languages write how well you can speak, read, listen to and write that language on a scale of 1 to 5.

1 = Beginner; 2 = Okay; 3 = Good; 4 = Very good; 5 = Fluent.

Appendix B.2: The new Vocabulary Levels Test (McLean & Kramer, 2016)

NVLT Part 1: 1000-word band

1. time: They have a lot of **time**.
a. money
b. food
c. hours
d. friends

2. stone: She sat on a **stone**.
a. hard thing
b. kind of chair
c. soft thing of the floor
d. part of a tree

3. poor: We are **poor**.
a. have no money
b. happy
c. very interested
d. tall

4. drive: She **drives** fast.
a. swims
b. learns
c. throws balls
d. uses a car

5. jump: She tried to **jump**.
a. lie on top of the water
b. get up off the ground
c. stop the car on the road
d. move very fast

6. shoe: Where is your other **shoe**?
a. the person who looks after you
b. the thing you keep your money in
c. the thing you use for writing
d. the thing you wear on your foot

7. test: We have a **test** in the morning.
a. meeting
b. travelling somewhere
c. a set of questions
d. an idea to do something

8. nothing: He said **nothing** to me.
a. very bad things
b. zero
c. very good things
d. something

9.cross: Don't **cross**.
a. go to the other side
b. push something
c. eat too fast
d. wait for something

10. actual: The **actual** one is larger.
a. real
b. old
c. round
d. other

11. any: Does she have **any** friends?
a. some
b. no
c. good
d. old

12. far: You have walked **far**!
a. for a long time
b. very fast
c. a long way
d. to your house

13. game: I like this **game**.
a. food
b. story
c. group of people
d. way of playing

14. cause: He **caused** the problem.
a. made
b. fixed
c. explained
d. understood

15. many: I have **many**.
a. none
b. enough
c. a few
d. a lot

16. where: **Where** did you go?
a. at what time
b. for what reason
c. to what place
d. in what way

17. school: This is a big **school**.
a. where money is kept
b. sea animal
c. place for learning
d. where people live

18. grow: All the children **grew**.
a. drew pictures
b. spoke
c. became bigger
d. cried a lot

19. flower: He gave me a **flower**.
a. night clothes
b. small clock
c. beautiful plant
d. type of food

20. handle: I can't **handle** it.
a. open
b. remember
c. deal with
d. believe

21. camp: He is in the **camp**.
a. sea
b. hospital
c. building where people sleep
d. place outside where people enjoy nature

22. lake: People like the **lake**.
a. area of water
b. very young child
c. leader
d. quiet place

23. past: It happened in the **past**.
a. before now
b. big surprise
c. night
d. summer

24. round: It is **round**.
a. friendly
b. very big
c. very quick
d. with no corners

NVLT Part 2: 2000-word band

1. maintain: Can they **maintain** it?
a. keep it like it is
b. make it larger
c. get a better one than it
d. get it

2. period: It was a difficult **period**.
a. small set of questions
b. time
c. thing to do
d. book

3. standard: Her **standards** are very high.
a. the back under her shoes
b. test scores
c. cost of something
d. level of how good she wants things to be

4. basis: This was used as the **basis**.
a. answer
b. resting place
c. next step
d. main part

5. upset: I am **upset**.
a. strong
b. famous
c. rich
d. angry

6. drawer: The **drawer** was empty.
a. place to keep cars
b. place used to keep things cold
c. animal house
d. box that goes in and out for clothes

7. pub: They went to the **pub**.
a. place where people drink and talk
b. place that keeps money
c. large building with many shops
d. building for swimming

8. circle: Make a **circle**.
a. rough picture
b. space with nothing in it
c. round shape
d. large hole

9. pro: He's a **pro**.
a. person whose job is to find secrets
b. stupid person
c. person who writes articles
d. someone who is very good at doing something and is paid to do it.

10. soldier: He is a **soldier**.
a. person who works in business
b. person who studies at school
c. person who works with wood
d. person who fights in a war

11. result: They were waiting for the **results**.
a. right time
b. questions
c. money
d. effects of something

12. resist: They **resisted** it.
a. made it work again
b. looked at it twice
c. thought hard about
d. acted against

13. lend: She often **lends** her books.
a. lets people use them
b. draws inside them
c. cleans them
d. writes her name on them

14. refuse: She **refused**.
a. went back
b. thought about something
c. said no
d. stayed late

15. speech: I enjoyed the **speech**.
a. type of presentation
b. very fast run
c. short piece of music
d. type of hot food

16. pressure: They used too much **pressure**.
a. money
b. time
c. hard pushing
d. bad words

17. refer: She **referred** to him.
a. supported him
b. let him go first
c. talked about him
d. answered him

18. army: They saw the **army**.
a. black and white animal
b. place where books are kept
c. person who lives nearby
d. people who protect a country

19. knee: Take care of your **knee**.
a. small child
b. part of your leg
c. plan for spending money
d. something that is yours

20. rope: He found a **rope**.
a. thick and strong string
b. something used to make holes
c. strong box for keeping money
d. metal tool used to climb up high

21. brand: This is a good **brand**.
a. dance party
b. first try
c. place to wait for others
d. name of a company

22. seal: They **sealed** it.
a. fixed it
b. closed it tightly
c. looked at it carefully
d. opened it quickly

23. warn: They were **warned**.
a. pushed away
b. welcomed inside
c. told about bad things
d. led into war

24. reserve: They have large **reserves**.
a. things kept to use later
b. machine for making bread
c. money from other people
d. group that runs a company

NVLT Part 3: 3000-word band

1. restore: It has been **restored**.
a. said again
b. given to a different person
c. given a lower price
d. made like new again

2. compound: They made a new **compound**.
a. agreement between two people
b. thing made of two or more parts
c. group that works together
d. guess based on past experience

3. latter: I agree with the **latter**.
a. man from the church
b. reason given before
c. second one of two things
d. answer to the spoken question

4. pave: It was **paved**.
a. stopped quickly
b. divided into many parts
c. given gold edges
d. covered with a hard surface

5. remedy: We found a good **remedy**.
a. way to fix a problem
b. place to eat in public
c. way to prepare food
d. rule about numbers

6. bacterium: They didn't find a single **bacterium**.
a. small living thing causing sickness
b. plant with red or orange flowers
c. animal that carries water on its back
d. thing that has been stolen and sold to a shop

9. silk: It's made of **silk**.
a. smooth and soft cloth
b. hard black wood
c. animal fur
d. very light metal

10. conceive: Who **conceived** the idea?
a. told it to others
b. explained it
c. thought of it first
d. said it was bad

11. legend: It is now a **legend**.
a. building for keeping old things
b. thing that is always done
c. story from the past
d. event that happens regularly

12. impose: This was **imposed**.
a. completely changed
b. in the middle of other things
c. made to look like something else
d. forced to happen by someone in power

13. solution: There is no **solution**.
a. time
b. support
c. problem
d. answer

14. celebrate: We have **celebrated** a lot recently.
a. found something for the first time
b. seen many new places
c. worked very hard
d. had a lot of parties

17. reward: He got a good **reward**.
a. things said about him by others
b. someone to help him in the house
c. money or gift for the things he did
d. large group of people to listen to him

18. review: The committee **reviewed** the plan.
a. examined it carefully for a decision
b. agreed to allow
c. made more just like it
d. threw it away

19. mode: The **mode** of production has changed
a. type
b. speed
c. attitude
d. amount

20. personnel: I don't like the **personnel** there.
a. type of chair that folds
b. machine that controls the heat
c. people who work there
d. person who owns a company

21. competent: She was very **competent**.
a. very fast
b. made angry easily
c. able to do things
d. easily hurt

22. devastate: The city was **devastated**
a. made beautiful for a special occasion
b. separated from the rest of the world
c. suffered great damage
d. made dirty by small animals

7. behavior: Look at her **behavior**!
a. people who have come to listen
b. the way she acts
c. large amount of money
d. small land with water around it

8. fuel: Do you have any **fuel**?
a. material used to make energy
b. a drug that stops pain
c. clothing used to keep you warm
d. a material put in walls to keep heat inside

15. independence: He has too much **independence**.
a. freedom from outside control
b. time by himself
c. physical strength
d. feeling of being better than others

16. tunnel: We need a **tunnel** here.
a. way through or under something
b. long piece of wood or metal to hold
c. mark on paper to show a short space
d. piece of material to cover a window

23. constituent: This is an important **constituent**.
a. building
b. agreement
c. idea
d. part

24. weave: She knows how to **weave**.
a. make cloth
b. join pieces of metal together
c. make people think something
d. trick people

NVLT Part 4: 4000-word band

1. patience: He has a lot of **patience**.
a. ability to wait
b. free time
c. faith in God
d. knowledge

2. strap: She broke the **strap**.
a. promise
b. top
c. plate
d. belt

3. weep: He **wept**.
a. finished school
b. cried
c. died quickly
d. thought deeply

4. haunt: The house is **haunted**.
a. full of decorations
b. allowed to be used for money
c. completely empty
d. full of ghosts

5. cube: I need one more **cube**.
a. pin
b. box
c. cup
d. postcard

6. peel: Shall I **peel** it?
a. let it sit in water for a long time
b. take the skin off it
c. make it white
d. cut it into thin pieces

9. romance: They had a short **romance**.
a. difference of opinion
b. holiday away from home
c. serious discussion
d. love relationship

10. ambition: He has no **ambition**.
a. strong desire to do well
b. ability to understand people's feelings
c. ability to make new things
d. enjoyment of life

11. dash: They **dashed** over it.
a. ran quickly
b. walked slowly
c. fought bravely
d. looked quickly

12. drown: People have **drowned** here.
a. eaten outside
b. died in water
c. dug a hole
d. cut down trees

13. originate: It **originated** here.
a. grew very well
b. changed shape
c. remained
d. first started

14. leaf: He touched the **leaf**.
a. part of a plant
b. soft shoe
c. top of a bottle
d. glass window

17. exert: Don't **exert** yourself!
a. praise too much
b. hurt yourself
c. work too hard
d. give yourself everything you want

18. marble: It was made of **marble**.
a. hard stone
b. hard wood
c. soft metal
d. soft cloth

19. diminish: It has **diminished**.
a. become dark
b. become less in size
c. become cloudy
d. grown colder

20. sheriff: The **sheriff** was friendly.
a. pilot
e. b. housekeeper
c. policeman
d. teacher

21. monarch: They saw the **monarch**
a. army group
b. gate
c. king or queen
d. criminal

22. plunge: It **plunged**.
a. danced around
b. was made quiet
c. dropped suddenly
d. stayed still

7. distress: He felt **distressed**.

a. unwanted

b. satisfied

c. unhappy

d. energetic

15. amateur: She is an **amateur** player.

a. someone who plays for fun, not money

b. player who replaces other hurt players

c. player representing her country

d. ball-sports player

23. mourn: They **mourned** for several years.

a. performed on the street

b. felt very sad

c. worked hard

d. used their money carefully

8. depart: She **departed** yesterday.

a. went away

b. said no

c. went down a hill

d. got worse

16. evacuate: They were **evacuated**.

a. moved to another place for safety

b. searched for guns or knives

c. frightened suddenly

d. made to look like criminals

24. fragile: These things are very **fragile**.

a. Special

b. hard to find

c. popular

d. easily broken

NVLT Part 5: 5000-word band

1. scrub: He is **scrubbing** it.
a. cleaning
b. repairing
c. worrying about
d. drawing pictures

2. dinosaur: The children were pretending to be **dinosaurs**.
a. people who look for gold
b. small people that fly
c. animals that make fire
d. animals that lived a long time ago

3. nun: We saw a **nun**.
a. small worm
b. big accident
c. woman who serves her religion
d. strange light in the sky

4. compost: We need some **compost**.
a. strong support
b. mental help
c. strong material that is used for building
d. soil used to help the garden

5. miniature: It is a **miniature**.
a. small version of something
b. brick house
c. very small living creature
d. detailed plan for a building

6. crab: Do you like **crabs**?
a. small sea animals
b. hard thin salty bread
c. original copy of a piece of music
d. insect which sings and jumps

9. rove: He is **roving**.
a. getting drunk
b. traveling around
c. making a musical sound with his lips
d. working hard using his body

10. divert: The rivers were **diverted**.
a. made to move in a different way
b. given bridges
c. made very dirty
d. made wider and deeper

11. trench: They looked at the **trench**.
a. mountain
b. long hole
c. pile of trash
d. beautiful sight

12. technician: She is a **technician**.
a. man with magical abilities
b. person who works with and fixes machines
c. doctor who cares for young children
d. person who is good at music

13. query: I have a **query**.
a. headache
b. large amount of money
c. question
d. good idea

14. mug: This **mug** needs a wash.
a. big cup
b. old car you like
c. clothes worn under other clothes
d. area in front of the door where rain and wind cannot reach

17. spider: We caught the **spider**.
a. disease that gives red spots
b. small animal with eight legs
c. small public bus
d. oily fish

18. circus: We went to the **circus**.
a. place for people who love God
b. traveling company of entertainers
c. place where people run races
d. music group

19. sofa: He bought a **sofa**.
a. soft seat for two or more people
b. cutting machine
c. long pipe for putting water on the garden
d. a small car with four wheels that a baby can ride in while someone pushes it

20. logo: They have a pretty **logo**.
a. tree with red fruit
b. reception
c. picture or word that represents a company
d. a holiday home

21. commemorate: We must **commemorate** his actions.
a. remember something or someone
b. pretend to agree with something
c. protest against something
d. say good things about him

22. crook: They were **crooks**.
a. people who are not honest
b. people who work at hospitals
c. people who cannot walk
d. people who design buildings

7. vocabulary: You will need more **vocabulary**.
a. words
b. skills
c. money
d. guns

8. corpse: The **corpse** was found in the park.
a. large and deep cup
b. mobile phone
c. artist's hat
d. dead body

15. static: It's **static** at the moment.
a. not popular
b. demanded by law
c. often said
d. not moving or changing

16. slaughter: We read about the **slaughter** in the paper.
a. problem
b. scientific research
c. killing
d. sports event

23. volt: How many **volts** were used?
a. large envelope for business letters
b. something used to add flavor to food
c. units measuring electrical power
d. material that attracts other metals

24. warfare: Modern **warfare** is frightening.
a. crime
b. dancing
c. fighting
d. pollution

NVLT Part 6: Academic Word List

1. concept: This is a difficult **concept**.

a. legal agreement

b. idea about what something is

c. way of doing things

d. a written explanation of a law

2. similar: These articles are similar.

a. about a certain thing

b. of great quality

c. easy to understand

d. close to the same

3. item: The next **item** is very important.

a. thing on a list

b. question sheet

c. meeting of people

d. way something looks

4. component: Each **component** is very important.

a. set of ideas which support something.

b. flat part that sits on top of another

c. small part of something bigger

d. the person you work with

5. compensate: The government should **compensate** the farmers.

a. give something good to balance something bad

b. stop them from joining a group

c. find where they are

d. bring them together

6. professional: She wants to be a **professional** musician.

a. someone who stays at home

b. someone who gets paid to play

c. someone on a list

d. someone known by many people

9. migrate: The animals began to **migrate**.

a. work together

b. move together to a different place

c. come together as a group

d. change together

10. priority: That is our **priority**.

a. deal between two people

b. most important thing

c. something that has been printed

d. person who comes next

11. reverse: Try it in **reverse**.

a. the other direction

b. the way things are arranged

c. with the correct sound

d. at the correct time

12. arbitrary: Her decision was **arbitrary**.

a. not chosen for a reason

b. necessary for success

c. not able to be changed

d. good enough for a purpose

13. mutual: The feeling was **mutual**.

a. easy to understand

b. fully developed

c. the same between two people

d. kept under control

14. alternative: Is there an **alternative**?

a. another choice

b. thing to do

c. something to say

d. activity with many people

17. site: He looked for a better **site**.

a. basic part of something

b. opinion about the price

c. place where something is

d. something brought from another country

18. institute: We must **institute** new changes

a. get with effort

b. control with laws

c. begin or create

d. search for

19. retain: How will the club **retain** its memebers?

a. mix them together

b. help them develop

c. help them work together

d. keep them

20. phase: This is one **phase** of the new system.

a. list of things in a special order

b. short part of a process

c. range of levels

d. rule that controls what something is

21. pursue: This year she will **pursue** the group's goals.

a. try to get

b. change

c. check over time

d. make easier

22. recover: The men **recovered** their strength.

a. showed other people

b. used for a reason

c. said that they know

d. got back

7. external: They worried about the **external** damage.

a. not known

b. outside

c. based on facts

d. following

8. clause: Please fix that **clause**.

a. part of a sentence

b. something you are trying to do

c. large picture

d. small object

15. colleague: That is my **colleague**.

a. something that people talk about

b. plan of things to do

c. person you work with

d. piece of writing

16. legal: Is this meeting place **legal**?

a. based on the law

b. free to be used

c. easy to see

d. important to someone

23. diverse: Having **diverse** information is important.

a. with no mistakes

b. very small amount

c. able to be changed

d. having different types

24. hierarchy: This **hierarchy** is very common.

a. set of ideas a group has

b. group with people at different levels

c. dangerous material

d. popular way of dressing

25. distort: The image is **distorted**.

a. having more than one meaning

b. exactly the same as something else

c. has a badly changed shape

d. from recent times

26. accumulate: He **accumulated** many friends.

a. understood the value

b. got more and more

c. said good things about

d. became the same as

27. abandon: He **abandoned** the project.

a. used it for his own gain

b. controlled in a clever way

c. stopped working on it

d. made it as small as possible

28. rigid: These rules are **rigid**.

a. how good something is

b. happening at the same time

c. continuing for a limited time

d. not able to be changed

29. notwithstanding: **Notwithstanding** John's feelings, Allison went to France.

a. without knowing

b. giving back in the same way

c. because of

d. not being stopped by

30. perspective: You have a good **perspective**.

a. events that happen again and again

b. way of seeing things

c. group of people you know

d. how other people see you

Note. The updated Vocabulary Levels Test, Version 4.5 From. S. McLean, & B. Kramer. (2015). The creation of a New Vocabulary Levels Test. *Shiken*, *19*(2), 1-11.

The complete test can be downloaded from https://www.brandonkramer.net/resources

Appendix B.3: An Example of the Narrative and Academic YARC Passages

The following is an excerpt from the narrative passage of the The York Assessment of Reading Comprehension (YARC, Snowling et al., 2009)

The Schoolboy

The 'Back to School' signs had been in the shop windows for weeks now. Norman Kirk had always loved the last night of the summer holidays: laying out his clothes for the next morning - shirt, tie, trousers, socks, and finally, the new shoes. It was a relief to get back into the familiar routine after the long. empty summer. He spread margarine into the corners of two white squares of bread and centred a slice of ham, carefully trimming the overhang. Closing the sandwich, he gently sawed a diagonal cut and then, wrapping it in a new sheet of foil, he laid it in the plastic box next to the green apple and the chocolate biscuit. He never tired of this choice of lunch. As he closed the fridge door, he glanced at his watch - nearly nine. A whole hour to spare before bed. Usually there was school work to look at, but not tonight.

Norman turned on the TV. On the first channel “Entertainment Tonight” blared out. A room full of unknown performers sat waiting to be discovered, each hoping that their act would be chosen and propel them to fame and fortune. Norman groaned, entertainment - what a joke - sitting in a darkened room would be more enjoyable. The second channel was showing an old episode of City Detectives. He'd loved this series when he was younger; the suspense of trying to guess who perpetrated the crime, the satisfaction of being right, the groan of a twisted plot.

future during the long, dark evenings of the autumn-winter term.

The following is an excerpt from the fictional passage of the The York Assessment of Reading Comprehension (YARC, Snowling et al., 2009)

Honey for You, Honey for Me

In Southern Africa there is a bird called the Honey Guide. It is a small bird with a long pink beak. Its favourite food is honey. From a distance, the Honey Guide looks drab and brown, but up close you can see a splash of pale yellow on the white chest feathers. It looks a little as if the bird has just enjoyed a meal of golden honey, and been none too careful about its table manners! However, the Honey Guide gets its name not just from the colour of its chest; it is very well adapted to feeding on the contents of beehives. It doesn't just eat the honey, but also bee eggs, larvae, pupae and even beeswax. In fact, they are one of only a handful of birds that can digest wax. The Honey Guide is what you might call a bee specialist.

It does, however, have one major problem: bees sting. The Honey Guide is not a big bird, and bee stings can be very dangerous to it, or even fatal. The bird has to find a way to get at the bees' hive without being badly stung. The Honey Guide has developed a very elegant solution to the problem. It uses humans. The Honey Guide searches around its territory in the African grasslands until it finds a likely-looking beehive. When it has found one, it flies off to find some helpful humans.

Appendix B.4: The Updated Vocabulary Levels Test (Webb et al., 2017)

This is test that looks at how well you know useful English words. Put a check under the word that goes with each meaning. Here is an example.

	game	island	mouth	movie	song	yard
land with water all around it						
part of your body used for eating and talking						
piece of music						

It should be answered in the following way.

