

**AN ASSESSMENT OF STUDENTS' ENGLISH VOCABULARY LEVELS AND
AN EXPLORATION OF THE VOCABULARY PROFILE OF TEACHERS'
SPOKEN DISCOURSE IN AN INTERNATIONAL HIGH SCHOOL**

by

Graham Robert Creighton

Submitted in accordance with the requirements for the degree of

MASTER OF ARTS

In the subject

APPLIED LINGUISTICS

at the

University of South Africa

Supervisor: Prof E.J. Pretorius

Co-supervisor: Dr. N.M. Klapwijk

October 2016

DECLARATION

Name: Graham Robert Creighton

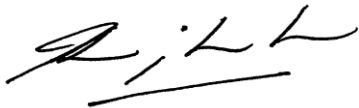
Student number: 3005-961-5

Degree: Master of Arts in Applied Linguistics

Exact wording of the title of the dissertation as appearing on the copies submitted for the examination:

**AN ASSESSMENT OF STUDENTS' ENGLISH VOCABULARY LEVELS AND
AN EXPLORATION OF THE VOCABULARY PROFILE OF TEACHERS'
SPOKEN DISCOURSE IN AN INTERNATIONAL HIGH SCHOOL**

I declare that the above dissertation is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.



SIGNATURE

21 March 2017

DATE

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor Lilli Pretorius, and my co-supervisor, Dr Nanda Klapwijk, for their support throughout my work on this dissertation. Not only am I grateful for their excellent advice regarding my dissertation, but also for their encouragement and understanding as I juggled study, work and family commitments.

I am also thankful to the support from my loving wife, Ikuko, and my parents Denis and Pam for their love throughout my life leading me to have the opportunity to study for a Master's degree. Thank you, too, to my two boys, Thomas and Andy, for always knowing how to make me smile.

Finally, I want to thank the students and my colleagues and friends at FIS who helped me conduct my research.

ABSTRACT

In many international schools where English is the language of learning and teaching there are large percentages of students whose first language is not English. Many of these students may have low vocabulary levels which inhibits their chances of taking full advantage of their education. Low vocabulary levels can be a particular problem for students in mainstream classes where fluent English speaking teachers are using English to teach content areas of Mathematics, Science and History. Not only do students have to comprehend the low-frequency, academic and technical vocabulary pertaining to the subject, but they also need to know the higher frequency vocabulary that makes up general English usage. If students' vocabulary levels fall too far below the vocabulary levels with which their teachers are speaking, then their chance of comprehending the topic is small, as is their chance of succeeding in their subjects.

This study has two broad aims. Firstly, I have set out to assess the English vocabulary levels of students at an international school where English is the language of learning and teaching. The majority of students at this school do not have English as their first language. The second aim of this study is to explore the vocabulary profile of the teachers' spoken discourse at the research school. By gaining a better understanding of the nature of teacher discourse – specifically the percentage of high, mid and low-frequency vocabulary, as well as academic vocabulary that they use – English as a Second Language (ESL) teachers will be in a stronger position to identify what the vocabulary learning task is and be able to assist students in reaching the vocabulary levels necessary to make sense of their lessons. This study revealed a large gap between the generally low vocabulary levels of ESL students and the vocabulary levels spoken by their teachers. As a result the need for explicit vocabulary instruction and learning is shown to be very important in English medium (international) schools, where there are large numbers of students whose first language is not English.

KEY TERMS

English as a second language (ESL), first language (L1), general service vocabulary, academic vocabulary, technical vocabulary, basic interpersonal communicative skills (BICS), cognitive/academic language proficiency (CALP), word family, high-frequency vocabulary, mid-frequency vocabulary, low-frequency vocabulary, corpus/corpora, productive vocabulary levels test (PVLTL), general service list (GSL), academic word list (AWL), content words, controlled productive vocabulary, formulaic sequences, function words, headword, lemma, word family, word frequency, technical vocabulary, English for academic purposes (EAP), incidental learning, intentional learning, lexis, collocation, lexico-grammar, concordance, quantitative research, qualitative research, corpus-driven approach, corpus-based approach