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

Question 1	choice	computer	garden	photograph	price	week
cost						
picture						
place where things grow outside						

Question 2	eye	father	night	van	voice	year
body part that sees						
parent who is a man						
part of the day with no sun						

Question 3	center	note	state	tomorrow	uncle	winter
brother of your mother or father						
middle						
short piece of writing						

Question 4	box	brother	horse	hour	house	plan
family member						
sixty minutes						
way of doing things						

Question 5	animal	bath	crime	grass	law	shoulder
green leaves that cover the ground						
place to wash						
top end of your arm						

Question 6	drink	educate	forget	laugh	prepare	suit
get ready						
make a happy sound						
not remember						
Question 7	check	fight	return	tell	work	write
do things to get money						
go back again						
make sure						

Question 8	bring	can	reply	stare	understand	wish
say or write an answer to						
carry to another place						
look at for a long time						

Question 9	alone	bad	cold	green	loud	main
most important						
not good						
not hot						

Question 10	awful	definite	exciting	general	mad	sweet
certain						
usual						
very bad						

Part 2:

Question 1	coach	customer	feature	pie	vehicle	weed
important part of something						
person who trains members of sports teams						
unwanted plant						

Question 2	average	discipline	knowledge	pocket	vegetable	trap
food grown in gardens						
information which a person has						
middle number						

Question 3	circle	justice	knife	onion	partner	pension
round shape						
something used to cut food						
using laws fairly						

Question 4	cable	section	sheet	site	staff	tank
part						
place						
something to cover a bed						

Question 5	apartment	cap	envelope	lawyer	speed	union
cover for letters						
kind of hat						
place to live inside a tall building						

Question 6	argue	contribute	quit	seek	vote	wrap
cover tightly and completely						
give to						
look for						

Question 7	avoid	contain	murder	search	switch	trade
have something inside						
look for						
try not to do						

Question 8	bump	complicate	include	organize	receive	warn
get something						
hit gently						
have as part of something						

Question 9	available	constant	electrical	medical	proud	super
feeling good about what you have done						
great						
happening all the time						

Question 10	smooth	junior	pure	rotten	environmental	wise
bad						
not rough						
younger in position						

Part 3:

Question 1	angle	apology	behavior	bible	celebration	portion
actions						
happy occasion						
statement saying you are sorry						

Question 2	anxiety	athlete	counsel	foundation	phrase	wealth
combination of words						
guidance						
large amount of money						

Question 3	agriculture	conference	frequency	liquid	regime	volunteer
farming						
government						
person who helps without payment						

Question 4	asset	heritage	novel	poverty	prosecution	suburb
having little money						
history						
useful thing						

Question 5	audience	intelligence	crystal	outcome	pit	welfare
ability to learn						
deep place						
people who watch and listen						

Question 6	consent	enforce	exhibit	retain	specify	target
agree						
say clearly						
show in public						

Question 7	accomplish	capture	debate	impose	proceed	prohibit
catch						
go on						
talk about what is correct						

Question 8	absorb	decline	exceed	link	nod	persist
continue to happen						
goes beyond the limit						
take in						

Question 9	approximate	frequent	graphic	pale	prior	vital
almost exact						
earlier						
happening often						

Question 10	consistent	enthusiastic	former	logical	marginal	mutual
not changing						
occurring earlier in time						
shared						

Part 4:

Question 1	cave	scenario	sergeant	stitch	vitamin	wax
healthy supplement						
opening in the ground or in the side of a hill						
situation						

Question 2	candle	diamond	gulf	salmon	soap	tutor
something used for cleaning						
teacher						
valuable stone						

Question 3	agony	kilogram	orchestra	scrap	slot	soccer
group of people who play music						
long, thin opening						
small unwanted piece						

Question 4	crust	incidence	ram	senator	venue	verdict
hard outside part						
judgment						
place						

Question 5	alley	embassy	hardware	nutrition	threshold	tabacco
government building						
plant that is smoked in cigarettes						
small street between buildings						

Question 6	fling	forbid	harvest	shrink	simulate	vibrate
do not allow						
make smaller						
throw						

Question 7	activate	disclose	hug	intimidate	plunge	weep
cry						
tell						
turn on						

Question 8	diminish	exaggerate	explode	penetrate	transplant	verify
break into pieces violently						
get smaller						
move something to another place						

Question 9	adjacent	crude	fond	sane	spherical	swift
beside						
not crazy						
quick						

Question 10	abnormal	bulky	credible	greasy	magnificent	optical
believable						
oily						
unusual						

Part 5:

Question 1	gown	maid	mustache	paradise	pastry	vinegar
hair on your upper lip						
perfect place						
small baked food						

Question 2	asthma	chord	jockey	monk	rectangle	vase
container for cut flowers						
group of musical notes that are played at the same time						
shape with two long and two short sides						

Question 3	batch	dentist	hum	lime	pork	scripture
green fruit						
low, constant sound						
meat from pigs						

Question 4	amnesty	claw	earthquake	perfume	sanctuary	wizard
liquid that is made to smell nice						
man who has magical powers						
safe place						

Question 5	altitude	diversion	hemisphere	pirate	robe	socket
height						
kind of clothing						
person who attacks ships						

Question 6	applaud	erase	jog	intrude	notify	wrestle
announce						
enter without permission						
remove						

Question 7	bribe	expire	immerse	meditate	persecute	shred
cut or tear into small pieces						
end						
think deeply						

Question 8	commemorate	growl	ignite	pierce	renovate	swap
catch fire						
exchange						
go into or through something						

Question 9	bald	eternal	imperative	lavish	moist	tranquil
calm and quiet						
having no hair						
slightly wet						

Question 10	diesel	incidental	mandatory	prudent	superficial	tame
not dangerous						
required						
using good judgment						

Note. The new Vocabulary Levels Test, from: S. Webb, Y. Sasao, & O. Balance. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*(1), 34-70.

The complete test can be downloaded from https://www.edu.uwo.ca/faculty-profiles/docs/other/webb/NVLT-VERSION-B.pdf

Appendix C: Bibliographical Information for the Textbooks

Adkins, J., & Lackovic, M. (2020). *English A: Literature for the IB Diploma* (2nd ed.). Pearson.

Allison, R., & Chanen, B. (2019). *English A: Language and literature course companion* (2nd ed.). Oxford University Press.

Allott, A., & Mindorf, D. (2014). *Biology course companion*. Oxford University Press.

Allum, J., & Talbot, C. (2014). *Physics for the IB Diploma* (2nd ed.). Hodder Education.

Amy, N., Henly, C. P., Johnson, A. S., & Waller, K. C. (2019). *English language for the IB Diploma*. Hodder Education.

Androulaki, A., & Whitted, B. (2019). *English A: Literature course companion for the IB Diploma* (2nd ed.). Oxford University Press.

Bastian, S., Kitching, J., & Sims, R. (2020). *Theory of Knowledge for the IB Diploma* (3rd ed.). Pearson Education.

Blink, J., & Dorton, L. (2020). *Economics course companion*. Oxford University Press.

Blyth, B., Bruder, G., Cirrito, F., Henry, M., Hung, B., Larson, W., McAuliffe, R., & Sanders, J. (2019). *Mathematics: Common core* (6th ed.). IBID Press.

Brown, C., & Ford, M. (2014). *Chemistry higher level* (2nd ed.). Pearson Education.

Bryan, C., & Thomas, G. (2020). *Theory of Knowledge essentials*. Pearson Education.

Bylikin, S., Horner, G., Murphy, B., & Tarcu, D. (2014). *Chemistry course companion*. Oxford University Press.

Clegg, C. J. (2014). *Biology for the IB Diploma* (2nd ed.). Hodder Education.

Dailey, A., & Webb, S. (2012). *Access to history for the IB Diploma: Causes, practices and effects of wars*. Hodder Education.

Damon, A., McGonegak, R., Tosto, P., & Ward, W. (2014). *Biology higher level* (2nd ed.). Pearson Education.

Fannon, P., Kadelburg, V., Woolley, B., & Ward, S. (2012). *Mathematics higher level for the IB Diploma* (2nd ed.). Cambridge University Press.

Fannon, P., Kadelburg, V., Woolley, B., & Ward, S. (2020). *Mathematics: Analysis and approaches HL*. Hodder Education.

Farmer, A. (2013). *Access to history for the IB Diploma: Independence movements*. Hodder Education.

Fensom, J. (2020). *Mathematics: For IB Diploma course preparation* (6th ed.). Oxford University Press.

Green, J., & Damji, S. (2008). *Chemistry* (3rd ed.). IBID Press.

Haese, M., Humphries, M., Sangwin, C., & Vo, N. (2019). *Mathematics: Analysis and approaches HL 2*. Haese Mathematics.

Haese, M., Humphries, M., Sangwin, C., & Vo, N. (2019). *Mathematics: Core topics HL 1*. Haese Mathematics.

Hamper, C. (2014). *Physics higher level* (2nd ed.). Pearson Education.

Henly, C. P., & Spraque, J. (2020). *Theory of Knowledge for the IB Diploma* (4th ed.). Hodder Education.

Henly, C. P., & Johnson, A. S. (2019). *Textual analysis for English language and literature for the IB Diploma*. Hodder Education.

Heydorn, W., & Jesudason, S. (2020). *Decoding Theory of Knowledge for the IB Diploma* (2nd ed.). Cambridge University Press.

Hoang, P. (2017). *Business management* (4th ed.). IBID Press.

Homer, D., & Bowen-Jones, M. (2014). *Physics course companion*. Oxford University Press.

Kerr, G., & Ruth, P. (2007). *Physics* (3rd ed.). IBID Press.

Kirsch, M. (2017). *Global politics course companion*. Oxford University Press.

Lomine, L., & Muchena, M. P., Robert A. (2014). *Business management course companion*. Oxford University Press.

McGee, M. (2020). *Economics: In terms of the good, the bad and the economist* (3rd ed.). IBID Press.

Murphy, R., & Gleek, C. (2016). *Global politics essentials*. Pearson Education.

Nutt, S., & Bottaro, J. (2011). *History for the IB Diploma: Nationalist and independence movements*. Cambridge University Press.

Owen, S., Ahmed, C., Martin, C., & Woodward, R. (2011). *Chemistry for the IB Diploma* (2nd ed.). Cambridge University Press.

Philpot, B. (2019). *English A: Language and literature for the IB Diploma* (2nd ed.). Cambridge University Press.

Smith, D. M. (2015). *Causes and effects of 20th-century wars course companion*. Oxford University Press.

Sprague, J. (2020). *Theory of Knowledge Skills for Success* (2nd ed.). Hodder Education.

Stimpson, P., & Smith, A. (2015). *Business management for the IB Diploma* (2nd ed.). Cambridge University Press.

Talbot, C., Harwood, R., & Coates, C. (2015). *Chemistry for the IB Diploma* (2nd ed.). Hodder Education.

Tandy, L., Gibbons, A., & Koszary, J. (2019). *English language and literature for the IB Diploma*. Hodder Education.

Tragakes, E. (2020). *Economics for the IB Diploma* (3rd ed.). Cambridge University Press.

Tsokos, K. A. (2014). *Physics for the IB Diploma* (6th ed.). Cambridge University Press.

Uzunova Dang, M., & Uzunoy Dang, A. S. (2020). *Theory of Knowledge course companion*. Oxford University Press.

van de Lagemaat, R., Heydorn, W., & Jesudason, S. (2020). *Theory of Knowledge for the IB Diploma* (3rd ed.). Cambridge University Press.

Walpole, B., Merson-Davies, A., & Dann, L. (2014). *Biology for the IB Diploma* (2nd ed.). Cambridge University Press.

Wathall, J. C., Harcet, J., Harrison, R., Heinrichs, L., & Torres-Skoumal, M. (2019). *Mathematics: Analysis and approaches higher level*. Oxford University Press.

Wazir, I., & Garry, T. (2019). *Mathematics: Analysis and approaches higher level* (6th ed.). Pearson Education.

Weem, M. P., Talbot, C., & Mayrhofer, A. (2007). *Biology*. IBID Press.

Wells, M. (2011). *History for the IB Diploma: Causes, practices and effects of wars*. Hodder Education.

Bibliographical Information for the Textbook Used in the Parallel Corpus

Primrose, D. M. (2019). *Biology: IB prepared*. Oxford University Press.

Appendix D: R-Scripts

Example of the R script for cleaning the text files in Chapter 4:

```
library(tidyverse)
library(tidytext)

## patterns
start_pattern <- START %R% "<" %R% "[^/]*?" %R% ">"
end_pattern <- "</" %R% ".*?" %R% ">" %R% END
all_pattern <- c(start_pattern, end_pattern)

## functions
## put text into a df and remove empty lines
import_text <- function(file){
  text <- readLines(file) %>% str_subset(., ".+")
  text_df <- enframe(text, name = "paragraph", value = "text")
  return(text_df)
}

## identifies heading
id_heading <- function(df){
  h_df <- df %>% mutate(text = sub("##", "<h>", text))
  return(h_df)
}

## Enter book_id
book_id <- function(text_df, file_name) {
  book <- str_split(file_name, fixed("/")) %>% unlist() %>%
    .[length(.)-1] %>% tolower()
  book_df <- text_df %>% mutate(book_id = book) %>% mutate(status = "cleaned")
  return(book_df)
}

## Enter a chpt column
chpt_column <- function(text_df){
  chpt_pattern <- "chpt " %R% one_or_more(DGT)
  chpt <- text_df[2,2]
  text_df["chpt"] <- chpt
  text_df <- text_df %>% mutate(chpt = str_match(tolower(chpt), chpt_pattern)) %>%
    mutate(chpt = parse_number(chpt))
  return(text_df)
}

## Enter page numbers
page_column <- function(text_df){
  pn_pattern <- "<pn:" %R% one_or_more(DGT) %R% ">"
  first_page <- str_locate(text_df$text, pattern = pn_pattern)
```

```
  pages <- cbind(text_df, first_page) %>% filter(!is.na(start)) %>%
    mutate(page_num = parse_number(text)) %>% select(paragraph, page_num)
  pn_df <- text_df %>%
    left_join(pages, by = "paragraph")  %>%
    fill(page_num) %>% drop_na(page_num)
  return(pn_df)
}

## Enter the type of text
tt_column <- function(text_df){
  type_df <- text_df %>% mutate(text, start_type = str_extract(text, start_pattern)) %>%
    mutate(end_type = str_extract(text, end_pattern)) %>%
    filter((start_type != "<pb>" & !str_detect(start_type, "<pn")) %>%
replace_na(TRUE)) %>%
    mutate(paragraph = row_number())

  new_type_df <- type_df %>% mutate(text_type = case_when(
    str_detect(start_type, "##") ~ "header",
    str_detect(start_type, "<h>") ~ "header",
    str_detect(start_type, "<fc>") ~ "figure_caption",
    (str_detect(start_type, "<c>")) & (str_detect(lag(start_type), "<t>")) ~
"table_caption",
    str_detect(start_type, "<c>") ~ "caption",
    str_detect(start_type, "<title>") ~ "title",
    str_detect(start_type, "<f") ~ "figure",
    str_detect(start_type, "<t>") ~ "table",
    str_detect(start_type, "<#>") ~ "missing",
    (str_detect(start_type, start_pattern)) & (str_detect(end_type, end_pattern)) ~
start_type,
    TRUE ~ "text"
  )
  )
  return(new_type_df)
}

## Enter the paragraph type
type_column <- function(text_df){
  types_df <- text_df %>% mutate(type = case_when(
    str_detect(text_type, "caption") ~ "caption",
    str_detect(text_type, "figure")  ~ "figure",
    str_detect(text_type, "table") ~ "table",
    str_detect(text_type, "image") ~ "image",
    str_detect(start_type, "<#>") ~ "missing",
    str_detect(start_type, "<b") ~ "text_box_start",
    str_detect(end_type, "</b") ~ "text_box_end",
    str_detect(start_type, "<a") ~ "activity_start",
    str_detect(end_type, "</a") ~ "activity_end",
    TRUE ~ "main"
```

```
  )
  )
  return(types_df)
}

## Clean the df
clean_up <- function(text_df){
  clean_df <- text_df %>% mutate(text = str_remove_all(text, paste(all_pattern, collapse
= "|"))) %>%
    mutate(text = case_when(
      (str_detect(text_type, "figure\\b")) & (text == "")  ~ "<figure>",
      str_detect(text_type, "table\\b")  & (text == "") ~ "<table>",
      str_detect(text_type, "image") & (text == "") ~ "<image>",
      str_detect(start_type, "<#>") ~ "<missing>",
      TRUE ~ text
    )
    ) %>% filter(text != "") %>% select(book_id, status, chpt, page_num, paragraph,
type, text_type, text)
  return(clean_df)
}

pipeline <- function(file_name){
  df <- import_text(file_name)
  h_df <- id_heading(df)
  df_book <- book_id(h_df, file_name)
  df_chpts <- chpt_column(df_book)
  df_pages <- page_column(df_chpts)
  df_tt <- tt_column(df_pages)
  df_types <- type_column(df_tt)
  df_clean <- clean_up(df_types)
  print(file_name)
  return(df_types)
}


## Select the files
selected_file <- file.choose()
directory <- dirname(selected_file)
files <- list.files(directory, pattern = ".txt" %R% END)

## Process the files
df_book <- data.frame()
for(file in files) {
  full_name <- paste(directory, file, sep = "/")
  df_temp <- pipeline(full_name)
  df_book <- bind_rows(df_book, df_temp)
  rm(df_temp)
}
```

The R script for compiling the IS-AVL word lists in Chapter 5:

```
rm(list=ls(all=TRUE))               # clears the memory
library(rstudioapi)                 # Loads the rstudioapi - needed to get current dir
library(tidytext)                   # Loads the tidytext library
library(tidyverse)                  # Loads the tidyverse library

## Gets the directory the script is in so that it is possible
## to determine all the other directories
main_dir <- getActiveDocumentContext()$path %>%
  str_remove(., "/Scripts/.*")

## Regular expressions used for getting information
pattern_book <- "[:upper:]{2}[:digit:]{2}"    # Gets book name from file name
pattern_sub <- "[:upper:]{2}"                 # Gets the subject from the file name
pattern_pos <- "(?<=_)[:lower:]*"             # Gets the part of speech

in_corpus <- 100/1000000                      # Sets the tpm ratio for in corpus
across_corpus <- 11.4/1000000                 # Sets the tpm for across corpus
all_corpus <- 28.5/1000000                    # Sets the tpm for total corpus

## This function gets the word frequency and
## then removes gsl words and low frequency words
get_wd_freq <- function(df_words){            # Sets up the function with df as input
  wd_freq <- df_words %>%                     # Loads the corpus dataframe
    group_by(subject) %>%                     # Groups by subject
    mutate(percentage = n/sum(n)) %>%         # calculates the frequency of each word
    anti_join(list_freq) %>%                  # removes the GSL words
    filter(percentage > across_corpus) %>%    # removes words with a freq <11.4 in subj
    left_join(list_bnc) %>%                   # Gets BNC/COCA frequency band information
    filter(!list %in% c("word_abbreviations", # Removes abbreviations from final list
                        "proper_nouns"))      # Removes proper nouns from the final list
}

## This function the coverage of the GSL over the corpus
get_gsl_cov <- function(df_words){           # Sets up the function with df as input
  wd_freq <- df_words %>%                    # Loads the corpus dataframe
    left_join(list_bnc) %>%                  # Gets BNC/COCA frequency band information
    group_by(subject) %>%                    # Groups by subject
    mutate(percentage = n/sum(n)) %>%        # Calculates the frequency of each word
    right_join(list_freq, by = "word")       # Returns only words in the GSL
}

## Get word lists
list_freq <- read.csv(paste0(main_dir,
                             "/lists/gsl_lists.csv")) %>%      # Imports the GSL to a df
  mutate(head = na_if(head, "")) %>% fill(head) %>%            # Creates headword column
  mutate(word = tolower(word))                                 # Converts to lower case
```

```
list_bnc <- read.csv(paste0(main_dir,
                            "/lists/bnc_coca.csv")) %>%        # Imports BNC/COCA to df
  mutate(head = na_if(head, "")) %>% fill(head) %>%            # Creates headword column
  mutate(word = tolower(word))                                 # Converts to lower case

list_awl <- read.csv(paste0(main_dir, "/lists/awl.csv")) %>%   # Imports the AWL to a df
  mutate(head = na_if(head, "")) %>% fill(head) %>%            # Creates headword column
  mutate(word = tolower(word))                                 # Converts to lower case

list_stop <- read.csv(paste0(main_dir, "/lists/stop_words.csv")) %>%   # Stop words
  mutate(word = tolower(word))                                         # To lower case

corpus_list <- c("literature", "biology",                      # List of all subjects
                 "chemistry", "physics", "maths",
                 "economics", "social", "tok")

## This sets the subject names so that they can be saved to a table later
subject_names <- data.frame(
  subject = c("EM", "EC", "SS", "BM", "CM", "PM", "MM", "TK"),
  subject_sort = c("1_LM", "6_EC", "7_SS", "2_BM", "3_CM", "4_PM",
                   "5_MM", "8_TK"),
  subject_name = c("literature", "economics", "social", "biology",
                   "chemistry", "physics", "maths", "tok")
)

## This is where all the tables and exported files will be saved for use later
table_directory <- paste0(main_dir, "/Tables and Figures") # A directory for tables
stage5 <- paste0(main_dir, "/_Step 5 Word Frequencies")    # A directory for word lists

## This identifies all the corpus files are stored and gets a list of files
directory <- paste0(main_dir, "/_Step 3 Clean Corpus") # Main dir: Change if necessary
files <- list.files(directory, pattern = ".*.csv")     # Gets a list of files in dir

## Builds the initial corpus dataframe
corpus_df <- data.frame()                              # Creates an empty dataframe
for(file in files) {                                   # This loops over all the files
  file_to_read <- paste0(directory, "/", file)         # Gets the file path
  df_file <- read.csv(file_to_read) %>%                # Reads the file
    mutate(subject = str_extract(book,
                                 pattern_sub)) %>%     # Gets the book and subject name
    select(book, subject, text)                        # Selects the columns we need
  corpus_df <- bind_rows(corpus_df, df_file)           # Binds the book df to corpus df
  rm(df_file)                                          # Cleans up
}

## Creates a tidy df with 1 row for each word
tidy_df <- corpus_df %>%                               # Loads initial corpus df
```

```
  unnest_tokens(word, text)                          # unnesting the words

## Creates a df of all the words in the corpus
df_words <- tidy_df %>%                               # Creates a df of words
  mutate(word = tolower(word)) %>%                    # Converts to lower case
  inner_join(subject_names) %>%                       # Gets the subject names
  mutate(subject = subject_name) %>%                  # Renames the column
  select(book, subject, word)                         # Selects columns we need

## Gets the frequency within the subcorpora
wd_count <- df_words %>%                              # Loads the corpus data frame
  left_join(list_bnc) %>%                             # Adds column for marginal words
  filter(!list %in% c("word_marginal")) %>%           # removes marginal words
  group_by(subject) %>%                               # Groups by subject
  count(word) %>%                                     # Counts the words in the subject
  arrange(desc(n)) %>%                                # Arranges in descending order
  ungroup() %>%                                       # Ungroups the dataframe
  get_wd_freq()                                       # Gets the frequency of the words

## Gets the frequency across all corpora
freq_all <- df_words %>%                            # Loads the corpus data frame
  anti_join(list_stop) %>%                          # Removes stop words
  left_join(list_bnc) %>%                           # Adds column for marginal words
  filter(!list %in% c("word_marginal")) %>%         # removes marginal words
  count(word) %>%                                   # Counts # of times word appears
  arrange(desc(n)) %>%                              # Arranges in descending order
  mutate(percentage = n/sum(n)) %>%                 # Gets frequency in corpus
  filter(percentage > all_corpus) %>%               # Removes low frequency words
  left_join(list_bnc, by = "word") %>%              # Adds headwords for later
  anti_join(list_freq, by = "word") %>%             # Removes GSL words
  mutate(tpm = percentage * 1000000)                # Gets the tokens per million
(tpm)

wl_all_corpora <- wd_count %>%                        # Gets frequency across corpora
  select(c("word", "subject", "percentage")) %>%      # Selects the columns we need
  pivot_wider(names_from = subject,
              values_from = percentage) %>%           # Converts the long to a wide df
  inner_join(freq_all) %>%                            # Gets the freq across corpora
  rename(all = percentage) %>%                        # Renames the column
  drop_na() %>%                                       # Removes the unnecessary rows
  anti_join(stop_words) %>%                           # Removes the stop words
  select("word", "all", "literature", "biology",      # Selects the columns we need
         "chemistry", "physics", "maths", "economics",
         "social", "tok", "head")

## Step 1: This get the 1st group of words for the list by adding words from the df
## that have a frequency greater than 28.5 in the whole corpus & greater than 11.4 in
## all of the subcorpora. These are added to all of the subject specific word lists
```

```
wl_group1 <- wl_all_corpora %>%
  pivot_longer("all":"tok", names_to = "subject",     # Creates a long df with the
               values_to = "percentage") %>%          # word, subj., and ave. for words
  select(word, subject, percentage, head)             # Selects the necessary columns

## Step 2: This is the 2nd group of words
## These are words from the AWL that have a freq higher than 28.5 across all corpora
## and greater than 11.4 tpm in a subcorpus and are added to each subcorpus individually
wl_group2 <- data.frame()                             # Creates an empty df
for(sub_name in corpus_list){                         # Loops over all the subjects
  df_all <- freq_all %>%                              # Gets words  >28.5 in corpus
    anti_join(list_stop, by = "word") %>%             # Removes stop words
    select(word)                                      # Selects the word column
  df_sub <- wd_count %>%                              # Gets the freq in the subject
    filter(subject == sub_name) %>%                   # Selects 1 subject
    inner_join(select(list_awl, word), by = "word") %>% # Gets all words in the AWL
    filter(percentage > across_corpus) %>%             # Removes low frequency words
    inner_join(select(df_all, word), by = "word") %>%  # Adds words freq across all
    select(word, subject, percentage)                 # Selects the columns we need
  wl_group2 <- bind_rows(wl_group2, df_sub)           # Adds the subject df
  rm(df_all, df_sub,sub_name)                         # Cleans up
}

# Calculates the number of words and coverage for this stage
wl_group2_count <- wl_group2 %>%                      # Imports the words from Step 2
  anti_join(wl_all_corpora, by = "word") %>%          # Removes words also in Step 1
  group_by(subject) %>%                               # Groups by subject
  summarize(words = n(),                              # Calculates number of words and
            coverage = sum(percentage)) %>%           # coverage for each subject
  inner_join(subject_names,                           # Gets the long name and sorting
             by = c("subject" = "subject_name")) %>%  # information for each subject
  arrange(subject_sort) %>%                           # Puts the subjects in order
  select(subject, words, coverage)                    # Selects the necessary rows

## Step 3: This is the 3rd group of words
## These are words with a frequency greater than 100 tpm in a subcorpus
## These words are added to that subcorpus
wl_group3 <- data.frame()                             # Creates an empty df
for(sub_name in corpus_list){                         # Loops over each subject
  df_sub <- wd_count %>%
    filter(subject == sub_name) %>%                   # Selects the subject
    filter(percentage > in_corpus) %>%                # Removes low frequency words
    select(word, subject, percentage)                 # Selects the columns we need
  wl_group3 <- bind_rows(wl_group3, df_sub)           # Creates a corpus df
  rm(df_sub,sub_name)                                 # Cleans up
}

## This gives the number and coverage of the words from this step
```

```
wl_group3_count <- wl_group3 %>%                          # Imports the words from Step 3
  anti_join(wl_all_corpora, by = "word") %>%              # Removes words also in Step 1
  anti_join(wl_group2, by = c("word", "subject")) %>%     # Removes words also in Step 2
  group_by(subject) %>%                                   # Groups by subject
  summarize(words = n(),                                  # Calculates number of words and
            coverage = sum(percentage)) %>%               # coverage for each subject
  inner_join(subject_names,                               # Gets the long name and sorting
             by = c("subject" = "subject_name")) %>%      # information for each subject
  arrange(subject_sort) %>%                               # Puts the subjects in order
  select(subject, words, coverage)                        # Selects the necessary columns

## Creates the final IS-AVL
wl_final <- data.frame()                                   # Creates an empty df
for(sub_name in corpus_list){                              # Loops over all subjects and
  df_wl1 <- wl_group1 %>% filter(subject == sub_name)      # Selects words from Step 1
  df_wl2 <- wl_group2 %>% filter(subject == sub_name)      # Step 2
  df_wl3 <- wl_group3 %>% filter(subject == sub_name)      # Step 3
  wl_final <- bind_rows(wl_final, df_wl1,                  # Combines these dfs into one
                        df_wl2, df_wl3)
  }
wl_final <- wl_final %>%                                  # Loads final list of words
  group_by(subject) %>%                                   # Groups by subject
  distinct(word, .keep_all = TRUE) %>%                    # Removes all duplicate
  ungroup()                                               # Ungroups
  rm(df_wl1, df_wl2, df_wl3, sub_name)                    # Cleans up

## Gets the final frequency counts for the words
freq_counts <- wl_final %>%                               # Loads final list of words
  group_by(subject) %>%                                   # Groups by subject
  summarize(words = n(),                                  # Gets # of words and coverage
            coverage  = sum(percentage)) %>%              # for each subject
  inner_join(subject_names,                               # Gets the long name and sorting
             by = c("subject" = "subject_name")) %>%      # information for each subject
  arrange(subject_sort) %>%                               # Puts the subjects in order
  select(subject, words, coverage)                        # Selects the necessary columns

## Gets the final frequency counts for the word families
freq_family <- wl_final %>%                               # Loads final list of words
  select(subject, word) %>%                               # Selects the sub and word columns
  inner_join(list_bnc) %>%                                # Gets the BNC/COCA Information
  group_by(subject) %>%                                   # Groups by subject
  distinct(head) %>%                                      # Finds all the unique head words
  summarize(families = n())                               # Counts the # of word families

## Gets the final frequency counts including gsl
gsl_freq <- df_words %>%                                  # Loads the corpus dataframe
  group_by(subject) %>%                                   # Groups by subject
  left_join(list_bnc) %>%                                 # Adds column for marginal words
```

```
filter(!list %in% c("word_marginal")) %>%          # removes marginal words
count(word) %>%                                    # Counts # of times word appears
arrange(desc(n)) %>%                               # Arranges in descending order
ungroup() %>%                                      # Ungroups
get_gsl_cov() %>%                                  # Sends to get_gsl_cov function
group_by(subject) %>%                              # Groups by subject
summarize(words = n(),                             # Gets coverage for each subject
          coverage = sum(percentage)) %>%
ungroup() %>%                                    # Ungroups
drop_na()                                        # Removes freq of GSL not in corpus
```

The R script for compiling the IS-AVL word lists in Chapter 6:

```
rm(list=ls(all=TRUE))               # clears the memory
library(rstudioapi)                 # Loads the rstudioapi - to get current directory
library(tidyverse)                  # Loads the tidyverse library

## Gets the directory the script is in so that it is possible
## to determine all the other directories
main_dir <- getActiveDocumentContext()$path %>%
  str_remove(., "/Scripts/.*")

## Regular expression patterns that are used in the processing of the text
pattern_sub <- "[:upper:]{2}"
pattern_word <- "[:lower:]*"
pattern_pos <- "(?<=_)[:lower:]*"

## The percentage of words per million used as a cutoff point for frequency for
## inclusion into the corpus
in_discipline <- 28.57/1000000

## Word lists used in the processing of the corpus.
## This 1st list allows for familiarizing the lemmas.
wl_fam <- read.csv(paste0(main_dir, "/Lists/bnc_coca.csv")) %>%   # Imports BNC/COCA
  mutate(head = na_if(head, "")) %>% fill(head) %>%               # Gets head words
  mutate(word = tolower(word))  %>%                               # Converts to lower
  select(word, head, list)                                        # Selects columns

## This second list is used to identify high-frequency words.
wl_freq <- read.csv(paste0(main_dir, "/Lists/bnc_coca.csv")) %>%   # Imports BNC/COCA
  mutate(word = tolower(word))  %>%                                # Converts to lower
  mutate(word = trimws(word)) %>%                                  # Removes white space
  filter(list == "K01") %>%                                        # Removes all non-1K
  select(word)                                                     # Selects column

## This sets the subject names so that they can be saved to a table later
list_subjects <- data.frame(
  subject = c("EM", "EC", "SS", "BM", "CM", "PM", "MM", "TK"),
  subject_sort = c("1_LM", "2_EC", "3_SS", "4_BM", "5_CM", "6_PM", "7_MM", "8_TK"),
  subject_name = c("Literature", "Economics", "Social Studies", "Biology",
                   "Chemistry", "Physics", "Maths", "TOK")
)

## This is where all the tables and word lists will be saved for use later
table_directory <- paste0(main_dir, "/Tables and Figures")  # A directory for tables
stage5 <- paste0(main_dir, "/_Step 5 Lemma Frequencies")    # A directory for the lists

## This identifies all the corpus files are stored and gets a list of files
directory <- paste0(main_dir, "/_Step 4 Lemmatized Corpus")  # Gets the directory
```

```
files <- list.files(directory, pattern = ".*.csv")          # Gets a list of files

## This creates the initial corpus data frame
corpus_df <- data.frame()                                   # Sets up empty data frame
for(file in files) {                                        # Loops over all the files
  file_to_read <- paste0(directory, "/", file)              # Gets the location of file
  df_file <- read.csv(file_to_read) %>%                     # Imports the file to a df
    mutate(book = book_name,                                # Gets the book & sub names
           subject = str_extract(book_name, pattern_sub)) %>%
    select(book, subject, token, pos)                       # Selects columns
  corpus_df <- bind_rows(corpus_df, df_file)                # Adds the df to corpus df
  rm(df_file)                                               # Cleans up
}

corpus_df <- corpus_df %>%
  mutate(word = str_extract(token, pattern_word)) %>%       # Gets the word from token
           semi_join(wl_fam)                                # Adds the frequency bands

## This calculates the frequency within the subjects
freq_disc <- corpus_df %>%                                  # Imports the corpus df
  group_by(subject) %>%                                     # Groups by subject
  count(token) %>%                                          # Counts how many times a tokens
appear
  mutate(perc = n/sum(n)) %>%                               # Calculates the % of total tokens
  ungroup() %>%                                             # Ungroups
  mutate(pos = str_extract(token, pattern_pos))             # Gets the part of speech

## Gets the total number of books in the corpus - used for range
number_of_books <- corpus_df %>%                            # Imports the corpus df
  group_by(subject) %>%                                     # Groups by subject
  distinct(book) %>%                                        # Gets a list of the books
  summarise(book = book, number = n()) %>%                  # Counts how many books
  select(subject, number) %>%                               # Selects the important
columns
  distinct(subject, .keep_all = T)                          # Gets rid of any duplicates

subject_list <- corpus_df$subject %>%                       # Gets a list of the subjects
  unique()

##The following code contains the six steps described in Chapter Six that I used
##to extract words for the IS-AVL

## Step 1: Remove lemmas with a frequency lower than 28.57 pm in each subcorpora
step_1 <- freq_disc %>%                                     # Imports the df with word frequencies
  filter(perc > in_discipline) %>%                          # Removes any words less than 28.57
  arrange(desc(perc))                                       # Sorts in descending order

step_1_coverage <- step_1 %>%                               # Imports the words from the step above
```

```
  group_by(subject) %>%                                    # Groups by subject
  summarize(words = n(), step_1_coverage = sum(perc)) %>%  # Gets coverage by subject
  inner_join(list_subjects) %>%                            # Gets list of subjects
  ungroup() %>%                                            # Ungroups
  arrange(subject_sort) %>%                                # Arranges by subject
  select(subject_name, words, step_1_coverage)             # Selects columns

## Step 2: Remove any words that are not nouns/verb/adj/adv
step_2 <- step_1 %>%                                      # Gets the df from Step 1
  filter(pos %in% c("noun", "verb", "adj", "adv"))        # Removes all function words

step_2_coverage <- step_2 %>%                            # Imports the words from Step 2
  group_by(subject) %>%                                  # Groups by subject
  summarize(lemmas = n(), step_2_coverage = sum(perc)) %>%   # Gets coverage for subject
  inner_join(list_subjects) %>%                             # Gets list of subjects
  ungroup() %>%                                             # Ungroups
  arrange(subject_sort) %>%                                 # Arranges by subject
  select(subject_name, lemmas, step_2_coverage)             # Selects the columns

step_2_coverage <- inner_join(step_1_coverage,        # Combines Step 1 & 2 coverage
                              step_2_coverage, by = "subject_name")

## Writes the tables to csv files for later use
write_csv(step_2_coverage,
          paste0(table_directory,"/", "table_6.1_lemma_coverage_step_2.csv"))

rm(step_1, step_1_coverage)                               # Cleans up

## Step 3: Remove any lemmas that appear in less than 50% of the books in that subject
text_range <- corpus_df %>%                          # Imports the corpus df
  group_by(subject, token) %>%                       # Groups by subject, then lemma
  summarize(freq = n(), n = n_distinct(book)) %>%    # Counts # of books each lemma is in
  inner_join(number_of_books) %>%                    # Imports the total number of books
  mutate(range = n / number) %>%                     # Calculates the range for lemmas
  filter(range > .50) %>%                            # Removes lemmas with a range < 50%
  mutate(step_3 = paste0(subject, "_", token))       # Creates a unified column

step_3 <- step_2 %>%                                 # Imports the stop 2 dataframe
  mutate(step_3 = paste0(subject, "_", token)) %>% # Creates a unified column
  semi_join(text_range, by = "step_3") %>%           # Removes Step 2 words not in Step 3
  select(-step_3) %>%                                # Removes the unified column
  mutate(lemma = str_extract(token,                  # Creates a column of lemmas
                             pattern_word)) %>%
  ungroup()                                          # Ungroups

step_3_coverage <- step_3 %>%                        # Imports df of words from Step 3
  group_by(subject) %>%                              # Group by subject
  summarize(lemmas = n(), coverage = sum(perc)) %>%  # Gets the # and coverage of lemmas
```

```
  inner_join(list_subjects) %>%                    # Gets the long subject name
  arrange(subject_sort) %>%                        # Sorts subjects
  select(subject_name, lemmas, coverage) %>%       # Selects the relevant columns
  ungroup()                                        # Ungroups

rm (text_range)                                    # Cleans up

## Step 4: Dispersion
## Note: Calculating the DP of the lemmas takes a very long time so it was done
## using a separate script and then imported into this workflow. That script is
## available in the appendix

dp_files <- dir(                                   # Gets the directory of DP files
  paste0(main_dir, "/_Step 5 DP Lists"),           # Gets a list of all the files
  full.names = T                                   # Includes directory information
)

dp_lists <- data.frame()                           # Creates and empty dataframe
for(file in dp_files){                             # Loops over files in DP directory
  dp_list <- read.csv(file) %>%                    # Imports each file into a df
    mutate(dp_check = 1 - dp_value) %>%            # Creates the correct DP value
    rename(token = lemma) %>%                      # Renames the lemma column
    filter(dp_check > .7)                          # Removes lemmas with a DP < .7
  dp_lists <- bind_rows(dp_lists, dp_list)         # Adds the df to the final DP df
  rm(dp_list)                                      # Cleans up
}

step_4 <- step_3 %>%                               # Imports the word list from Step 3
  semi_join(dp_lists, by = c("subject", "token"))  # Removes words not in the DP list

step_4_coverage <- step_4 %>%                      # Imports the df of words from Step
4
  group_by(subject) %>%                            # Group by subject
  summarize(lemmas = n(), coverage = sum(perc)) %>%  # Gets the # and cov of lemmas
  inner_join(list_subjects) %>%                    # Gets the long subject
  arrange(subject_sort) %>%                        # Sorts subjects
  select(subject_name, lemmas, coverage) %>%       # Selects the relevant columns
  ungroup()                                        # Ungroups

## Writes the tables to csv files for later use
write_csv(step_4_coverage, paste0(table_directory,"/",
                                  "table_6.2_lemma_coverage_step_4.csv"))

rm(dp_lists)                                       # Cleans up

## Step 5: Range Ration
## A word must appear at more than 20% of its expected frequency in more than 50%
## of the texts in that subcorpus
```

```
freq_books <- corpus_df %>%                  # Imports the corpus df
  group_by(book) %>%                         # Groups by book
  count(token) %>%                           # Counts times lemmas appears in each book
  mutate(perc = n/sum(n)) %>%                # Gets frequency of each lemma per book
  ungroup() %>%                                          # Ungroups
  mutate(pos = str_extract(token, pattern_pos)) %>%      # Gets the part of speech
  mutate(subject = str_extract(book, pattern_sub)) %>%   # Gets the subject
  rename(perc_book = perc) %>%                           # Renames the column
  inner_join(freq_disc, by = c("subject", "token")) %>%  # Gets freq for each subject
  select(subject, book, token, perc_book, perc) %>%      # Selects the columns
  mutate(range_ratio = perc_book/perc) %>%          # Determines range ratio per book
  filter(range_ratio >= .2) %>%                     # Removes lemmas/book freq < 20%
  group_by(subject, token) %>%                      # Groups by subject and lemma
  summarise(number = n()) %>%                       # Counts books with > 20% freq.
  inner_join(number_of_books, by = "subject") %>%        # Gets total number of books
  mutate(range_books = number.x/number.y) %>%            # Calculates % of total > 20%
  filter(range_books >= .5) %>%                          # Removes range less than 50%
  ungroup()                                              # Ungroups

step_5 <- step_4 %>%                                     # Gets lemmas from Step 4
  semi_join(freq_books, by = c("subject", "token"))      # Removes lems not in Step 5

step_5_coverage <- step_5 %>%                            # Imports lemmas from Step 4
  group_by(subject) %>%                             # Group by subject
  summarize(lemmas = n(), coverage = sum(perc)) %>%   # Gets the # and cov of lemmas
  inner_join(list_subjects) %>%                     # Gets the long subject name
  arrange(subject_sort) %>%                         # Sorts subjects
  select(subject_name, lemmas, coverage) %>%        # Selects the relevant columns
  ungroup()                                         # Ungroups

rm(freq_books)                                      # Cleans up

## Writes the table to csv files for later use
write_csv(step_5_coverage, paste0(table_directory,"/",
                                  "table_6.3_lemma_coverage_step_5.csv"))


## Step 6: Remove the most frequent 1,000 words from the BNC/COCA
## While this is different from Green and Lambert(2015) it produces a more
## academic word list. It also adds any high frequency academic words back into
## the word lists
final_df <- step_5  %>%                                  # Gets lemmas from Step 5
  mutate(word = str_extract(token, pattern_word)) %>%    # Changes token to a lemma
  inner_join(wl_fam) %>%                                 # Gets marginal word info
  anti_join(wl_freq) %>%                                 # Removes BNC/COCA 1K words
  filter(!list %in% c("word_marginal", "word_compound",  # Removes marginal words
                      "word_abbreviations", "proper_nouns"))
```

```
high_freq_academic <- read.csv(paste0(main_dir,              # High freq acad. words
                                      "/Lists/hf_academic.csv")) %>%
  select(subject, token) %>%                                 # Selects subject and token
  inner_join(step_5, by = c("subject", "token"))             # Gets word frequency data

final_df <- bind_rows(final_df, high_freq_academic)          # Adds high freq academic

final_coverage<- final_df %>%                                 # Imports the df from Final Step
  group_by(subject) %>%                                      # Group by subject
  summarize(lemmas = n(), coverage = sum(perc)) %>%   # Gets the num and coverage of
lemmas
  inner_join(list_subjects) %>%                             # Gets the long sub name and order
  arrange(subject_sort) %>%                                 # Sorts subjects in correct order
  select(subject_name, lemmas, coverage) %>%                # Selects the relevant columns
  ungroup()                                                 # Ungroups

## Writes the tables to csv files for later use
write_csv(final_coverage, paste0(table_directory,"/", "table_6.4_final.csv"))
```

The script used to calculate dispersion in Chapter 5

```
rm(list=ls(all=TRUE)) # clear memory
library(tidyverse)

all.corpus.words <- c() # creates a df for words
all.corpus.files <- c() # creates a df for split corpus name

## Gets the directory the script is in so that it is possible
## to determine all the other directories
main_dir <- getActiveDocumentContext()$path %>%
  str_remove(., "/Scripts/.*")

## This is where all the tables and word lists will be saved for use later
table_directory <- paste0(main_dir, "/Tables and Figures") # A directory for tables

## This identifies all the corpus files are stored and gets a list of files
directory <- paste0(main_dir, "/_Step 4 Lemmatized Corpus") # Gets the directory
files <- list.files(directory, pattern = ".*.csv") # Gets a list of files
save_files <- paste0(main_dir, "/_Step 5 DP Lists/") # sets save directory

## Function for splitting the corpus into equal parts
## It takes 2 arguments, the corpus df and the number of parts
## If no number is given, the default is 10 based on Biber et al. 2016
get_split <- function(df, parts = 10){
  split_df <-  split(df, rep(1:parts, length.out = nrow(df),
                             each = ceiling(nrow(df)/parts)))
  df_split <- data.frame()
  for(num in 1:parts){
    df_part <- split_df[[num]] %>%
      mutate(part = num)
    df_split <- bind_rows(df_split, df_part)
  }
return(df_split)
}

## Function for getting the dispersion
## It takes 2 values
## 1. The subject df: There need to be 2 columns, token and part
## 2. Name of the subject, this is used to create a subject column in the df
get_disp <- function(df, sub) {
  subject_dp <- data.frame()
  all.words <- df$token
  all.files <- df$part
  list.lemmas <- unique(all.words)
  number_left <- length(list.lemmas)
  for(lemma in list.lemmas){
```

```
    wheres.the.word <-            # make wheres.the.word
      table(                      # the table with
        all.files,                # the files in the rows and
        all.words==lemma)         # whether the lemma or not in the columns

    obs.perc <-                         # make obs.perc the result of
      wheres.the.word[,"TRUE"] /        # dividing the freq of "staining" per file by
      sum(wheres.the.word[,"TRUE"])     # the frequency of "staining"

    exp.perc <-                   # make exp.perc the result of
      rowSums(wheres.the.word) /  # dividing the sizes of the files in words by
      sum(wheres.the.word)        # the corpus size

    dp <- sum(abs(obs.perc - exp.perc)) / 2 # calculates dp

    lemma_dp <- data.frame(               # creates a df with
      subject = sub,                      # subject name
      lemma = lemma,                      # lemma
      dp_value = dp                       # DP
    )
    subject_dp <- bind_rows(subject_dp, lemma_dp) # add to final subject df
    number_left = number_left - 1                 # calculates number left
  }
  write.csv(subject_dp, paste0(save_files, sub, "_dp_df.csv"))
  return(subject_dp)
}

corpus_df <- data.frame()   # creates empty df
files <- dir(
    directory,              # creates a list of all the files in the directory
    full.names=TRUE)        # and retain the complete paths to the files

for(file in files) {                # imports lemmatized corpus
  df_file <- read.csv(file)         # reads each file
  corpus_df <- bind_rows(           # adds file to final df
    corpus_df,
    df_file
  )
  rm(df_file)                       # cleans up
}

all_corpus_subjects <- corpus_df %>%       # creates a list of subjects
  mutate(subject = str_extract(book_name,  # gets the subject name
  "[:upper:]{2}")) %>%
  distinct(subject) %>%                    # removes duplicates
  pull(subject)                            # gets subjects

corpus_split <- data.frame()               # splits each subject into parts
```

```
for(sub in all_corpus_subjects){           # gets list of subjects
  df_split_sub <- corpus_df %>%            # gets a split df for each sub
    mutate(subject = str_extract(book_name, # gets subject name
    "[:upper:]{2}")) %>%
    filter(subject == sub)                 # selects one subject
  df_split_sub <- get_split(df_split_sub) %>% # creates split df
    select(subject, part, token)           # selects columns
  corpus_split <- bind_rows(corpus_split,  # creates final df
  df_split_sub)
  rm(df_split_sub)                         # cleans up
}
rm(corpus_df)                              # cleans up

corpus_dp <- data.frame()                 # gets dp value
for(sub in all_corpus_subjects){          # gets list of subjects
  df_dp_sub <- corpus_split %>%           # imports split df
    filter(subject == sub)                # gets 1 subject
  subject_dp <- get_disp(df_dp_sub, sub)  # gets dp value for subject
  corpus_dp <- bind_rows(                 # adds subject dp to
    corpus_dp,                            # final corpus dp df
    subject_dp
  )
  rm(df_dp_sub, subject_dp)               # cleans up
}
```

Appendix E: The IS-AVLs by Discipline

The IS-AVL for Literature (610 words) Frequency in Tokens Per Millon (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	text.n	5290	31	focus.v	575	61	refer.v	360
2	language.n	2489	32	interpretation.n	543	62	paragraph.n	351
3	poem.n	1969	33	identity.n	536	63	narrative.n	344
4	literary.adj	1795	34	discuss.v	526	64	process.n	339
5	example.n	1281	35	purpose.n	522	65	connect.v	334
6	character.n	1165	36	knowledge.n	508	66	role.n	330
7	culture.n	1067	37	style.n	504	67	similar.adj	322
8	literature.n	1043	38	event.n	479	68	communication.n	320
9	image.n	984	39	represent.v	470	69	scene.n	312
10	novel.n	963	40	sentence.n	467	70	reveal.v	308
11	create.v	932	41	reflect.v	463	71	final.adj	305
12	explore.v	913	42	develop.v	457	72	reference.n	305
13	author.n	911	43	social.adj	457	73	area.n	303
14	include.v	903	44	connection.n	451	74	graphic.adj	301
15	issue.n	871	45	poet.n	436	75	tone.n	299
16	feature.n	837	46	compare.v	424	76	aspect.n	299
17	context.n	832	47	topic.n	413	77	convey.v	293
18	chapter.n	827	48	interpret.v	402	78	identify.v	290
19	audience.n	798	49	section.n	402	79	publish.v	289
20	non.adj	778	50	skill.n	399	80	century.n	288
21	perspective.n	721	51	society.n	397	81	translation.n	276
22	poetry.n	717	52	symbol.n	394	82	detail.n	272
23	structure.n	713	53	theme.n	391	83	impact.n	270
24	fiction.n	664	54	specific.adj	387	84	drama.n	268
25	effect.n	626	55	concept.n	385	85	approach.n	262
26	global.adj	623	56	provide.v	385	86	common.adj	262
27	describe.v	621	57	visual.adj	382	87	technique.n	262
28	response.n	604	58	affect.v	380	88	focus.n	261
29	narrator.n	595	59	value.n	380	89	device.n	260
30	extract.n	579	60	political.adj	362	90	guide.v	257

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	attitude.n	254	121	period.n	195	151	involve.v	167
92	genre.n	254	122	violence.n	195	152	significant.adj	164
93	original.adj	254	123	recognize.v	192	153	influence.v	163
94	gender.n	253	124	attention.n	192	154	opportunity.n	163
95	various.adj	247	125	famous.adj	191	155	remain.v	163
96	community.n	246	126	translate.v	191	156	stylistic.adj	163
97	representation.n	243	127	relevant.adj	190	157	memory.n	162
98	encourage.v	240	128	version.n	189	158	plot.n	161
99	communicate.v	236	129	intend.v	188	159	category.n	161
100	background.n	233	130	introduction.n	187	160	complex.adj	161
101	apply.v	230	131	transform.v	187	161	significance.n	161
102	belief.n	228	132	likely.adj	186	162	link.n	159
103	engage.v	225	133	approach.v	185	163	produce.v	159
104	phrase.n	222	134	comment.n	180	164	object.n	158
105	exploration.n	220	135	contain.v	180	165	politic.n	155
106	metaphor.n	219	136	directly.adv	179	166	gain.v	154
107	discussion.n	217	137	respond.v	179	167	conclusion.n	153
108	therefore.adv	216	138	pattern.n	177	168	formal.adj	152
109	content.n	210	139	critical.adj	176	169	direct.adj	151
110	exist.v	210	140	receive.v	176	170	multiple.adj	151
111	define.v	209	141	stanza.n	175	171	allusion.n	150
112	description.n	209	142	emotion.n	174	172	dramatic.adj	150
113	require.v	209	143	imply.v	173	173	dialogue.n	149
114	individual.n	207	144	associate.v	172	174	performance.n	149
115	strategy.n	206	145	challenge.n	172	175	creative.adj	148
116	range.n	205	146	job.n	172	176	broad.adj	147
117	argue.v	201	147	title.n	172	177	prose.n	146
118	imagery.n	200	148	conflict.n	171	178	determine.v	145
119	familiar.adj	199	149	quality.n	171	179	physical.adj	145
120	examine.v	198	150	contrast.n	168	180	appeal.n	144

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	function.n	144	211	link.v	124	241	development.n	104
182	emotional.adj	142	212	popular.adj	124	242	authority.n	102
183	transformation.n	142	213	extend.v	123	243	effort.n	102
184	traditional.adj	140	214	finally.adv	123	244	assume.v	100
185	achieve.v	137	215	mood.n	123	245	inspire.v	100
186	previous.adj	136	216	critic.n	122	246	length.n	100
187	repeat.v	136	217	protagonist.n	121	247	practise.v	100
188	direction.n	135	218	verse.n	121	248	belong.v	99
189	lack.n	135	219	justice.n	120	249	negative.adj	98
190	rhyme.n	135	220	religious.adj	120	250	obvious.adj	98
191	decision.n	134	221	similarity.n	120	251	struggle.n	98
192	aware.adj	133	222	aim.v	119	252	violent.adj	98
193	playwright.n	133	223	encounter.v	119	253	attempt.n	97
194	future.n	131	224	contrast.v	118	254	narrative.adj	97
195	condition.n	130	225	series.n	118	255	capture.v	96
196	suffer.v	130	226	awareness.n	116	256	repetition.n	96
197	immediately.adv	130	227	importance.n	116	257	advice.n	94
198	perform.v	130	228	standard.adj	116	258	religion.n	94
199	tool.n	130	229	creativity.n	115	259	vocabulary.n	94
200	appreciate.v	129	230	establish.v	115	260	environment.n	93
201	intention.n	129	231	frame.n	115	261	contextual.adj	92
202	highlight.v	128	232	occur.v	115	262	material.n	92
203	prose.v	128	233	remind.v	113	263	status.n	91
204	thus.adv	128	234	effectively.adv	112	264	tension.n	91
205	contribute.v	127	235	structural.adj	112	265	originally.adv	90
206	influence.n	126	236	theatre.n	112	266	poetic.adj	90
207	result.n	126	237	feature.v	110	267	tale.n	89
208	seek.v	126	238	method.n	106	268	loss.n	88
209	introduce.v	125	239	lyric.n	105	269	accord.v	87
210	journey.n	124	240	peace.n	105	270	account.n	86

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	desire.n	86	301	implication.n	75	331	surround.v	68
272	reaction.n	86	302	aim.n	74	332	citizen.n	67
273	avoid.v	85	303	claim.v	74	333	escape.v	67
274	deliver.v	85	304	myth.n	74	334	expose.v	67
275	entirely.adv	85	305	struggle.v	74	335	murder.n	67
276	fictional.adj	84	306	column.n	73	336	objective.n	67
277	figurative.adj	84	307	eventually.adv	72	337	obviously.adv	67
278	international.adj	84	308	former.adj	72	338	specifically.adv	67
279	unique.adj	84	309	hero.n	72	339	village.n	67
280	design.v	83	310	invite.v	72	340	debate.n	66
281	fail.v	83	311	spread.v	72	341	essential.adj	66
282	indicate.v	83	312	successful.adj	72	342	oppose.v	66
283	ensure.v	82	313	characteristic.n	71	343	shift.v	66
284	manner.n	82	314	irony.n	71	344	circumstance.n	65
285	shift.n	82	315	literal.adj	71	345	combine.v	65
286	entire.adj	81	316	creature.n	70	346	internal.adj	65
287	strength.n	80	317	private.adj	70	347	literally.adv	65
288	commit.v	79	318	wealth.n	70	348	success.n	65
289	connotation.n	79	319	alive.adj	69	349	vary.v	65
290	edit.v	78	320	ancient.adj	69	350	cite.v	64
291	frequently.adv	78	321	merely.adv	69	351	consist.v	64
292	organize.v	78	322	portray.v	69	352	detailed.adj	64
293	rely.v	78	323	respect.n	69	353	faith.n	64
294	observe.v	77	324	behaviour.n	69	354	generate.v	64
295	access.n	76	325	couple.n	69	355	director.n	63
296	differ.v	76	326	novel.adj	69	356	duty.n	63
297	structure.v	76	327	similarly.adv	69	357	promote.v	63
298	ultimately.adv	76	328	distance.n	68	358	prove.v	63
299	university.n	76	329	maintain.v	68	359	destroy.v	62
300	accurate.adj	75	330	soul.n	68	360	assumption.n	61

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	available.adj	61	391	mental.adj	55	421	weak.adj	50
362	divide.v	61	392	search.v	55	422	decade.n	49
363	equal.adj	61	393	tongue.n	55	423	evoke.v	49
364	evolve.v	61	394	abstract.adj	54	424	excellent.adj	49
365	humour.n	61	395	anger.n	54	425	format.n	49
366	romantic.adj	61	396	appreciation.n	54	426	furthermore.adv	49
367	creation.n	60	397	current.adj	54	427	reject.v	49
368	exercise.n	60	398	innocent.adj	54	428	relevance.n	49
369	major.adj	60	399	acknowledge.v	54	429	result.v	49
370	pleasure.n	60	400	justify.v	54	430	team.n	49
371	combination.n	59	401	perceive.v	54	431	alliteration.n	48
372	path.n	59	402	property.n	54	432	attend.v	48
373	brief.adj	58	403	active.adj	53	433	circle.n	48
374	enemy.n	58	404	intellectual.adj	53	434	impression.n	48
375	remove.v	58	405	narrow.adj	53	435	regard.n	48
376	review.v	58	406	regular.adj	53	436	stress.v	48
377	adult.n	57	407	surface.n	53	437	tiny.adj	48
378	battle.n	57	408	association.n	52	438	accompany.v	47
379	increase.v	57	409	direct.v	52	439	lesson.n	47
380	mirror.n	57	410	foreign.adj	52	440	survive.v	47
381	punctuation.n	57	411	mystery.n	52	441	distant.adj	46
382	celebrate.v	56	412	basis.n	51	442	document.n	46
383	equally.adv	56	413	civil.adj	51	443	equality.n	46
384	presence.n	56	414	enhance.v	51	444	fruit.n	46
385	progress.n	56	415	adopt.v	50	445	humorous.adj	46
386	regardless.adv	56	416	blank.adj	50	446	manipulate.v	46
387	scheme.n	56	417	cloud.n	50	447	objective.adj	46
388	consciousness.n	55	418	overview.n	50	448	root.n	46
389	distinct.adj	55	419	separate.adj	50	449	scale.n	46
390	linguistic.adj	55	420	varied.adj	50	450	bold.adj	45

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	commonly.adv	45	481	ethical.adj	41	511	otherwise.adv	39
452	display.v	45	482	extremely.adv	41	512	potential.adj	39
453	monologue.n	45	483	failure.n	41	513	preserve.v	39
454	overall.adj	45	484	inequality.n	41	514	shock.n	39
455	professor.n	45	485	overcome.v	41	515	value.v	39
456	recognise.v	45	486	punishment.n	41	516	warning.n	39
457	subtle.adj	45	487	screen.n	41	517	youth.n	39
458	thin.adj	45	488	sharp.adj	41	518	alter.v	38
459	assignment.n	44	489	award.v	40	519	bind.v	38
460	critically.adv	44	490	borrow.v	40	520	creator.n	38
461	deny.v	44	491	confront.v	40	521	custom.n	38
462	false.adj	44	492	conscious.adj	40	522	grief.n	38
463	future.adj	44	493	convince.v	40	523	satire.n	38
464	opposite.n	44	494	earn.v	40	524	climate.n	38
465	previously.adv	44	495	fan.n	40	525	complicated.adj	38
466	reduce.v	44	496	ignore.v	40	526	computer.n	38
467	separate.v	44	497	juxtaposition.n	40	527	inspiration.n	38
468	shame.n	44	498	oppression.n	40	528	location.n	38
469	limit.n	43	499	palm.n	40	529	platform.n	38
470	opposite.adj	43	500	risk.n	40	530	attack.v	37
471	replace.v	43	501	secret.adj	40	531	confuse.v	37
472	solve.v	43	502	threat.n	40	532	dare.v	37
473	destruction.n	42	503	correct.adj	39	533	declare.v	37
474	embrace.v	42	504	demand.v	39	534	indirect.adj	37
475	explicitly.adv	42	505	flame.n	39	535	outcome.n	37
476	lack.v	42	506	ideal.adj	39	536	reference.v	37
477	practical.adj	42	507	incident.n	39	537	sympathy.n	37
478	principle.n	42	508	independent.adj	39	538	terror.n	37
479	valuable.adj	42	509	item.n	39	539	extreme.adj	36
480	weapon.n	42	510	journalist.n	39	540	philosophical.adj	36

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
541	pride.n	36	565	sink.v	34	589	culturally.adv	30
542	recall.v	36	566	victory.n	34	590	exception.n	30
543	recommend.v	36	567	conjunction.n	33	591	fault.n	30
544	sin.n	36	568	deserve.v	33	592	forth.adv	30
545	vast.adj	36	569	due.adj	33	593	fulfil.v	30
546	capable.adj	35	570	inevitably.adv	33	594	fundamental.adj	30
547	mix.v	35	571	possess.v	33	595	intense.adj	30
548	murder.v	35	572	sake.n	33	596	library.n	30
549	numerous.adj	35	573	disaster.n	32	597	mission.n	30
550	opposition.n	35	574	evident.adj	32	598	motivation.n	30
551	regime.n	35	575	match.v	32	599	purely.adv	30
552	trace.v	35	576	presumably.adv	32	600	region.n	30
553	wise.adj	35	577	pure.adj	32	601	shade.n	30
554	angel.n	34	578	rage.n	32	602	admire.v	29
555	blame.v	34	579	security.n	32	603	confident.adj	29
556	cheek.n	34	580	essence.n	31	604	defend.v	29
557	comedy.n	34	581	motion.n	31	605	disappear.v	29
558	commitment.n	34	582	row.n	31	606	fantasy.n	29
559	guest.n	34	583	tragic.adj	31	607	locate.v	29
560	margin.n	34	584	accurately.adv	30	608	mirror.v	29
561	pace.n	34	585	bridge.n	30	609	psychological.adj	29
562	physically.adv	34	586	cast.v	30	610	sorrow.n	29
563	refuse.v	34	587	chain.n	30			
564	shine.v	34	588	command.n	30			

The IS-AVL for Economics (481 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	firm.n	3443	31	involve.v	742	61	impact.n	489
2	example.n	2480	32	likely.adj	732	62	receive.v	482
3	increase.v	2324	33	refer.v	709	63	efficiency.n	481
4	production.n	2201	34	source.n	695	64	expenditure.n	480
5	consumer.n	2161	35	result.n	686	65	identify.v	477
6	profit.n	2148	36	concept.n	678	66	analysis.n	474
7	output.n	1679	37	area.n	665	67	operate.v	474
8	value.n	1587	38	opportunity.n	651	68	exist.v	466
9	produce.v	1414	39	flow.n	648	69	risk.n	455
10	resource.n	1399	40	average.adj	639	70	global.adj	442
11	include.v	1341	41	create.v	630	71	private.adj	441
12	revenue.n	1309	42	period.n	620	72	measure.v	439
13	reduce.v	1118	43	activity.n	616	73	environment.n	437
14	factor.n	1109	44	improve.v	612	74	scale.n	427
15	unit.n	1087	45	achieve.v	610	75	occur.v	424
16	decision.n	1041	46	international.adj	561	76	positive.adj	415
17	provide.v	1026	47	datum.n	557	77	discuss.v	412
18	develop.v	981	48	require.v	547	78	technology.n	411
19	social.adj	980	49	result.v	546	79	population.n	403
20	capital.n	979	50	issue.n	543	80	skill.n	397
21	benefit.n	959	51	wage.n	533	81	effective.adj	393
22	labour.n	930	52	potential.adj	522	82	compare.v	390
23	investment.n	861	53	affect.v	520	83	available.adj	378
24	industry.n	857	54	negative.adj	518	84	role.n	376
25	method.n	855	55	competitive.adj	513	85	material.n	370
26	financial.adj	820	56	define.v	512	86	disadvantage.n	369
27	objective.n	809	57	common.adj	508	87	ensure.v	347
28	advantage.n	782	58	external.adj	506	88	percentage.n	337
29	calculate.v	771	59	sector.n	505	89	major.adj	334
30	loss.n	752	60	non.adj	500	90	approach.n	332

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	direct.adj	332	121	limited.adj	264	151	overall.adj	199
92	determine.v	332	122	primary.adj	259	152	purchase.n	199
93	gain.v	329	123	individual.adj	257	153	future.n	196
94	individual.n	326	124	environmental.adj	254	154	response.n	196
95	earn.v	322	125	significant.adj	253	155	fund.n	195
96	focus.v	319	126	incentive.n	251	156	element.n	194
97	purchase.v	312	127	political.adj	250	157	introduce.v	192
98	influence.v	310	128	extent.n	249	158	directly.adv	190
99	appropriate.adj	308	129	property.n	244	159	assessment.n	189
100	condition.n	308	130	represent.v	238	160	maintain.v	189
101	standard.n	307	131	efficient.adj	237	161	expand.v	185
102	relatively.adv	306	132	physical.adj	237	162	indicate.v	185
103	apply.v	303	133	importance.n	236	163	assess.v	185
104	cycle.n	295	134	aim.v	234	164	due.adj	184
105	remain.v	295	135	consequence.n	231	165	event.n	184
106	specific.adj	291	136	final.adj	230	166	describe.v	183
107	accord.v	289	137	improvement.n	228	167	compete.v	182
108	evaluate.v	287	138	tool.n	225	168	sustainability.n	179
109	encourage.v	286	139	similar.adj	222	169	input.n	176
110	failure.n	286	140	component.n	222	170	attract.v	175
111	productivity.n	286	141	limit.v	220	171	basis.n	175
112	range.n	285	142	adopt.v	218	172	contribute.v	175
113	computer.n	283	143	consume.v	217	173	item.n	175
114	legal.adj	282	144	limitation.n	216	174	enable.v	174
115	argue.v	280	145	outline.v	216	175	evidence.n	174
116	outcome.n	279	146	lack.n	215	176	regulation.n	170
117	future.adj	277	147	benefit.v	213	177	security.n	166
118	promote.v	277	148	invest.v	213	178	prevent.v	165
119	examine.v	266	149	pressure.n	206	179	avoid.v	164
120	goal.n	264	150	trend.n	204	180	principle.n	163

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	raw.adj	161	211	respond.v	132	241	pursue.v	108
182	equity.n	160	212	effectively.adv	131	242	prefer.v	107
183	increasingly.adv	160	213	reflect.v	131	243	skilled.adj	107
184	initial.adj	159	214	hire.v	130	244	detail.n	106
185	community.n	158	215	allocate.v	128	245	facility.n	104
186	essential.adj	157	216	differ.v	128	246	average.n	103
187	account.v	156	217	solution.n	128	247	discussion.n	103
188	commercial.adj	154	218	productive.adj	125	248	status.n	103
189	vary.v	154	219	technological.adj	125	249	reference.n	102
190	fail.v	152	220	implement.v	124	250	predict.v	102
191	effort.n	151	221	influence.n	124	251	aware.adj	100
192	gain.n	151	222	mobile.adj	124	252	satisfy.v	100
193	select.v	151	223	reduction.n	123	253	conclusion.n	99
194	restaurant.n	148	224	fluctuation.n	121	254	alternative.adj	98
195	requirement.n	146	225	pattern.n	121	255	context.n	98
196	estimate.v	145	226	integration.n	120	256	progress.n	98
197	topic.n	145	227	region.n	120	257	similarly.adv	97
198	comparison.n	145	228	attempt.v	118	258	commonly.adv	96
199	sum.n	145	229	divide.v	118	259	replace.v	96
200	provision.n	142	230	strength.n	118	260	potential.n	95
201	constant.adj	142	231	characteristic.n	115	261	claim.v	93
202	relevant.adj	142	232	variety.n	115	262	distinguish.v	93
203	original.adj	141	233	calculation.n	114	263	perceive.v	93
204	energy.n	140	234	link.v	114	264	belief.n	93
205	rely.v	139	235	eliminate.v	113	265	effectiveness.n	93
206	design.v	139	236	manufacture.v	113	266	enhance.v	92
207	proportion.n	139	237	accurate.adj	112	267	combination.n	91
208	partner.n	136	238	regard.v	112	268	satisfaction.n	91
209	scheme.n	136	239	finance.v	111	269	fuel.n	90
210	associate.v	135	240	challenge.n	109	270	intend.v	90

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	trade.v	90	301	decline.v	78	331	grant.v	66
272	transaction.n	90	302	agency.n	78	332	regional.adj	66
273	sufficient.adj	89	303	category.n	77	333	remove.v	66
274	previous.adj	88	304	rapid.adj	77	334	typical.adj	66
275	contain.v	87	305	broad.adj	76	335	minimum.n	65
276	damage.n	86	306	combine.v	76	336	research.v	65
277	finally.adv	86	307	availability.n	75	337	selection.n	65
278	prove.v	86	308	exceed.v	75	338	survive.v	65
279	solve.v	86	309	existence.n	74	339	attempt.n	64
280	unlikely.adj	86	310	publish.v	74	340	correct.adj	64
281	entire.adj	85	311	engage.v	73	341	moral.adj	64
282	implication.n	85	312	famous.adj	73	342	decline.n	63
283	lack.v	85	313	rapidly.adv	73	343	procedure.n	63
284	ignore.v	84	314	eventually.adv	72	344	geographical.adj	63
285	adjust.v	84	315	monitor.v	72	345	series.n	63
286	core.n	84	316	efficiently.adv	71	346	observe.v	62
287	circumstance.n	83	317	maintenance.n	71	347	summary.n	62
288	consist.v	83	318	weakness.n	71	348	damage.v	61
289	direction.n	83	319	investigate.v	69	349	flight.n	61
290	justify.v	83	320	technical.adj	69	350	separate.adj	60
291	complex.adj	82	321	identical.adj	68	351	achievement.n	60
292	competitiveness.n	81	322	modern.adj	68	352	fund.v	60
293	suffer.v	81	323	significantly.adv	68	353	stability.n	60
294	acquire.v	81	324	attractive.adj	67	354	illegal.adj	59
295	application.n	81	325	graph.n	67	355	purchasing.n	58
296	extend.v	81	326	independent.adj	67	356	reveal.v	58
297	fee.n	80	327	priority.n	67	357	secure.v	58
298	provider.n	80	328	consistent.adj	66	358	initially.adv	57
299	comment.n	79	329	error.n	66	359	majority.n	57
300	limit.n	79	330	extreme.adj	66	360	presence.n	57

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	recommend.v	57	391	steel.n	47	421	automatically.adv	41
362	convert.v	56	392	valuable.adj	47	422	impact.v	41
363	accurately.adv	55	393	concentrate.v	47	423	specifically.adv	41
364	earning.n	55	394	transfer.v	47	424	extensive.adj	41
365	otherwise.adv	55	395	commitment.n	46	425	struggle.v	41
366	university.n	54	396	harm.v	46	426	adult.n	40
367	burden.n	54	397	tourist.n	46	427	discourage.v	40
368	match.v	54	398	vulnerable.adj	45	428	medical.adj	40
369	dominate.v	53	399	non.n	44	429	risky.adj	40
370	guarantee.v	52	400	threaten.v	44	430	dominant.adj	39
371	insufficient.adj	52	401	emerge.v	44	431	recommendation.n	39
372	legally.adv	52	402	equally.adv	44	432	scope.n	39
373	immediate.adj	51	403	fundamental.adj	44	433	advanced.adj	38
374	improved.adj	51	404	access.v	43	434	narrow.adj	38
375	approach.v	50	405	connection.n	43	435	highlight.v	38
376	extremely.adv	50	406	estimate.n	43	436	initiative.n	37
377	favour.n	50	407	manner.n	43	437	intellectual.adj	37
378	fulfil.v	50	408	previously.adv	43	438	prospect.n	37
379	version.n	50	409	succeed.v	43	439	revise.v	37
380	advance.n	50	410	alternatively.adv	42	440	challenge.v	36
381	frequently.adv	50	411	beneficial.adj	42	441	connect.v	36
382	propose.v	50	412	contrast.v	42	442	strict.adj	36
383	resolve.v	50	413	creative.adj	42	443	adequate.adj	35
384	search.n	50	414	distinct.adj	42	444	facilitate.v	35
385	flow.v	49	415	emphasis.n	42	445	favour.v	35
386	numerical.adj	49	416	habit.n	42	446	participate.v	35
387	satisfied.adj	49	417	length.n	42	447	convince.v	35
388	decade.n	48	418	multiply.v	42	448	emergency.n	35
389	desire.n	48	419	pose.v	42	449	meat.n	35
390	weak.adj	48	420	anticipate.v	41	450	pre.adj	35

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	reaction.n	35	462	disagreement.n	32	473	complicated.adj	30
452	conclude.v	34	463	origin.n	32	474	dynamic.adj	30
453	guide.v	34	464	originally.adv	32	475	excessive.adj	30
454	negatively.adv	34	465	inappropriate.adj	32	476	bind.v	29
455	opposite.n	34	466	precise.adj	32	477	capture.v	29
456	sustain.v	34	467	refuse.v	32	478	interact.v	29
457	familiar.adj	33	468	tackle.v	32	479	overcome.v	29
458	involvement.n	33	469	interaction.n	31	480	repeat.v	29
459	season.n	33	470	massive.adj	31	481	separate.v	29
460	translate.v	33	471	subjective.adj	31			
461	attend.v	32	472	withdraw.v	31			

The IS-AVL for Social Studies (526 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	political.adj	1957	31	operation.n	508	61	technology.n	362
2	economic.adj	1802	32	extent.n	490	62	process.n	361
3	military.adj	1678	33	remain.v	490	63	defeat.v	359
4	army.n	1395	34	trade.n	488	64	declare.v	355
5	troop.n	1223	35	enemy.n	479	65	threat.n	355
6	include.v	1126	36	resistance.n	473	66	decision.n	352
7	social.adj	1012	37	foreign.adj	472	67	affect.v	342
8	attack.n	916	38	western.adj	469	68	successful.adj	338
9	policy.n	834	39	soldier.n	467	69	influence.n	336
10	economy.n	770	40	factor.n	464	70	strength.n	336
11	area.n	750	41	unit.n	461	71	authority.n	335
12	major.adj	708	42	weapon.n	461	72	skill.n	333
13	population.n	679	43	supply.n	453	73	refer.v	330
14	increase.v	673	44	campaign.n	449	74	maintain.v	324
15	establish.v	662	45	significant.adj	428	75	thus.adv	324
16	develop.v	653	46	involve.v	423	76	defence.n	318
17	region.n	647	47	argument.n	422	77	material.n	315
18	role.n	635	48	success.n	414	78	treaty.n	315
19	create.v	633	49	destroy.v	412	79	defeat.n	313
20	century.n	632	50	effort.n	411	80	scale.n	312
21	result.n	622	51	gain.v	409	81	crisis.n	310
22	event.n	598	52	period.n	402	82	effective.adj	309
23	issue.n	598	53	fail.v	400	83	require.v	306
24	impact.n	593	54	independent.adj	399	84	alliance.n	304
25	achieve.v	566	55	practice.n	399	85	result.v	304
26	battle.n	565	56	prevent.v	388	86	value.n	304
27	provide.v	549	57	response.n	379	87	revolution.n	300
28	effect.n	540	58	produce.v	374	88	attempt.n	297
29	victory.n	528	59	military.n	371	89	anti.adj	294
30	territory.n	510	60	regime.n	370	90	occupy.v	291

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	ally.n	288	121	dominate.v	247	151	limit.v	198
92	divide.v	288	122	attempt.v	243	152	operate.v	198
93	focus.v	288	123	increasingly.adv	240	153	southern.adj	196
94	seek.v	288	124	failure.n	237	154	internal.adj	195
95	armed.adj	286	125	loss.n	234	155	threaten.v	193
96	aid.n	285	126	opportunity.n	233	156	method.n	192
97	advantage.n	283	127	analysis.n	230	157	receive.v	192
98	common.adj	282	128	conclusion.n	225	158	series.n	192
99	discuss.v	282	129	aspect.n	224	159	occupation.n	189
100	limited.adj	282	130	ideology.n	224	160	popular.adj	187
101	industry.n	278	131	ensure.v	219	161	purpose.n	187
102	civilian.n	277	132	compare.v	218	162	range.n	187
103	defend.v	275	133	democracy.n	216	163	available.adj	186
104	encourage.v	275	134	belief.n	215	164	expand.v	186
105	capital.n	271	135	basis.n	213	165	structure.n	184
106	debate.n	271	136	exist.v	213	166	traditional.adj	184
107	pressure.n	271	137	introduction.n	213	167	occur.v	180
108	importance.n	269	138	contribute.v	212	168	perspective.n	180
109	demand.n	268	139	former.adj	212	169	commit.v	178
110	evidence.n	268	140	section.n	212	170	determine.v	178
111	intervention.n	268	141	similar.adj	212	171	elect.v	177
112	struggle.n	266	142	outcome.n	208	172	weakness.n	177
113	condition.n	265	143	accord.v	207	173	advance.v	173
114	emerge.v	263	144	tension.n	205	174	avoid.v	173
115	various.adj	262	145	democratic.adj	204	175	negotiate.v	173
116	oppose.v	257	146	prove.v	204	176	organize.v	173
117	suffer.v	256	147	identify.v	202	177	design.v	170
118	therefore.adv	251	148	approach.n	199	178	claim.v	169
119	reduce.v	248	149	consequence.n	199	179	opinion.n	169
120	tactic.n	248	150	target.n	199	180	direct.adj	164

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	introduce.v	160	211	finally.adv	135	241	desire.n	114
182	modern.adj	160	212	refuse.v	134	242	extend.v	114
183	civilian.adj	158	213	adopt.v	132	243	spread.v	114
184	effectively.adv	158	214	remove.v	132	244	agricultural.adj	113
185	replace.v	158	215	future.adj	129	245	immediate.adj	113
186	context.n	157	216	destruction.n	128	246	contrast.v	111
187	improve.v	155	217	sustain.v	128	247	counter.n	111
188	withdraw.v	152	218	imperial.adj	126	248	weaken.v	111
189	involvement.n	151	219	status.n	126	249	initial.adj	110
190	product.n	151	220	conduct.v	125	250	vote.n	110
191	impose.v	149	221	contain.v	125	251	examine.v	108
192	supply.v	149	222	lack.v	125	252	grant.v	108
193	weak.adj	148	223	liberal.adj	125	253	affair.n	107
194	represent.v	146	224	overall.adj	125	254	document.n	107
195	apply.v	143	225	assess.v	123	255	essential.adj	107
196	focus.n	143	226	intend.v	123	256	participate.v	107
197	rely.v	143	227	resist.v	123	257	politically.adv	107
198	solution.n	143	228	critical.adj	122	258	potential.adj	107
199	domestic.adj	142	229	dictatorship.n	122	259	village.n	107
200	style.n	142	230	vital.adj	122	260	contrast.n	105
201	describe.v	140	231	directly.adv	120	261	restore.v	105
202	opponent.n	140	232	equal.adj	120	262	strengthen.v	105
203	progress.n	140	233	export.n	120	263	abandon.v	103
204	regard.v	140	234	vote.v	119	264	essentially.adv	103
205	relatively.adv	138	235	attitude.n	117	265	speech.n	103
206	territorial.adj	138	236	entire.adj	117	266	benefit.v	102
207	demonstrate.v	137	237	vast.adj	117	267	broad.adj	102
208	final.adj	137	238	demand.v	116	268	engage.v	102
209	financial.adj	137	239	establishment.n	116	269	exploit.v	102
210	diplomatic.adj	135	240	significance.n	116	270	surround.v	102

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	damage.n	100	301	currency.n	87	331	compete.v	73
272	ignore.v	100	302	formal.adj	87	332	deliver.v	73
273	private.adj	100	303	presence.n	87	333	enormous.adj	73
274	tradition.n	100	304	reject.v	87	334	equip.v	73
275	creation.n	99	305	succeed.v	87	335	image.n	73
276	pursue.v	99	306	import.n	85	336	ordinary.adj	73
277	respond.v	99	307	reinforce.v	85	337	specifically.adv	73
278	attention.n	97	308	wound.v	85	338	associate.v	72
279	benefit.n	97	309	capacity.n	84	339	indicate.v	72
280	rivalry.n	97	310	map.n	84	340	obtain.v	72
281	secure.v	97	311	route.n	84	341	boundary.n	70
282	politician.n	96	312	hostility.n	82	342	combination.n	70
283	account.n	94	313	instability.n	82	343	commitment.n	70
284	assessment.n	94	314	successfully.adv	82	344	deny.v	70
285	committee.n	94	315	superiority.n	82	345	deploy.v	70
286	decade.n	94	316	circumstance.n	81	346	expose.v	70
287	future.n	94	317	implement.v	81	347	intention.n	70
288	gain.n	93	318	subsequent.adj	81	348	authoritarian.adj	68
289	persuade.v	93	319	secret.adj	79	349	confirm.v	68
290	combine.v	91	320	autonomy.n	78	350	connect.v	68
291	fundamental.adj	91	321	loan.n	78	351	construct.v	68
292	immediately.adv	91	322	perceive.v	78	352	dominant.adj	68
293	issue.v	91	323	quality.n	78	353	expense.n	68
294	superior.adj	91	324	target.v	78	354	acquire.v	67
295	active.adj	90	325	variety.n	78	355	detail.n	67
296	previous.adj	90	326	convince.v	76	356	determination.n	67
297	relative.adj	90	327	favour.v	76	357	direct.v	67
298	estimate.v	88	328	prosperity.n	76	358	effectiveness.n	67
299	primary.adj	88	329	integrate.v	75	359	mass.adj	67
300	contribution.n	87	330	recover.v	75	360	conclude.v	65

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	guarantee.v	65	391	distribute.v	58	421	trade.v	52
362	permanent.adj	65	392	female.adj	58	422	unpopular.adj	52
363	potential.n	65	393	initiative.n	58	423	improvement.n	50
364	provoke.v	65	394	severe.adj	58	424	practise.v	50
365	reserve.n	65	395	victim.n	58	425	shift.v	50
366	era.n	64	396	withdrawal.n	58	426	aggression.n	49
367	gap.n	64	397	direction.n	56	427	discipline.n	49
368	message.n	64	398	eliminate.v	56	428	hostile.adj	49
369	moral.adj	64	399	exception.n	56	429	outline.v	49
370	officially.adv	64	400	inspire.v	56	430	spread.n	49
371	propose.v	64	401	mission.n	56	431	alternative.n	47
372	rank.n	64	402	provision.n	56	432	apparent.adj	47
373	solve.v	64	403	transport.n	56	433	compose.v	47
374	transform.v	64	404	brief.adj	55	434	decline.n	47
375	undermine.v	64	405	debate.v	55	435	gather.v	47
376	aware.adj	62	406	exchange.n	55	436	incident.n	47
377	differ.v	62	407	obvious.adj	55	437	inferior.adj	47
378	disaster.n	62	408	pose.v	55	438	suffering.n	47
379	overcome.v	62	409	potentially.adv	55	439	urge.v	47
380	struggle.v	62	410	survival.n	55	440	aggressive.adj	46
381	valuable.adj	62	411	vehicle.n	55	441	condemn.v	46
382	vary.v	62	412	extension.n	53	442	contact.n	46
383	emphasis.n	61	413	notion.n	53	443	harsh.adj	46
384	insist.v	61	414	ongoing.adj	53	444	proportion.n	46
385	initiate.v	59	415	reveal.v	53	445	scheme.n	46
386	neutral.adj	59	416	association.n	52	446	capable.adj	44
387	root.n	59	417	belong.v	52	447	confront.v	44
388	accompany.v	58	418	detailed.adj	52	448	constitute.v	44
389	connection.n	58	419	formally.adv	52	449	decline.v	44
390	criticism.n	58	420	motivate.v	52	450	permit.v	44

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	consolidate.v	43	477	agent.n	38	503	rush.v	33
452	distinct.adj	43	478	briefly.adv	38	504	stock.n	33
453	endure.v	43	479	burden.n	38	505	abroad.adv	32
454	entirely.adv	43	480	extensive.adj	38	506	attract.v	32
455	original.adj	43	481	fate.n	38	507	exert.v	32
456	predominantly.adv	43	482	non.n	38	508	philosophy.n	32
457	branch.n	41	483	spirit.n	38	509	prohibit.v	32
458	confrontation.n	41	484	arrest.n	37	510	trading.n	32
459	constant.adj	41	485	compromise.n	37	511	transformation.n	32
460	inhabitant.n	41	486	diminish.v	37	512	adult.n	30
461	justification.n	41	487	lesson.n	37	513	approval.n	30
462	male.adj	41	488	professional.adj	37	514	atrocity.n	30
463	manufacture.v	41	489	sale.n	37	515	calculate.v	30
464	prefer.v	41	490	site.n	37	516	lend.v	30
465	profit.n	41	491	alternative.adj	35	517	manner.n	30
466	rarely.adv	41	492	deliberately.adv	35	518	numerous.adj	30
467	successor.n	41	493	encounter.v	35	519	popularity.n	30
468	undertake.v	41	494	exercise.n	35	520	correct.adj	29
469	vision.n	41	495	independently.adv	35	521	gradually.adv	29
470	alter.v	40	496	likelihood.n	35	522	legislation.n	29
471	derive.v	40	497	stance.n	35	523	mutual.adj	29
472	explicitly.adv	40	498	stretch.v	35	524	phrase.n	29
473	guarantee.n	40	499	ambitious.adj	33	525	unstable.adj	29
474	manufacturing.n	40	500	campaign.v	33	526	warning.n	29
475	militarily.adv	40	501	clash.v	33			
476	monopoly.n	40	502	coal.n	33			

The IS-AVL for Biology (845 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	cell.n	8469	31	involve.v	965	61	transport.n	627
2	molecule.n	2786	32	dioxide.n	964	62	specific.adj	622
3	gene.n	2642	33	genetic.adj	957	63	condition.n	616
4	protein.n	2501	34	increase.v	945	64	substance.n	608
5	produce.v	2454	35	ion.n	895	65	pressure.n	595
6	organism.n	2383	36	bacteria.n	894	66	non.adj	593
7	example.n	2111	37	glucose.n	887	67	plasma.n	589
8	dna.n	2062	38	disease.n	885	68	individual.n	586
9	specie.n	1964	39	effect.n	872	69	common.adj	581
10	chromosome.n	1962	40	chain.n	868	70	reduce.v	573
11	acid.n	1907	41	factor.n	854	71	datum.n	572
12	structure.n	1868	42	function.n	824	72	bind.v	571
13	enzyme.n	1861	43	temperature.n	810	73	evidence.n	567
14	energy.n	1811	44	cycle.n	806	74	require.v	566
15	carbon.n	1807	45	amino.adj	774	75	determine.v	564
16	membrane.n	1778	46	electron.n	763	76	provide.v	542
17	occur.v	1631	47	surface.n	756	77	absorb.v	534
18	reaction.n	1564	48	experiment.n	744	78	nucleus.n	534
19	process.n	1455	49	develop.v	736	79	stem.n	534
20	contain.v	1259	50	production.n	736	80	characteristic.n	528
21	result.n	1195	51	respiration.n	719	81	compound.n	526
22	concentration.n	1159	52	product.n	715	82	root.n	523
23	population.n	1134	53	active.adj	707	83	development.n	521
24	oxygen.n	1109	54	site.n	706	84	meiosis.n	518
25	muscle.n	1074	55	photosynthesis.n	705	85	organic.adj	518
26	sequence.n	1070	56	hydrogen.n	693	86	theory.n	516
27	allele.n	1057	57	environment.n	687	87	role.n	515
28	area.n	1034	58	release.v	680	88	diagram.n	505
29	include.v	1032	59	hormone.n	675	89	nucleotide.n	502
30	tissue.n	994	60	strand.n	635	90	compare.v	501

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	offspring.n	498	121	gamete.n	403	151	volume.n	335
92	atom.n	492	122	division.n	397	152	cancer.n	334
93	material.n	490	123	similar.adj	395	153	convert.v	334
94	therefore.adv	485	124	lung.n	390	154	layer.n	333
95	affect.v	473	125	model.n	390	155	fluid.n	332
96	solution.n	468	126	period.n	389	156	sex.n	332
97	activity.n	467	127	soil.n	389	157	available.adj	331
98	link.v	467	128	normal.adj	388	158	calculate.v	331
99	research.n	466	129	mutation.n	378	159	homologous.adj	329
100	describe.v	460	130	obtain.v	375	160	remain.v	329
101	generation.n	459	131	value.n	375	161	chemical.adj	326
102	sugar.n	448	132	represent.v	372	162	atmosphere.n	325
103	cytoplasm.n	446	133	region.n	370	163	pathogen.n	323
104	potential.n	439	134	sample.n	367	164	formation.n	322
105	basis.n	435	135	measure.v	365	165	presence.n	321
106	source.n	435	136	prevent.v	364	166	component.n	320
107	selection.n	429	137	tube.n	364	167	sperm.n	316
108	section.n	427	138	divide.v	363	168	embryo.n	315
109	evolution.n	422	139	pattern.n	363	169	risk.n	315
110	result.v	421	140	nerve.n	358	170	chloroplast.n	313
111	identify.v	415	141	maintain.v	348	171	range.n	310
112	chemical.n	414	142	increase.n	346	172	phosphate.n	307
113	attach.v	411	143	lipid.n	346	173	exist.v	304
114	consist.v	411	144	remove.v	346	174	frequency.n	303
115	length.n	410	145	vessel.n	346	175	ratio.n	303
116	replication.n	409	146	nutrient.n	344	176	impulse.n	302
117	response.n	407	147	secrete.v	342	177	primary.adj	302
118	virus.n	407	148	variation.n	342	178	separate.v	301
119	method.n	405	149	mechanism.n	339	179	mass.n	300
120	fibre.n	404	150	gland.n	335	180	loss.n	298

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	contraction.n	297	211	graph.n	268	241	capillary.n	235
182	diffusion.n	297	212	variety.n	268	242	bacterial.adj	234
183	genotype.n	295	213	vein.n	268	243	survive.v	233
184	liver.n	291	214	phenotype.n	266	244	gradient.n	232
185	patient.n	291	215	direction.n	265	245	adapt.v	231
186	likely.adj	289	216	genome.n	265	246	ribosome.n	229
187	relatively.adv	285	217	amino.n	259	247	significant.adj	229
188	carbohydrate.n	282	218	insect.n	258	248	exchange.n	228
189	male.n	282	219	leaf.n	258	249	stimulate.v	228
190	polar.adj	282	220	transcription.n	258	250	event.n	226
191	indicate.v	279	221	substrate.n	257	251	image.n	226
192	dominant.adj	278	222	copy.n	256	252	genetically.adv	224
193	positive.adj	277	223	essential.adj	256	253	label.v	224
194	technique.n	277	224	identical.adj	256	254	due.adj	222
195	hypothesis.n	276	225	feature.n	254	255	catalyse.v	221
196	intestine.n	276	226	polymerase.n	252	256	mouse.n	221
197	code.n	273	227	diet.n	251	257	digestion.n	219
198	complex.adj	273	228	molecular.adj	249	258	habitat.n	219
199	create.v	273	229	sodium.n	247	259	dissolve.v	218
200	mammal.n	273	230	transport.v	247	260	internal.adj	216
201	metabolism.n	272	231	pigment.n	246	261	negative.adj	215
202	organ.n	272	232	reproduction.n	246	262	combination.n	214
203	eukaryotic.adj	270	233	surround.v	245	263	combine.v	214
204	synthesis.n	270	234	phase.n	243	264	consequence.n	212
205	female.adj	269	235	radiation.n	243	265	transfer.v	212
206	flow.n	269	236	receive.v	243	266	associate.v	210
207	independent.adj	269	237	insulin.n	241	267	unit.n	210
208	infection.n	269	238	community.n	239	268	absorption.n	209
209	recessive.adj	269	239	fossil.n	239	269	biological.adj	209
210	female.n	268	240	male.adj	237	270	mineral.n	209

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	prokaryote.n	209	301	series.n	185	331	osmosis.n	166
272	generate.v	206	302	element.n	184	332	pea.n	166
273	major.adj	205	303	individual.adj	184	333	pollen.n	166
274	polypeptide.adj	205	304	investigation.n	184	334	secretion.n	166
275	supply.n	205	305	observation.n	183	335	analysis.n	164
276	eventually.adv	203	306	original.adj	183	336	diffuse.v	164
277	bacterium.n	202	307	transfer.n	183	337	regulate.v	163
278	code.v	202	308	compose.v	182	338	calcium.n	162
279	observe.v	202	309	content.n	182	339	origin.n	162
280	various.adj	202	310	expose.v	182	340	repeat.v	162
281	species.n	200	311	fragment.n	182	341	thin.adj	162
282	starch.n	200	312	modify.v	182	342	chlorophyll.n	159
283	reproduce.v	199	313	adult.n	181	343	flow.v	159
284	aerobic.adj	197	314	random.adj	181	344	external.adj	157
285	eukaryote.n	196	315	decrease.v	178	345	influence.v	157
286	inheritance.n	196	316	particle.n	178	346	survival.n	157
287	principle.n	196	317	sexual.adj	178	347	vesicle.n	157
288	environmental.adj	195	318	cellular.adj	177	348	physical.adj	156
289	channel.n	194	319	fruit.n	177	349	classify.v	155
290	organelle.n	193	320	separate.adj	177	350	detect.v	155
291	potassium.n	191	321	distribution.n	174	351	replace.v	155
292	metabolic.n	190	322	nucleic.adj	174	352	knowledge.n	153
293	digest.v	189	323	pump.v	174	353	nervous.adj	152
294	release.n	189	324	activate.v	171	354	perform.v	151
295	translation.n	189	325	predict.v	171	355	pregnancy.n	150
296	discuss.v	187	326	examine.v	169	356	define.v	149
297	establish.v	187	327	apparatus.n	168	357	inherit.v	149
298	helix.n	185	328	synthesize.v	168	358	interaction.n	149
299	limit.v	185	329	vary.v	168	359	resistant.adj	149
300	locate.v	185	330	visible.adj	168	360	biology.n	147

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	link.n	147	391	experimental.adj	134	421	undergo.v	122
362	unique.adj	147	392	normally.adv	134	422	uterus.n	122
363	follicle.n	146	393	respond.v	134	423	typical.adj	121
364	opposite.adj	146	394	correct.adj	133	424	mixture.n	120
365	rapidly.adv	146	395	variable.n	133	425	rapid.adj	119
366	infect.v	145	396	limb.n	132	426	accumulate.v	118
367	procedure.n	145	397	band.n	131	427	concept.n	118
368	culture.n	144	398	select.v	131	428	contribute.v	118
369	differ.v	144	399	capillary.adj	130	429	histone.n	118
370	inhibit.v	144	400	damage.n	130	430	replicate.v	118
371	reduction.n	144	401	gut.n	130	431	cardiac.adj	117
372	issue.n	143	402	intensity.n	130	432	genetic.n	117
373	outline.v	143	403	potential.adj	130	433	modification.n	117
374	cellulose.n	141	404	scale.n	130	434	technology.n	117
375	apply.v	140	405	distance.n	128	435	fuel.n	115
376	digestive.adj	140	406	ensure.v	128	436	linkage.n	115
377	enable.v	140	407	constant.adj	127	437	modern.adj	115
378	medium.n	140	408	duct.n	127	438	supply.v	115
379	century.n	139	409	accord.v	126	439	damage.v	114
380	cholesterol.n	139	410	composition.n	126	440	design.v	114
381	entire.adj	138	411	directly.adv	126	441	glycogen.n	114
382	estimate.v	137	412	radioactive.adj	126	442	matrix.n	114
383	location.n	137	413	commonly.adv	125	443	absence.n	113
384	measurement.n	137	414	transmission.n	125	444	average.adj	113
385	mitochondrion.n	137	415	approximately.adv	124	445	aquatic.adj	112
386	symptom.n	137	416	multiple.adj	124	446	immediately.adv	112
387	advantage.n	136	417	reproductive.adj	124	447	outcome.n	112
388	breeding.n	136	418	stable.adj	124	448	soluble.adj	112
389	arise.v	134	419	branch.n	122	449	circular.adj	111
390	biochemical.adj	134	420	proportion.n	122	450	construct.v	111

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	diameter.n	111	481	smooth.adj	102	511	assume.v	95
452	ray.n	111	482	successful.adj	102	512	liquid.n	95
453	recognize.v	111	483	current.adj	101	513	passive.adj	95
454	tip.n	111	484	efficient.adj	101	514	ethical.adj	94
455	block.v	109	485	electrical.adj	101	515	industrial.adj	94
456	inorganic.adj	109	486	finally.adv	101	516	coli.n	93
457	relative.adj	109	487	function.v	101	517	harmful.adj	93
458	contact.n	108	488	net.adj	101	518	interact.v	92
459	detail.n	108	489	propose.v	101	519	spectrum.n	92
460	pre.adj	108	490	publish.v	101	520	incidence.n	90
461	resource.n	108	491	anaerobic.adj	100	521	occupy.v	90
462	artificial.adj	107	492	connect.v	100	522	zygote.n	90
463	clone.n	107	493	height.n	100	523	adjacent.adj	89
464	isolate.v	107	494	object.n	100	524	alter.v	89
465	ovary.n	107	495	photosynthetic.adj	100	525	comparison.n	89
466	similarity.n	107	496	structural.adj	100	526	contract.n	89
467	tumour.n	107	497	critical.adj	99	527	exposure.n	89
468	embryonic.adj	106	498	maximum.adj	99	528	rare.adj	89
469	extremely.adv	106	499	permeable.adj	99	529	conclusion.n	88
470	transcribe.v	106	500	suffer.v	99	530	limited.adj	88
471	progesterone.n	105	501	achieve.v	97	531	medical.adj	88
472	standard.adj	105	502	column.n	97	532	prove.v	88
473	contract.v	103	503	consume.v	97	533	reflect.v	88
474	demonstrate.v	103	504	fungus.n	97	534	respiratory.adj	88
475	mature.adj	103	505	narrow.adj	97	535	cord.n	87
476	strand.v	103	506	promote.v	97	536	final.adj	87
477	subunit.n	103	507	reveal.v	97	537	locus.n	87
478	breed.v	102	508	testis.n	97	538	plate.n	87
479	diabetes.n	102	509	pore.n	96	539	split.v	87
480	effective.adj	102	510	appropriate.adj	95	540	toxic.adj	87

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
541	destroy.v	86	571	intermediate.adj	80	601	nutrient.adj	74
542	hydrolysis.n	86	572	marine.adj	80	602	semi.adj	74
543	improve.v	86	573	previous.adj	80	603	separation.n	74
544	terrestrial.adj	86	574	previously.adv	80	604	actively.adv	73
545	activation.n	84	575	data.n	78	605	direct.adj	73
546	gain.v	84	576	design.n	78	606	maternal.adj	73
547	summary.n	84	577	distinguish.v	78	607	sequence.v	73
548	avoid.v	83	578	exception.n	78	608	significantly.adv	73
549	complex.n	83	579	induce.v	78	609	analyse.v	71
550	coronary.adj	83	580	majority.n	78	610	block.n	71
551	grain.n	83	581	removal.n	78	611	dimensional.adj	71
552	importance.n	83	582	toxin.n	78	612	ingest.v	71
553	lack.n	83	583	accurate.adj	77	613	limit.n	71
554	lower.v	83	584	alternative.adj	77	614	capacity.n	70
555	optimum.adj	83	585	mix.v	77	615	characteristic.adj	70
556	phenomenon.n	83	586	overall.adj	77	616	potato.n	70
557	sufficient.adj	83	587	phosphate.adj	77	617	syndrome.n	70
558	template.n	83	588	vapour.n	77	618	team.n	70
559	approach.n	82	589	trap.v	76	619	aspect.n	69
560	computer.n	82	590	balance.n	75	620	match.v	69
561	extend.v	82	591	derive.v	75	621	process.v	69
562	facilitated.adj	82	592	dialysis.n	75	622	severe.adj	69
563	polysaccharide.n	82	593	equal.adj	75	623	variable.adj	69
564	retain.v	82	594	fuse.v	75	624	category.n	68
565	reticulum.n	82	595	blindness.n	74	625	frequently.adv	68
566	shell.n	82	596	compete.v	74	626	fusion.n	68
567	tertiary.adj	82	597	conversion.n	74	627	initial.adj	68
568	primer.n	81	598	decision.n	74	628	measure.n	68
569	attract.v	80	599	differentiate.v	74	629	speed.n	68
570	glycerol.n	80	600	excess.adj	74	630	argue.v	67

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
631	conduct.v	67	661	originate.v	58	691	gap.n	54
632	differentiation.n	67	662	villus.n	58	692	interpret.v	54
633	fail.v	67	663	belong.v	57	693	label.n	54
634	liquid.adj	67	664	calculation.n	57	694	villi.n	54
635	purpose.n	67	665	emerge.v	57	695	absent.adj	52
636	selective.adj	67	666	fibrous.adj	57	696	advance.n	52
637	extract.v	65	667	influence.n	57	697	chemically.adv	52
638	rice.n	65	668	insoluble.adj	57	698	correspond.v	52
639	disorder.n	64	669	resemble.v	57	699	excrete.v	52
640	existence.n	64	670	ribosome.adj	57	700	originally.adv	52
641	extent.n	64	671	solid.adj	57	701	correlate.v	51
642	pure.adj	64	672	universal.adj	57	702	effectively.adv	51
643	representation.n	64	673	extreme.adj	56	703	raw.adj	51
644	season.n	64	674	numerous.adj	56	704	recycle.v	51
645	translate.v	64	675	reject.v	56	705	attempt.v	50
646	upper.adj	64	676	restrict.v	56	706	cohesion.n	50
647	twin.n	63	677	successfully.adv	56	707	fertile.adj	50
648	heritable.adj	62	678	contrast.n	55	708	germinate.v	50
649	testosterone.n	62	679	decrease.n	55	709	pancreas.n	50
650	independently.adv	61	680	disappear.v	55	710	regular.adj	50
651	inject.v	61	681	distribute.v	55	711	relax.v	50
652	shift.n	61	682	facilitate.v	55	712	acquire.v	49
653	confirm.v	59	683	interior.n	55	713	alpha.n	49
654	isotope.n	59	684	mitochondria.n	55	714	communication.n	49
655	sac.n	59	685	obvious.adj	55	715	description.n	49
656	statistical.adj	59	686	positively.adv	55	716	lack.v	49
657	currently.adv	58	687	subsequent.adj	55	717	plot.v	49
658	endoplasmic.adj	58	688	distinct.adj	54	718	rely.v	49
659	error.n	58	689	escape.v	54	719	arrow.n	48
660	metre.n	58	690	failure.n	54	720	desire.v	48

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
721	entry.n	48	751	haploid.n	44	781	decay.v	37
722	fixation.n	48	752	interbreed.v	44	782	detailed.adj	37
723	future.n	48	753	stain.n	44	783	ideal.adj	37
724	medicine.n	48	754	benefit.v	43	784	junction.n	37
725	paradigm.n	48	755	luteum.n	43	785	multiply.v	37
726	switch.v	48	756	mucus.n	43	786	needle.n	37
727	task.n	48	757	divergence.n	42	787	permanent.adj	37
728	abundant.adj	46	758	ecology.n	42	788	sexually.adv	37
729	attachment.n	46	759	frequent.adj	42	789	abnormal.adj	36
730	autosomal.adj	46	760	macromolecule.n	42	790	bean.n	36
731	cow.n	46	761	otherwise.adv	42	791	disrupt.v	36
732	donor.n	46	762	project.n	42	792	embed.v	36
733	negatively.adv	46	763	speed.v	40	793	encourage.v	36
734	peptide.adj	46	764	beetle.n	39	794	award.v	34
735	potentially.adv	46	765	breathe.v	39	795	binding.n	34
736	search.n	46	766	cattle.n	39	796	corpus.n	34
737	sequencing.n	46	767	constantly.adv	39	797	deposit.v	34
738	strength.n	46	768	harm.n	39	798	direct.v	34
739	argument.n	45	769	isolated.adj	39	799	efficiently.adv	34
740	attack.n	45	770	mammalian.adj	39	800	excess.n	34
741	circumstance.n	45	771	metal.n	39	801	extracellular.adj	34
742	commercial.adj	45	772	pregnant.adj	39	802	lipase.n	34
743	conclude.v	45	773	reference.n	39	803	osmotic.adj	34
744	excessive.adj	45	774	demand.n	38	804	peripheral.adj	34
745	fundamental.adj	45	775	equally.adv	38	805	phagocytosis.n	34
746	inactive.adj	45	776	iodine.n	38	806	quality.n	34
747	lysosome.n	45	777	row.n	38	807	recover.v	34
748	rarely.adv	45	778	ultimately.adv	38	808	surrounding.n	34
749	sheet.n	45	779	weak.adj	38	809	trace.v	34
750	extensive.adj	44	780	access.n	37	810	wound.n	34

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
811	implication.n	33	823	statistic.n	32	835	extreme.n	30
812	imply.v	33	824	valuable.adj	32	836	fold.n	30
813	persist.v	33	825	volcanic.adj	32	837	harvest.v	30
814	primarily.adv	33	826	branch.v	31	838	invertebrate.n	30
815	quaternary.adj	33	827	correctly.adv	31	839	microscopic.adj	30
816	sum.n	33	828	dye.n	31	840	parallel.adj	30
817	suspend.v	33	829	feather.n	31	841	regularly.adv	30
818	current.n	32	830	migration.n	31	842	resist.v	30
819	fuse.n	32	831	requirement.n	31	843	splitting.n	30
820	granule.n	32	832	research.v	31	844	sticky.adj	30
821	interval.n	32	833	spontaneously.adv	31	845	transplant.n	30
822	specimen.n	32	834	disadvantage.n	30			

The IS-AVL for Chemistry (681 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	reaction.n	7901	31	occur.v	1262	61	negative.adj	770
2	electron.n	5317	32	process.n	1208	62	standard.adj	768
3	atom.n	4437	33	oxidation.n	1205	63	non.adj	764
4	ion.n	4423	34	property.n	1119	64	require.v	753
5	energy.n	4290	35	substance.n	1116	65	result.n	741
6	acid.n	4251	36	formula.n	1108	66	reactant.n	737
7	bond.n	3965	37	constant.adj	1106	67	unit.n	707
8	molecule.n	3748	38	volume.n	1080	68	relative.adj	704
9	solution.n	3235	39	effect.n	1052	69	chemical.adj	702
10	example.n	3127	40	sodium.n	1050	70	measure.v	701
11	hydrogen.n	2651	41	datum.n	1037	71	chloride.n	700
12	concentration.n	2473	42	particle.n	1035	72	dioxide.n	695
13	carbon.n	2422	43	molecular.adj	1025	73	catalyst.n	684
14	temperature.n	2305	44	chemical.n	990	74	organic.adj	684
15	equation.n	2301	45	therefore.adv	988	75	mechanism.n	676
16	value.n	2266	46	describe.v	962	76	spectrum.n	670
17	structure.n	2195	47	theory.n	937	77	formation.n	651
18	metal.n	2144	48	reduce.v	936	78	include.v	649
19	compound.n	1968	49	atomic.adj	933	79	covalent.n	648
20	mass.n	1942	50	react.v	913	80	transition.n	644
21	element.n	1932	51	pressure.n	890	81	condition.n	639
22	produce.v	1636	52	mixture.n	876	82	graph.n	617
23	cell.n	1605	53	increase.n	854	83	polar.adj	602
24	product.n	1511	54	represent.v	841	84	basis.n	596
25	oxygen.n	1486	55	bonding.n	818	85	chlorine.n	588
26	contain.v	1455	56	positive.adj	818	86	salt.n	587
27	calculate.v	1415	57	weak.adj	807	87	model.n	586
28	involve.v	1322	58	electrode.n	804	88	sample.n	584
29	increase.v	1308	59	oxide.n	799	89	proton.n	571
30	determine.v	1272	60	aqueous.adj	772	90	nitrogen.n	554

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	nucleus.n	544	121	lattice.n	433	151	molar.adj	371
92	series.n	544	122	discuss.v	429	152	percentage.n	368
93	deduce.v	539	123	common.adj	428	153	calcium.n	362
94	experiment.n	539	124	complex.adj	425	154	initial.adj	362
95	specie.n	536	125	combustion.n	422	155	acidic.adj	359
96	copper.n	530	126	ratio.n	420	156	component.n	357
97	liquid.n	521	127	magnesium.n	419	157	exist.v	354
98	solid.adj	520	128	similar.adj	418	158	ethanol.n	353
99	presence.n	513	129	solvent.n	414	159	specific.adj	352
100	identify.v	501	130	factor.n	413	160	lone.adj	349
101	decrease.v	497	131	provide.v	407	161	stable.adj	347
102	alcohol.n	491	132	iron.n	402	162	region.n	346
103	sulfur.n	478	133	radiation.n	401	163	aluminium.n	342
104	remove.v	467	134	potassium.n	400	164	frequency.n	342
105	convert.v	464	135	bond.v	396	165	melting.n	342
106	reduction.n	461	136	range.n	396	166	refer.v	340
107	chemistry.n	460	137	experimental.adj	395	167	calculation.n	339
108	isomer.n	459	138	nuclear.adj	391	168	iodine.n	332
109	equal.adj	457	139	benzene.n	390	169	structural.adj	331
110	result.v	457	140	strength.n	387	170	technique.n	331
111	compare.v	456	141	indicate.v	383	171	crystal.n	327
112	overall.adj	456	142	surface.n	381	172	react.n	326
113	physical.adj	453	143	curve.n	380	173	functional.adj	323
114	absorb.v	450	144	obtain.v	378	174	kinetic.adj	322
115	density.n	443	145	release.v	378	175	scale.n	320
116	hydroxide.adv	443	146	apply.v	377	176	halogen.n	319
117	method.n	441	147	cathode.n	377	177	remain.v	316
118	predict.v	437	148	boiling.n	375	178	hydrochloric.adj	315
119	titration.n	435	149	dissolve.v	375	179	develop.v	314
120	agent.n	434	150	significant.adj	373	180	pure.adj	314

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	direction.n	312	211	soluble.adj	263	241	electricity.n	217
182	principle.n	312	212	production.n	262	242	masse.n	216
183	reactant.adj	308	213	potential.adj	261	243	monoxide.n	214
184	diagram.n	307	214	cation.n	258	244	final.adj	211
185	anode.n	305	215	correct.adj	257	245	industry.n	211
186	define.v	305	216	layer.n	252	246	observe.v	211
187	affect.v	303	217	development.n	250	247	research.n	210
188	activity.n	300	218	decrease.n	248	248	concept.n	209
189	alkene.n	299	219	exothermic.adj	248	249	balanced.adj	207
190	hydrocarbon.n	298	220	length.n	247	250	ray.n	207
191	undergo.v	296	221	average.adj	246	251	endothermic.adj	205
192	phase.n	295	222	electrolysis.n	246	252	replace.v	205
193	area.n	293	223	interaction.n	246	253	assume.v	205
194	carbonate.n	292	224	quantity.n	245	254	analysis.n	204
195	angle.n	291	225	dissociation.n	240	255	pattern.n	201
196	peak.n	291	226	zinc.n	240	256	loss.n	200
197	isotope.n	291	227	chemist.n	237	257	separate.v	200
198	empirical.adj	285	228	reactive.adj	237	258	derive.v	196
199	directly.adv	283	229	relatively.adv	234	259	vary.v	196
200	primary.adj	283	230	symbol.n	234	260	function.n	195
201	consist.v	282	231	available.adj	229	261	atmosphere.n	194
202	liquid.adj	276	232	axis.n	227	262	plane.n	193
203	combine.v	273	233	gain.v	224	263	composition.n	190
204	magnetic.adj	272	234	emission.n	223	264	methane.n	188
205	absorption.n	271	235	generate.v	223	265	carboxylic.adj	187
206	accord.v	270	236	knowledge.n	223	266	proportional.adj	187
207	electrical.adj	270	237	major.adj	223	267	attract.v	186
208	ammonia.n	269	238	yield.n	219	268	silver.n	186
209	illustrate.v	268	239	associate.v	218	269	various.adj	186
210	evidence.n	266	240	effective.adj	217	270	intermediate.adj	185

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	excess.adj	183	301	observation.n	157	331	commonly.adv	138
272	tertiary.adj	183	302	solid.n	155	332	concentrated.adj	138
273	behaviour.n	181	303	appropriate.adj	152	333	conversion.n	138
274	shift.n	178	304	distribution.n	152	334	uv.n	138
275	arrow.n	177	305	equivalent.adj	150	335	coefficient.n	136
276	neutral.adj	177	306	industrial.adj	150	336	laboratory.n	136
277	balance.v	176	307	introduce.v	150	337	aldehyde.n	134
278	divide.v	176	308	graphite.n	149	338	feature.n	133
279	measure.n	176	309	iodide.n	149	339	representation.n	133
280	establish.v	173	310	original.adj	149	340	dissociate.v	131
281	approach.n	171	311	transfer.n	149	341	multiply.v	130
282	combination.n	171	312	likely.adj	148	342	variety.n	130
283	correspond.v	171	313	molten.adj	148	343	electromagnetic.adj	129
284	opposite.adj	171	314	decomposition.n	146	344	orientation.n	129
285	due.adj	170	315	repeat.v	146	345	transfer.v	129
286	individual.adj	170	316	definition.n	145	346	electrolyte.n	129
287	differ.v	169	317	external.adj	145	347	identical.adj	129
288	fluorine.n	169	318	limit.v	145	348	substitute.v	129
289	silicon.n	167	319	linear.adj	144	349	theoretical.adj	129
290	bromide.n	166	320	nitrate.adj	144	350	create.v	128
291	maximum.adj	166	321	capacity.n	143	351	radical.n	128
292	achieve.v	163	322	ester.n	143	352	stability.n	128
293	alkali.adj	162	323	summarize.v	143	353	surround.v	128
294	reverse.adj	162	324	container.n	142	354	balance.n	127
295	visible.adj	159	325	gradient.n	142	355	essential.adj	127
296	conductivity.n	158	326	methyl.n	140	356	approximately.adv	125
297	electric.adj	158	327	proportion.n	140	357	typical.adj	125
298	spin.n	157	328	sum.n	140	358	voltage.n	125
299	ammonium.n	157	329	mix.v	139	359	conduct.v	124
300	extent.n	157	330	reactivity.n	139	360	distance.n	124

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	hydrolysis.n	124	391	experimentally.adv	110	421	solve.v	99
362	polarity.n	124	392	technology.n	110	422	conductor.n	98
363	respectively.adv	123	393	cancel.v	110	423	direct.adj	98
364	current.n	122	394	dilute.n	110	424	ketone.n	98
365	tetrahedral.adj	122	395	heterogeneous.adj	110	425	signal.n	97
366	description.n	121	396	importance.n	109	426	positively.adv	96
367	rapidly.adv	121	397	language.n	109	427	account.v	94
368	propose.v	120	398	circuit.n	108	428	emit.v	94
369	characteristic.adj	119	399	hydrated.adj	108	429	flow.n	94
370	computer.n	119	400	precise.adj	108	430	row.n	94
371	distinguish.v	119	401	catalyse.v	107	431	assumption.n	93
372	alloy.n	119	402	chemically.adv	107	432	saturate.v	93
373	dimensional.adj	119	403	homogeneous.adj	107	433	supply.n	93
374	plot.v	119	404	triple.adj	106	434	halide.n	92
375	classify.v	118	405	lithium.n	105	435	normal.adj	92
376	photon.n	117	406	platinum.n	105	436	account.n	91
377	behave.v	116	407	contribute.v	104	437	consume.v	91
378	detect.v	116	408	melt.v	104	438	instrument.n	91
379	spectra.n	116	409	nitric.adj	104	439	summary.n	91
380	flow.v	115	410	connect.v	103	440	apparatus.n	90
381	excess.n	114	411	reversible.adj	103	441	minimum.adj	90
382	quantitative.adj	114	412	sufficient.adj	102	442	variation.n	90
383	split.v	114	413	alkaline.n	100	443	arise.v	89
384	tube.n	114	414	donate.v	100	444	effectively.adv	89
385	image.n	113	415	mathematical.adj	100	445	regular.adj	89
386	prediction.n	113	416	current.adj	100	446	contact.n	88
387	splitting.n	113	417	mass.adj	100	447	improve.v	88
388	accurate.adj	112	418	modern.adj	100	448	label.v	88
389	insoluble.adj	111	419	electrochemical.adj	99	449	methanol.n	88
390	analyse.v	110	420	object.n	99	450	typically.adv	88

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	nickel.n	87	481	significantly.adv	77	511	concentrate.v	66
452	extend.v	86	482	standard.n	77	512	initially.adv	66
453	interact.v	86	483	aspect.n	76	513	interpret.v	66
454	spectrometer.n	85	484	determination.n	76	514	economy.n	65
455	planar.adj	84	485	negatively.adv	76	515	monitor.v	65
456	butane.n	83	486	ultraviolet.adj	76	516	origin.n	65
457	expand.v	83	487	abundance.n	75	517	issue.n	64
458	intensity.n	83	488	construct.v	75	518	vessel.n	64
459	reverse.v	83	489	convenient.adj	75	519	impurity.n	63
460	century.n	82	490	ensure.v	75	520	magnitude.n	63
461	compose.v	82	491	estimate.v	75	521	purpose.n	63
462	fundamental.adj	82	492	profile.n	75	522	convention.n	62
463	characteristic.n	81	493	donor.n	74	523	desire.v	62
464	independent.adj	81	494	influence.v	74	524	dissolve.n	62
465	kinetic.n	81	495	inert.adj	73	525	familiar.adj	62
466	lower.v	81	496	reflect.v	73	526	international.adj	62
467	contrast.n	80	497	bridge.n	72	527	motion.n	62
468	exception.n	80	498	ethane.n	72	528	redox.adj	62
469	separate.adj	80	499	notation.n	72	529	approximate.adj	62
470	speed.n	80	500	universal.adj	72	530	atmospheric.adj	62
471	barium.n	79	501	attack.n	71	531	confirm.v	62
472	chromium.n	79	502	investigate.v	71	532	demonstrate.v	62
473	eventually.adv	78	503	overcome.v	71	533	equal.v	62
474	multiple.adj	78	504	attractive.adj	69	534	flame.n	62
475	practice.n	78	505	hydrogenation.n	69	535	frequently.adv	62
476	accompany.v	77	506	hydroxide.n	69	536	limit.n	62
477	assign.v	77	507	procedure.n	68	537	perform.v	62
478	boron.n	77	508	removal.n	68	538	vertical.adj	62
479	comparison.n	77	509	supply.v	68	539	distribute.v	61
480	consequence.n	77	510	decompose.v	67	540	gain.n	61

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
541	accurately.adv	60	571	propane.n	52	601	unstable.adj	46
542	device.n	60	572	relevant.adj	52	602	upper.adj	46
543	limited.adj	60	573	bracket.n	51	603	correspond.n	45
544	trace.n	60	574	escape.v	51	604	hydride.n	45
545	approximation.n	59	575	literature.n	51	605	identity.n	45
546	equally.adv	59	576	purple.adj	51	606	mechanical.adj	45
547	planar.n	59	577	sheet.n	51	607	numerical.adj	45
548	extremely.adv	58	578	similarity.n	51	608	subtract.v	45
549	interpretation.n	58	579	variable.adj	51	609	biochemical.adj	44
550	prove.v	58	580	crystalline.n	50	610	evaluate.v	44
551	successful.adj	58	581	detailed.adj	50	611	horizontal.adj	44
552	fractional.adj	57	582	evaporation.n	50	612	correctly.adv	43
553	broad.adj	56	583	select.v	50	613	gradually.adv	43
554	introduction.n	56	584	anhydrous.adj	49	614	inductive.adj	43
555	catalysis.n	55	585	essentially.adv	49	615	replacement.n	43
556	diatomic.adj	55	586	ignore.v	49	616	rapid.adj	43
557	focus.v	55	587	unique.adj	49	617	valid.adj	43
558	originally.adv	55	588	dilute.v	48	618	wire.n	43
559	beaker.n	54	589	display.v	48	619	comment.n	42
560	investigation.n	54	590	avoid.v	47	620	contribution.n	42
561	promote.v	54	591	efficient.adj	47	621	implication.n	42
562	trioxide.n	54	592	expose.v	47	622	increasingly.adv	42
563	approach.v	53	593	fertilizer.n	47	623	range.v	42
564	distinct.adj	53	594	imply.v	47	624	inversely.adv	41
565	kg.n	53	595	isolate.v	47	625	mixed.adj	41
566	influence.n	52	596	overall.adv	47	626	normally.adv	41
567	significance.n	52	597	permanent.adj	47	627	design.v	40
568	specify.v	52	598	sketch.v	47	628	dimension.n	40
569	design.n	52	599	construction.n	46	629	eliminate.v	40
570	economic.adj	52	600	pink.adj	46	630	yield.v	40

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
631	cyclic.adj	39	648	vice.n	35	665	determined.adj	31
632	height.n	39	649	artificial.adj	34	666	metal.adj	31
633	purity.n	39	650	exert.v	34	667	mg.n	31
634	assess.v	38	651	impure.n	34	668	restrict.v	31
635	spectrometry.n	38	652	precisely.adv	34	669	simultaneously.adv	31
636	alkaline.adj	37	653	rely.v	34	670	tablet.n	31
637	commercial.adj	37	654	subsequent.adj	34	671	vehicle.n	31
638	hexane.n	37	655	belong.v	33	672	electrically.adv	30
639	indication.n	36	656	complicated.adj	33	673	encounter.v	30
640	tiny.adj	36	657	microscopic.adj	33	674	multi.adj	30
641	unreactive.adj	36	658	retain.v	33	675	oppose.v	30
642	volatility.n	36	659	average.n	33	676	prefer.v	30
643	alternatively.adv	35	660	immediately.adv	33	677	separately.adv	30
644	progress.n	35	661	titanium.n	33	678	sharp.adj	30
645	publish.v	35	662	insert.v	32	679	process.v	29
646	requirement.n	35	663	magnet.n	32	680	reveal.v	29
647	temporary.adj	35	664	specifically.adv	32	681	subscript.n	29

The IS-AVL for Physics (578 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	energy.n	7422	31	determine.v	1094	61	positive.adj	668
2	mass.n	2938	32	atom.n	1053	62	quantity.n	667
3	speed.n	2554	33	unit.n	1027	63	resistor.n	654
4	particle.n	2440	34	nucleus.n	1026	64	solution.n	653
5	electron.n	2191	35	acceleration.n	1018	65	apply.v	646
6	example.n	2145	36	effect.n	1010	66	maximum.adj	635
7	temperature.n	1996	37	gravitational.adj	1009	67	magnitude.n	632
8	distance.n	1865	38	area.n	911	68	result.n	628
9	calculate.v	1714	39	increase.v	906	69	average.adj	623
10	equation.n	1707	40	kinetic.adj	890	70	proton.n	622
11	direction.n	1659	41	length.n	889	71	volume.n	612
12	graph.n	1618	42	equal.adj	881	72	density.n	610
13	object.n	1600	43	circuit.n	872	73	thermal.adj	596
14	velocity.n	1597	44	pressure.n	848	74	datum.n	591
15	value.n	1564	45	cm.n	810	75	involve.v	579
16	constant.adj	1546	46	molecule.n	796	76	radius.n	571
17	surface.n	1516	47	intensity.n	766	77	voltage.n	570
18	current.n	1447	48	represent.v	765	78	material.n	567
19	potential.adj	1416	49	experiment.n	753	79	require.v	562
20	frequency.n	1359	50	therefore.adv	733	80	strength.n	561
21	measure.v	1330	51	wire.n	729	81	theory.n	558
22	electric.adj	1326	52	photon.n	713	82	negative.adj	555
23	motion.n	1274	53	momentum.n	708	83	reaction.n	550
24	produce.v	1251	54	slit.n	708	84	model.n	548
25	source.n	1199	55	kg.n	706	85	pattern.n	544
26	magnetic.adj	1194	56	current.adj	700	86	component.n	540
27	diagram.n	1174	57	displacement.n	699	87	physics.n	533
28	radiation.n	1167	58	emit.v	696	88	electrical.adj	517
29	angle.n	1111	59	describe.v	686	89	diffraction.n	510
30	resistance.n	1097	60	decay.n	668	90	path.n	506

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	define.v	505	121	orbit.n	401	151	specific.adj	338
92	electromagnetic.adj	505	122	relative.adj	399	152	similar.adj	336
93	connect.v	502	123	satellite.n	395	153	estimate.v	335
94	occur.v	495	124	flow.v	391	154	flow.n	335
95	period.n	489	125	amplitude.n	390	155	curve.n	329
96	internal.adj	483	126	observe.v	389	156	ideal.adj	328
97	process.n	482	127	spectrum.n	389	157	initial.adj	325
98	assume.v	477	128	tube.n	389	158	rotate.v	320
99	reflect.v	466	129	incident.n	385	159	atomic.adj	317
100	measurement.n	456	130	hydrogen.n	381	160	convert.v	313
101	parallel.adj	449	131	masse.n	381	161	medium.n	311
102	axis.n	447	132	vertical.adj	381	162	series.n	311
103	accelerate.v	444	133	interaction.n	379	163	predict.v	309
104	absorb.v	442	134	property.n	375	164	result.v	308
105	conductor.n	442	135	range.n	372	165	equilibrium.n	303
106	height.n	438	136	scale.n	372	166	factor.n	300
107	metal.n	436	137	vary.v	371	167	section.n	297
108	contain.v	429	138	proportional.adj	370	168	conservation.n	293
109	opposite.adj	428	139	alpha.n	369	169	normal.adj	293
110	provide.v	423	140	beam.n	369	170	concept.n	285
111	increase.n	421	141	radioactive.adj	366	171	element.n	285
112	interference.n	415	142	phase.n	364	172	sin.n	282
113	planet.n	415	143	compare.v	364	173	gravity.n	276
114	sphere.n	415	144	release.v	364	174	gradient.n	272
115	include.v	413	145	atmosphere.n	360	175	screen.n	272
116	principle.n	413	146	region.n	358	176	separation.n	270
117	reduce.v	413	147	discuss.v	356	177	fundamental.adj	269
118	variation.n	412	148	liquid.n	354	178	decrease.v	266
119	horizontal.adj	409	149	circular.adj	348	179	remain.v	264
120	plane.n	408	150	km.n	341	180	common.adj	258

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	substance.n	255	211	fusion.n	219	241	correct.adj	185
182	magnet.n	253	212	equivalent.adj	218	242	linear.adj	185
183	rod.n	252	213	obtain.v	218	243	definition.n	184
184	diameter.n	250	214	thin.adj	216	244	deduce.v	182
185	electricity.n	249	215	non.adj	214	245	interval.n	182
186	create.v	247	216	separate.v	214	246	structure.n	182
187	device.n	247	217	absorption.n	213	247	production.n	181
188	ratio.n	246	218	approximately.adv	213	248	combine.v	180
189	loss.n	246	219	physicist.n	212	249	refractive.adj	179
190	stationary.adj	246	220	arrow.n	208	250	visible.adj	178
191	original.adj	245	221	resolve.v	208	251	detail.n	175
192	directly.adv	242	222	index.n	207	252	equal.v	174
193	efficiency.n	239	223	supply.v	207	253	evidence.n	174
194	uniform.adj	239	224	carbon.n	205	254	remove.v	174
195	emission.n	237	225	receive.v	205	255	affect.v	172
196	method.n	237	226	detect.v	204	256	peak.n	171
197	significant.adj	237	227	plot.v	201	257	typical.adj	171
198	transmit.v	237	228	maximum.n	200	258	identical.adj	170
199	function.n	236	229	scatter.v	200	259	conserve.v	169
200	hence.adv	236	230	solid.adj	200	260	gamma.n	169
201	sample.n	236	231	reflection.n	198	261	surrounding.n	169
202	transfer.n	232	232	condition.n	196	262	individual.adj	168
203	application.n	229	233	metre.n	196	263	minimum.adj	167
204	output.n	229	234	activity.n	195	264	rocket.n	167
205	calculation.n	229	235	exert.v	195	265	observation.n	164
206	final.adj	226	236	gain.v	195	266	vacuum.n	163
207	generator.n	222	237	isotope.n	192	267	available.adj	162
208	symbol.n	222	238	refer.v	191	268	orbit.v	162
209	consist.v	221	239	perpendicular.adj	188	269	generate.v	161
210	develop.v	219	240	random.adj	187	270	helium.n	160

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	pole.n	158	301	vibration.n	139	331	mechanic.n	121
272	contact.n	157	302	shift.n	138	332	illustrate.v	120
273	container.n	157	303	overall.adj	137	333	limit.n	120
274	indicate.v	157	304	undergo.v	137	334	infinite.adj	119
275	investigate.v	157	305	introduce.v	136	335	assumption.n	118
276	primary.adj	157	306	identify.v	135	336	dissipate.v	118
277	various.adj	157	307	motor.n	135	337	measure.n	115
278	derive.v	156	308	instrument.n	134	338	attach.v	113
279	exist.v	156	309	phenomenon.n	134	339	narrow.adj	113
280	sum.n	155	310	advantage.n	133	340	perform.v	112
281	percentage.n	152	311	combination.n	133	341	transition.n	112
282	calculate.n	151	312	critical.adj	133	342	oppose.v	111
283	mass.adj	149	313	repeat.v	132	343	propose.v	111
284	associate.v	148	314	coefficient.n	131	344	aperture.n	110
285	resolution.n	147	315	cylinder.n	131	345	design.v	110
286	spread.v	146	316	experimental.adj	131	346	dimension.n	110
287	constant.n	145	317	achieve.v	130	347	gravitation.n	110
288	product.n	145	318	demonstrate.v	130	348	negligible.adj	110
289	harmonic.adj	144	319	laboratory.n	130	349	vibrate.v	110
290	computer.n	144	320	cloud.n	129	350	interact.v	109
291	due.adj	144	321	divide.v	128	351	attraction.n	107
292	rotation.n	144	322	operate.v	128	352	collide.v	104
293	physical.adj	142	323	decay.v	127	353	multiply.v	104
294	positron.n	142	324	sketch.v	127	354	stable.adj	104
295	iron.n	141	325	mechanical.adj	126	355	conduct.v	103
296	behaviour.n	140	326	spherical.adj	125	356	orbital.adj	103
297	cable.n	140	327	parallel.n	124	357	practice.n	103
298	copper.n	140	328	upwards.adv	124	358	spectra.n	103
299	external.adj	140	329	enable.v	122	359	terminal.adj	102
300	origin.n	140	330	balance.n	121	360	compression.n	101

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	label.v	101	391	behave.v	91	421	neutral.adj	81
362	switch.n	101	392	extend.v	91	422	sectional.adj	81
363	apparatus.n	100	393	joule.n	90	423	volt.n	81
364	relatively.adv	100	394	limit.v	90	424	attract.v	80
365	replace.v	100	395	approximate.adj	89	425	improve.v	78
366	likely.adj	99	396	distribution.n	89	426	interfere.v	78
367	mathematical.adj	99	397	background.n	88	427	account.n	77
368	separate.adj	99	398	direct.adj	88	428	camera.n	76
369	approach.v	98	399	eventually.adv	88	429	classical.adj	75
370	consequence.n	98	400	construct.v	87	430	complicated.adj	75
371	existence.n	98	401	theoretical.adj	87	431	confirm.v	75
372	chemical.adj	97	402	characteristic.n	86	432	generation.n	75
373	chemical.n	97	403	connection.n	86	433	precise.adj	75
374	ignore.v	97	404	constructive.adj	86	434	decrease.n	74
375	substitute.v	97	405	quantum.n	86	435	convenient.adj	73
376	instantaneous.adj	96	406	static.adj	86	436	discrete.adj	73
377	estimate.n	95	407	data.n	85	437	equally.adv	73
378	interpret.v	95	408	aluminium.n	84	438	gap.n	72
379	massive.adj	95	409	analysis.n	84	439	triangle.n	72
380	neutrino.n	95	410	appropriate.adj	84	440	analyse.v	71
381	surround.v	95	411	horizontally.adv	84	441	experimentally.adv	71
382	oxygen.n	94	412	examination.n	83	442	obvious.adj	71
383	prediction.n	94	413	quote.v	83	443	physic.n	71
384	normally.adv	93	414	complex.adj	82	444	sketch.n	70
385	revolution.n	93	415	deliver.v	82	445	band.n	69
386	diffract.v	93	416	liquid.adj	82	446	importance.n	68
387	previous.adj	93	417	scalar.adj	82	447	presence.n	68
388	beta.n	92	418	approach.n	81	448	attractive.adj	67
389	commonly.adv	92	419	dense.adj	81	449	essential.adj	67
390	technique.n	92	420	destructive.adj	81	450	feature.n	67

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	obey.v	65	481	repulsion.n	58	511	ms.n	47
452	outline.v	65	482	steel.n	58	512	trap.v	47
453	reverse.v	65	483	content.n	57	513	variety.n	47
454	displace.v	64	484	prove.v	57	514	construction.n	46
455	examine.v	64	485	purpose.n	57	515	extremely.adv	46
456	gain.n	64	486	repel.v	57	516	graphically.adv	46
457	switch.v	64	487	collide.n	56	517	kilogram.n	46
458	alternative.adj	63	488	immediately.adv	55	518	illuminate.v	45
459	melt.v	63	489	international.adj	55	519	precisely.adv	45
460	overcome.v	63	490	notation.n	55	520	radial.adj	45
461	respectively.adv	63	491	sensitive.adj	55	521	role.n	45
462	resultant.n	63	492	unstable.adj	55	522	accurately.adv	44
463	upper.adj	63	493	finally.adv	54	523	ensure.v	44
464	bubble.n	62	494	previously.adv	54	524	outcome.n	44
465	practical.adj	62	495	release.n	54	525	qualitatively.adv	43
466	version.n	62	496	approximation.n	53	526	verify.v	43
467	sensor.n	61	497	copy.v	53	527	mathematically.adv	42
468	sufficient.adj	61	498	representation.n	53	528	conclude.v	42
469	independent.adj	60	499	universal.adj	53	529	consistent.adj	42
470	distinguish.v	59	500	mode.n	52	530	famous.adj	42
471	electrostatic.adj	59	501	plastic.n	52	531	characteristic.adj	41
472	maintain.v	59	502	publish.v	52	532	exceed.v	41
473	aspect.n	59	503	familiar.adj	51	533	inversely.adv	41
474	atmospheric.adj	59	504	receiver.n	51	534	plot.n	41
475	eject.v	59	505	alternatively.adv	50	535	contribute.v	40
476	solenoid.n	59	506	basis.n	50	536	halve.v	40
477	context.n	58	507	repulsive.adj	50	537	hz.n	40
478	discussion.n	58	508	combined.adj	48	538	scattering.n	40
479	explosion.n	58	509	concentrate.v	48	539	tiny.adj	40
480	project.v	58	510	emerge.v	48	540	aware.adj	39

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
541	otherwise.adv	39	554	expose.v	36	567	millimetre.n	32
542	periodic.adj	39	555	occupy.v	36	568	array.n	30
543	alter.v	38	556	significance.n	36	569	lower.v	30
544	confine.v	38	557	conclusion.n	35	570	perception.n	30
545	deflect.v	38	558	rapid.adj	35	571	regular.adj	30
546	entire.adj	38	559	numerical.adj	34	572	branch.n	29
547	sodium.n	38	560	originally.adv	34	573	circumference.n	29
548	derivation.n	37	561	upwards.adj	34	574	classify.v	29
549	implication.n	37	562	gram.n	33	575	distribute.v	29
550	insulate.v	37	563	vessel.n	33	576	exchange.v	29
551	steady.adj	37	564	comment.n	32	577	orientation.n	29
552	conducting.n	36	565	geometry.n	32	578	oscillate.n	29
553	correctly.adv	36	566	influence.n	32			

The IS-AVL for Maths (379 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	function.n	6268	31	factor.n	935	61	select.v	565
2	equation.n	5521	32	formula.n	915	62	object.n	557
3	value.n	5274	33	method.n	911	63	coefficient.n	549
4	graph.n	3815	34	root.n	904	64	horizontal.adj	549
5	example.n	2576	35	domain.n	894	65	quadratic.adj	548
6	solution.n	2296	36	represent.v	893	66	speed.n	548
7	area.n	2107	37	product.n	888	67	parallel.adj	546
8	axis.n	1663	38	determine.v	881	68	vertical.adj	545
9	hence.adv	1618	39	direction.n	858	69	calculator.n	539
10	solve.v	1599	40	positive.adj	840	70	common.adj	537
11	curve.n	1439	41	interval.n	816	71	linear.adj	534
12	diagram.n	1391	42	circle.n	805	72	section.n	523
13	variable.n	1355	43	volume.n	799	73	geometric.adj	518
14	triangle.n	1329	44	range.n	795	74	maximum.adj	517
15	cm.n	1292	45	population.n	790	75	proof.n	500
16	sin.n	1223	46	equal.adj	785	76	mathematical.adj	499
17	result.n	1198	47	deviation.n	777	77	perpendicular.adj	498
18	length.n	1162	48	sketch.v	777	78	property.n	496
19	sequence.n	1158	49	therefore.adv	756	79	ratio.n	496
20	sum.n	1116	50	height.n	749	80	evaluate.v	485
21	random.adj	1104	51	define.v	739	81	asymptote.n	479
22	coordinate.n	1081	52	normal.adj	637	82	multiply.v	475
23	series.n	1077	53	constant.adj	632	83	increase.v	459
24	distance.n	1075	54	apply.v	618	84	measure.n	455
25	prove.v	1019	55	exercise.n	608	85	measure.v	441
26	chapter.n	986	56	contain.v	598	86	intersection.n	439
27	calculate.v	978	57	negative.adj	587	87	radius.n	438
28	standard.adj	961	58	describe.v	584	88	region.n	437
29	unit.n	952	59	intercept.n	573	89	occur.v	428
30	complex.adj	938	60	divide.v	571	90	include.v	424

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	km.n	417	121	minimum.adj	288	151	expand.v	222
92	require.v	416	122	origin.n	282	152	axis.adj	221
93	arithmetic.adj	412	123	process.n	280	153	inverse.n	217
94	ax.n	405	124	similar.adj	279	154	decimal.adj	216
95	definition.n	398	125	substitute.v	279	155	denominator.n	214
96	fraction.n	381	126	compare.v	277	156	label.v	214
97	differentiate.v	372	127	original.adj	276	157	significant.adj	214
98	integration.n	370	128	decrease.v	275	158	graph.v	213
99	initial.adj	367	129	rational.adj	272	159	correspond.v	212
100	substitution.n	366	130	scale.n	271	160	experiment.n	212
101	correct.adj	364	131	binomial.adj	269	161	pattern.n	211
102	outcome.n	364	132	independent.adj	269	162	intersect.v	210
103	variable.adj	361	133	model.v	269	163	discuss.v	209
104	model.n	359	134	temperature.n	269	164	rotate.v	208
105	argument.n	357	135	discrete.adj	264	165	algebraic.adj	207
106	non.adj	354	136	increase.n	264	166	constant.n	204
107	vertex.n	352	137	application.n	261	167	multiple.n	201
108	transformation.n	349	138	indicate.v	258	168	denote.v	200
109	metre.n	347	139	illustrate.v	250	169	approach.n	199
110	produce.v	340	140	differentiation.n	248	170	appropriate.adj	197
111	estimate.v	338	141	mode.n	241	171	quantity.n	197
112	notation.n	336	142	condition.n	236	172	relative.adj	196
113	principle.n	330	143	remain.v	235	173	minimum.n	195
114	period.n	325	144	identify.v	231	174	provide.v	194
115	surface.n	318	145	construct.v	229	175	satisfy.v	191
116	assume.v	312	146	exist.v	228	176	combination.n	187
117	inverse.adj	298	147	infinite.adj	226	177	result.v	185
118	involve.v	295	148	percentage.n	226	178	dimension.n	184
119	rectangle.n	294	149	previous.adj	223	179	mathematician.n	182
120	displacement.n	292	150	calculation.n	222	180	repeat.v	182

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	operation.n	178	211	accurate.adj	139	241	graphically.adv	112
182	reflect.v	177	212	feature.n	134	242	copy.v	111
183	corresponding.adj	176	213	likely.adj	134	243	invest.v	111
184	otherwise.adv	176	214	numerator.n	134	244	gram.n	109
185	cube.n	175	215	conclusion.n	132	245	reduce.v	108
186	approximation.n	173	216	combine.v	131	246	eliminate.v	106
187	dimensional.adj	173	217	plot.v	131	247	finite.adj	105
188	respectively.adv	173	218	refer.v	131	248	deduce.v	104
189	predict.v	171	219	simultaneous.adj	131	249	integral.adj	104
190	reflection.n	167	220	solid.adj	131	250	behaviour.n	101
191	parallelogram.n	166	221	subtract.v	130	251	computer.n	101
192	approximately.adv	165	222	composite.adj	128	252	conjugate.n	101
193	digit.n	164	223	final.adj	128	253	summary.n	101
194	mass.n	164	224	symbol.n	128	254	classify.v	100
195	unique.adj	164	225	compound.n	126	255	introduce.v	100
196	replace.v	162	226	consist.v	126	256	analysis.n	99
197	prime.adj	161	227	extend.v	124	257	assumption.n	97
198	interpret.v	158	228	effect.n	121	258	investigate.v	97
199	estimate.n	157	229	theorem.n	121	259	path.n	97
200	perform.v	157	230	proportion.n	119	260	consecutive.adj	94
201	maximum.n	156	231	verify.v	118	261	distinct.adj	93
202	multiplication.n	156	232	convert.v	116	262	numerical.adj	93
203	generate.v	153	233	justify.v	116	263	average.n	91
204	revolution.n	153	234	derive.v	115	264	split.v	91
205	trial.n	149	235	receive.v	115	265	affect.v	90
206	quotient.n	145	236	comment.n	114	266	approximate.adj	90
207	theory.n	145	237	cubic.adj	114	267	rectangular.adj	90
208	concept.n	143	238	fundamental.adj	114	268	separate.adj	90
209	column.n	141	239	create.v	112	269	practice.n	88
210	directly.adv	141	240	establish.v	112	270	interpretation.n	87

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	language.n	86	301	various.adj	64	331	purpose.n	51
272	profit.n	86	302	consistent.adj	63	332	correspond.n	50
273	sketch.n	86	303	passenger.n	63	333	differ.v	50
274	connect.v	85	304	branch.n	62	334	multiple.adj	50
275	imply.v	85	305	conclude.v	62	335	associate.v	49
276	specific.adj	85	306	formulae.adj	62	336	boundary.n	49
277	theorem.v	85	307	formulae.n	62	337	accurately.adv	48
278	vary.v	84	308	spread.v	62	338	wire.n	48
279	vertically.adv	83	309	balloon.n	61	339	current.adj	47
280	equally.adv	82	310	cell.n	61	340	individual.n	46
281	intersect.adj	81	311	simultaneously.adv	61	341	tennis.n	46
282	separate.v	81	312	task.n	60	342	alternative.adj	45
283	algebraically.adv	80	313	earn.v	56	343	essential.adj	45
284	familiar.adj	77	314	individual.adj	56	344	specify.v	45
285	examine.v	76	315	avoid.v	55	345	subtraction.n	45
286	selection.n	73	316	commonly.adv	55	346	description.n	43
287	algebra.n	72	317	entire.adj	55	347	gap.n	43
288	context.n	72	318	identical.adj	55	348	op.n	43
289	bacteria.n	71	319	replacement.n	55	349	purchase.v	43
290	equal.v	71	320	statistical.adj	55	350	normal.n	42
291	calculate.n	70	321	achieve.v	54	351	frequently.adv	41
292	explore.v	69	322	enable.v	54	352	alternatively.adv	40
293	ladder.n	69	323	equilateral.adj	54	353	convenient.adj	40
294	exceed.v	68	324	male.n	54	354	cot.n	40
295	characteristic.n	67	325	remove.v	54	355	detail.n	40
296	chord.n	67	326	separately.adv	54	356	triangular.adj	40
297	procedure.n	67	327	attempt.n	53	357	centimetre.n	38
298	accuracy.n	66	328	dollar.n	53	358	convergent.n	38
299	restriction.n	64	329	link.v	53	359	decrease.n	38
300	substance.n	64	330	major.adj	52	360	finally.adv	38

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	sum.v	38	368	current.n	33	375	focus.v	31
362	physics.n	37	369	sufficient.adj	33	376	reciprocal.n	31
363	code.n	36	370	immediately.adv	32	377	switch.v	31
364	material.n	36	371	repeat.n	32	378	design.v	30
365	rely.v	36	372	symmetrical.adj	32	379	vehicle.n	30
366	visual.adj	36	373	arise.v	31			
367	display.n	35	374	cancel.v	31			

The IS-AVL for Theory of Knowledge (685 words) Frequency in Tokens Per Million (TPM)

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
1	knowledge.n	8943	31	tool.n	694	61	therefore.adv	425
2	example.n	2821	32	produce.v	680	62	development.n	421
3	language.n	2614	33	datum.n	678	63	affect.v	416
4	perspective.n	1499	34	issue.n	666	64	analysis.n	416
5	claim.n	1366	35	identify.v	657	65	involve.v	408
6	belief.n	1266	36	provide.v	657	66	approach.n	398
7	area.n	1183	37	context.n	635	67	observation.n	388
8	method.n	1166	38	argue.v	613	68	individual.adj	385
9	ethical.adj	1162	39	mathematical.adj	607	69	specific.adj	380
10	object.n	1162	40	source.n	593	70	various.adj	379
11	theory.n	1087	41	result.n	586	71	discussion.n	375
12	evidence.n	1079	42	cultural.adj	566	72	section.n	371
13	culture.n	1061	43	experiment.n	549	73	common.adj	365
14	community.n	1049	44	behaviour.n	513	74	tradition.n	357
15	develop.v	1009	45	century.n	510	75	opinion.n	354
16	explore.v	1008	46	model.n	510	76	effect.n	353
17	include.v	892	47	influence.v	506	77	response.n	352
18	argument.n	887	48	discuss.v	505	78	scope.n	349
19	value.n	864	49	exist.v	496	79	expert.n	348
20	social.adj	863	50	practice.n	490	80	activity.n	347
21	concept.n	843	51	assumption.n	489	81	skill.n	341
22	chapter.n	842	52	apply.v	487	82	refer.v	340
23	extent.n	774	53	decision.n	486	83	justify.v	337
24	create.v	752	54	research.n	486	84	bias.n	332
25	describe.v	750	55	ethic.n	480	85	moral.adj	332
26	role.n	741	56	require.v	474	86	significant.adj	332
27	process.n	726	57	interpretation.n	472	87	similar.adj	324
28	society.n	712	58	conclusion.n	458	88	authority.n	323
29	individual.n	707	59	claim.v	452	89	assume.v	320
30	event.n	699	60	implication.n	436	90	emotion.n	318

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
91	non.adj	313	121	arise.v	253	151	contain.v	204
92	compare.v	311	122	pattern.n	252	152	seek.v	202
93	faith.n	305	123	economic.adj	246	153	examine.v	201
94	observe.v	299	124	influence.n	245	154	challenge.v	199
95	construct.v	297	125	challenge.n	244	155	rely.v	197
96	reflect.v	296	126	factor.n	244	156	objective.adj	195
97	principle.n	289	127	link.n	241	157	select.v	194
98	represent.v	286	128	image.n	240	158	global.adj	193
99	impact.n	284	129	structure.n	239	159	predict.v	193
100	physical.adj	284	130	basis.n	238	160	prove.v	193
101	connection.n	283	131	article.n	237	161	application.n	191
102	define.v	283	132	material.n	236	162	introduction.n	190
103	focus.v	282	133	product.n	234	163	academic.adj	188
104	accord.v	281	134	production.n	234	164	accurate.adj	187
105	feature.n	278	135	task.n	231	165	debate.n	187
106	interpret.v	276	136	belong.v	228	166	topic.n	187
107	complex.adj	274	137	phenomenon.n	228	167	universal.adj	187
108	purpose.n	273	138	account.n	227	168	remain.v	186
109	reliable.adj	271	139	environment.n	227	169	psychology.n	184
110	modern.adj	270	140	aspect.n	225	170	future.n	182
111	establish.v	269	141	description.n	225	171	appropriate.adj	181
112	gain.v	269	142	physics.n	223	172	condition.n	181
113	link.v	265	143	communicate.v	222	173	limit.v	181
114	false.adj	262	144	justification.n	222	174	aware.adj	180
115	definition.n	260	145	population.n	217	175	philosopher.n	180
116	identity.n	257	146	available.adj	215	176	publish.v	180
117	consequence.n	257	147	metaphor.n	214	177	avoid.v	179
118	likely.adj	257	148	relevant.adj	213	178	conflict.n	178
119	determine.v	256	149	universe.n	209	179	demonstrate.v	178
120	access.n	253	150	standard.n	205	180	range.n	178

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
181	distinction.n	177	211	function.n	153	241	institution.n	133
182	existence.n	176	212	multiple.adj	153	242	specie.n	133
183	prediction.n	176	213	category.n	152	243	fundamental.adj	132
184	cognitive.adj	174	214	occur.v	152	244	rational.adj	132
185	engage.v	173	215	attention.n	149	245	evolution.n	131
186	attempt.n	171	216	effort.n	149	246	ancient.adj	130
187	correct.adj	171	217	brain.n	148	247	insight.n	129
188	investigate.v	171	218	direct.adj	148	248	appreciate.v	128
189	gender.n	170	219	achieve.v	147	249	character.n	128
190	reveal.v	170	220	content.n	147	250	familiar.adj	128
191	contribute.v	169	221	detail.n	147	251	objectivity.n	128
192	increase.v	169	222	essential.adj	147	252	result.v	127
193	quality.n	169	223	practical.adj	147	253	version.n	127
194	connect.v	168	224	promote.v	147	254	background.n	126
195	generation.n	167	225	circumstance.n	146	255	disagree.v	126
196	acquire.v	166	226	status.n	146	256	author.n	125
197	perform.v	165	227	significance.n	145	257	literature.n	125
198	construction.n	164	228	economic.n	144	258	speech.n	125
199	criterion.n	163	229	limit.n	144	259	contribution.n	124
200	characteristic.n	162	230	extend.v	142	260	critical.adj	124
201	communication.n	162	231	emotional.adj	140	261	entirely.adv	124
202	final.adj	160	232	enable.v	140	262	unique.adj	124
203	property.n	160	233	period.n	139	263	logic.n	123
204	importance.n	159	234	benefit.n	136	264	oppose.v	123
205	original.adj	159	235	directly.adv	136	265	current.adj	122
206	opportunity.n	158	236	maintain.v	136	266	obvious.adj	122
207	measure.v	157	237	outcome.n	135	267	respond.v	122
208	design.v	156	238	illustrate.v	134	268	search.n	122
209	creation.n	154	239	analyse.v	134	269	distinguish.v	121
210	associate.v	153	240	effective.adj	133	270	primary.adj	120

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
271	accuracy.n	119	301	shift.n	107	331	aim.v	97
272	disagreement.n	119	302	ensure.v	106	332	empirical.adj	97
273	job.n	119	303	interaction.n	106	333	positive.adj	97
274	logical.adj	119	304	investigation.n	106	334	value.v	97
275	generate.v	118	305	reduce.v	106	335	intention.n	96
276	abstract.adj	116	306	awareness.n	105	336	similarity.n	96
277	perceive.v	116	307	finally.adv	105	337	appeal.v	95
278	planet.n	116	308	methodology.n	105	338	famous.adj	95
279	encounter.v	115	309	subjective.adj	105	339	invent.v	95
280	similarly.adv	115	310	origin.n	104	340	scale.n	94
281	attitude.n	114	311	critic.n	103	341	divide.v	93
282	peer.n	114	312	progress.n	103	342	error.n	93
283	reliability.n	114	313	broad.adj	102	343	convey.v	92
284	conduct.v	113	314	encourage.v	101	344	direction.n	92
285	formal.adj	113	315	intend.v	101	345	university.n	92
286	improve.v	113	316	introduce.v	101	346	future.adj	91
287	translate.v	113	317	respect.n	101	347	visual.adj	91
288	variety.n	112	318	behave.v	100	348	assess.v	90
289	differ.v	111	319	design.n	100	349	capture.v	90
290	fail.v	110	320	exploration.n	100	350	consistent.adj	90
291	independent.adj	109	321	phrase.n	100	351	contrast.n	90
292	popular.adj	109	322	successful.adj	100	352	poem.n	90
293	psychologist.n	109	323	valid.adj	100	353	vary.v	90
294	representation.n	109	324	expertise.n	99	354	attempt.v	89
295	solution.n	109	325	ignore.v	99	355	guide.v	89
296	negative.adj	108	326	philosophy.n	99	356	medical.adj	88
297	indicate.v	107	327	pursuit.n	99	357	military.adj	88
298	remove.v	107	328	biological.adj	98	358	vast.adj	88
299	repeat.v	107	329	focus.n	98	359	distinct.adj	87
300	series.n	107	330	major.adj	98	360	prejudice.n	87

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
361	receive.v	87	391	atom.n	79	421	merely.adv	71
362	relative.adj	87	392	conclude.v	79	422	psychological.adj	71
363	ultimately.adv	87	393	disease.n	79	423	convince.v	70
364	underlie.v	87	394	diverse.adj	79	424	dimension.n	70
365	complexity.n	86	395	eventually.adv	78	425	exercise.n	70
366	strength.n	86	396	imply.v	78	426	statistical.adj	70
367	style.n	86	397	item.n	78	427	surround.v	70
368	aim.n	86	398	manipulate.v	78	428	access.v	69
369	lack.n	86	399	observer.n	78	429	domain.n	69
370	comparison.n	85	400	team.n	78	430	industry.n	69
371	equal.adj	85	401	constitute.v	77	431	internal.adj	69
372	legal.adj	85	402	previous.adj	77	432	minority.n	69
373	nevertheless.adv	85	403	propose.v	77	433	revolution.n	69
374	document.n	84	404	technical.adj	77	434	unethical.adj	69
375	equally.adv	84	405	coherent.adj	76	435	virtue.n	69
376	pursue.v	84	406	expand.v	76	436	vote.v	69
377	reject.v	84	407	majority.n	76	437	deliberately.adv	68
378	review.n	83	408	navigate.v	76	438	correlation.n	67
379	success.n	83	409	approach.v	75	439	foundation.n	67
380	gather.v	82	410	possess.v	75	440	potential.adj	67
381	limited.adj	82	411	invite.v	74	441	recognition.n	67
382	participate.v	82	412	limitation.n	74	442	effectively.adv	66
383	advantage.n	81	413	normal.adj	74	443	evolve.v	66
384	creative.adj	80	414	combine.v	73	444	sufficient.adj	66
385	emerge.v	80	415	sophisticated.adj	73	445	cite.v	65
386	otherwise.adv	80	416	suffer.v	73	446	commonly.adv	65
387	pose.v	80	417	consist.v	72	447	literally.adv	65
388	professional.adj	80	418	invention.n	72	448	loss.n	65
389	relatively.adv	80	419	articulate.v	71	449	alternative.adj	64
390	acknowledge.v	79	420	embed.v	71	450	conscious.adj	64

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
451	valuable.adj	64	481	weak.adj	57	511	appeal.n	51
452	commit.v	63	482	accurately.adv	56	512	instruction.n	51
453	diagram.n	63	483	distance.n	56	513	lack.v	51
454	operate.v	63	484	energy.n	56	514	classical.adj	50
455	site.n	63	485	exhibit.v	56	515	ethically.adv	50
456	adult.n	62	486	extreme.adj	56	516	examination.n	50
457	contrast.v	62	487	extremely.adv	56	517	interact.v	50
458	survive.v	62	488	outline.v	56	518	potential.n	50
459	critically.adv	61	489	pre.adj	56	519	resolve.v	50
460	precise.adj	61	490	regardless.adv	56	520	traditionally.adv	50
461	procedure.n	61	491	separate.adj	56	521	alter.v	49
462	compete.v	61	492	cell.n	55	522	estimate.v	49
463	trial.n	61	493	failure.n	55	523	non.n	49
464	entire.adj	60	494	illusion.n	55	524	rigorous.adj	49
465	harm.n	60	495	pressure.n	55	525	testimony.n	49
466	deny.v	59	496	primarily.adv	55	526	accessible.adj	48
467	derive.v	59	497	detailed.adj	54	527	accompany.v	48
468	favour.n	59	498	initial.adj	54	528	dynamic.n	48
469	financial.adj	59	499	label.v	54	529	ethnic.adj	48
470	highlight.v	59	500	research.v	54	530	unlikely.adj	48
471	root.n	59	501	shift.v	54	531	ancestor.n	47
472	weakness.n	59	502	complicated.adj	53	532	comment.n	47
473	dismiss.v	58	503	laboratory.n	53	533	copy.n	47
474	legitimate.adj	58	504	replace.v	53	534	currently.adv	47
475	overcome.v	58	505	temperature.n	53	535	desirable.adj	47
476	precisely.adv	58	506	calculate.v	52	536	display.v	47
477	immediately.adv	57	507	explicit.adj	52	537	feature.v	47
478	previously.adv	57	508	impose.v	52	538	impression.n	47
479	surface.n	57	509	knowledge.v	52	539	independence.n	47
480	trade.n	57	510	account.v	51	540	landscape.n	47

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
541	speed.n	47	571	falsify.v	43	601	interview.n	39
542	struggle.v	47	572	mass.n	43	602	knowledgeable.adj	39
543	sum.n	47	573	persuade.v	43	603	print.v	39
544	expose.v	46	574	submit.v	43	604	rarely.adv	39
545	inevitable.adj	46	575	capital.n	42	605	review.v	39
546	straightforward.adj	46	576	controversial.adj	42	606	somewhat.adv	39
547	witness.n	46	577	declare.v	42	607	subjectivity.n	39
548	abandon.v	45	578	distribute.v	42	608	creature.n	38
549	average.adj	45	579	lesson.n	42	609	fundamentally.adv	38
550	excellent.adj	45	580	motivate.v	42	610	alive.adj	37
551	patient.n	45	581	poverty.n	42	611	attend.v	37
552	presence.n	45	582	remind.v	42	612	blind.adj	37
553	summary.n	45	583	consciousness.n	41	613	explicitly.adv	37
554	accuse.v	44	584	conservative.adj	41	614	exploit.v	37
555	capable.adj	44	585	eliminate.v	41	615	foreign.adj	37
556	ceremony.n	44	586	external.adj	41	616	formation.n	37
557	implicit.adj	44	587	function.v	41	617	incorporate.v	37
558	increase.n	44	588	illness.n	41	618	inevitably.adv	37
559	length.n	44	589	prefer.v	41	619	random.adj	37
560	literal.adj	44	590	publication.n	41	620	refuse.v	37
561	punishment.n	44	591	purely.adv	41	621	agency.n	36
562	range.v	44	592	chemical.n	40	622	compelling.adj	36
563	significantly.adv	44	593	component.n	40	623	deliver.v	36
564	association.n	43	594	income.n	40	624	effectiveness.n	36
565	attribute.v	43	595	operation.n	40	625	expansion.n	36
566	broadly.adv	43	596	stance.n	40	626	formulate.v	36
567	contradict.v	43	597	confident.adj	39	627	peace.n	36
568	deliberate.adj	43	598	distort.v	39	628	restrict.v	36
569	division.n	43	599	essentially.adv	39	629	vulnerable.adj	36
570	due.adj	43	600	evolutionary.adj	39	630	constraint.n	35

Rank	Lemma	TPM	Rank	Lemma	TPM	Rank	Lemma	TPM
631	independently.adv	35	650	defend.v	33	669	opposite.n	31
632	insist.v	35	651	individually.adv	33	670	progress.v	31
633	map.v	35	652	match.v	33	671	succeed.v	31
634	objective.n	35	653	precision.n	33	672	unit.n	31
635	overlap.v	35	654	profound.adj	33	673	violate.v	31
636	calculation.n	34	655	proposition.n	33	674	dispute.n	30
637	commitment.n	34	656	respect.v	33	675	irrelevant.adj	30
638	famously.adv	34	657	scene.n	33	676	poet.n	30
639	geography.n	34	658	appreciation.n	32	677	prime.adj	30
640	guarantee.v	34	659	demand.v	32	678	rapidly.adv	30
641	modelling.n	34	660	guidance.n	32	679	speculate.v	30
642	numerous.adj	34	661	successfully.adv	32	680	subsequent.adj	30
643	plausible.adj	34	662	target.n	32	681	definitive.adj	29
644	regular.adj	34	663	actively.adv	31	682	molecule.n	29
645	subtle.adj	34	664	altogether.adv	31	683	occupy.v	29
646	unconscious.adj	34	665	correctly.adv	31	684	predictive.adj	29
647	accident.n	33	666	extensive.adj	31	685	release.v	29
648	advanced.adj	33	667	immoral.adj	31			
649	debate.v	33	668	merit.n	31			