TABLE OF CONTENTS

Chapter 1: Introduction

1.1	Introduction	1
1.2	Context of the research problem	2
1.3	Research problem	4
1.4	Research aims	6
1.5	Research questions	6
1.6	Overview of the study	7
1.7	Significance of the study	8
1.8	Parameters of the study	9
1.9	Definition of terms	9
1.10	Summary of the remainder of this dissertation	10

Chapter 2: Literature review

2.1	Introduction	11
2.2	Vocabulary	11
2.2.1	Some key definitions	12
	‘Type’ and ‘Token’	12
	‘Word family’ and ‘Lemma’	12
2.2.2	Types of vocabulary	13
2.2.2.1	Word class	14
2.2.2.2	Formulaic sequences	14
2.2.2.3	Content and function words	15
2.2.2.4	Frequency	15
2.2.2.5	How much vocabulary does a native English speaker have?	16
2.2.2.6	What constitutes high frequency vocabulary?	17
2.2.2.7	Mid-Low frequency words	18
2.2.2.8	How much vocabulary does an ESL high school student need?	20
2.2.2.9	General, academic and technical vocabulary	20
2.2.2.9.1	General vocabulary	21
	<i>The general service list (GSL)</i>	21
	<i>The new general service list (NGSL)</i>	21
2.2.2.9.2	Academic vocabulary	22
	<i>The academic word list (AWL)</i>	23
	<i>The new academic word list (NAWL)</i>	24
2.2.2.9.3	Technical vocabulary	24

2.2.2.10	Written and spoken vocabulary	25
2.2.2.11	Receptive and productive vocabulary	26
	<i>Laufer and Nation's (1999) productive vocabulary levels test(PVLT)</i>	27
2.2.3	What vocabulary does a high school student need and its implication for teaching?	28
2.2.4	How many repetitions do learners need exposure to for the uptake of new words?	30
2.2.5	Incidental vs intentional learning of vocabulary	31
	<i>Depth of knowledge</i>	32
	<i>Pedagogical implications</i>	33
2.3	Corpus linguistics	35
2.3.1	What is a corpus?	34
2.3.2	Types of corpora	35
2.3.2.1	Generalised corpora	35
2.3.2.2	Specialised corpora	36
2.3.2.3	The value of a small specialised corpus	38
2.3.2.4	Written vs spoken corpora	40
2.3.3		
2.3.3	Analysing a corpus	41
2.3.3.1	VocabProfile	41
2.3.3.2	WordSmith tools	41
2.3.4	Corpora and language teaching	43
2.3.5	Issues surrounding building a spoken corpus	46
2.4	Conclusion	47

Chapter 3: Research method

3.1	Introduction	48
3.2	Type of research	48
3.2.1	Approach	48
3.2.1.1	Size, diversity, representativeness and balance	49
3.2.1.2	Reliability and validity of the productive vocabulary levels test PVLt.....	50
3.2.2	Theoretical purpose	52
3.3	School context	52
3.4	Participants	53
3.4.1	Teachers	53

3.4.2	Students	54
3.5	Research instruments	56
3.5.1	Voice recorder	56
3.5.2	VocabProfile	56
3.5.3	WordSmith tools	57
3.5.4	Productive vocabulary levels test (PVLТ)	57
3.6	Pilot study	58
3.7	Research procedures for main study.....	60
3.7.1	The administration of the PVLТ	60
3.7.2	Data capturing and analysis of the vocabulary data	62
3.7.3	Procedures for the administration of the recordings	62
3.8	Data processing and analysis of corpus	65
3.9	Ethical considerations	65
3.10	Conclusion	65

Chapter 4: Findings

4.1	Introduction	67
4.2	Research question 1	67
4.3	Research question 2	70
4.3.1	Research question 2(a).....	70
4.3.2	Research question 2(b)	72
4.3.3	Research question 2(c)	77
4.3.4	Research question 2(d)	79
4.3.5	Research question 2(e)	88
4.4	Discussion	93
4.4.1	What do the findings suggest about the type of vocabulary students need in order to comprehend the words spoken by their teachers?	93
	<i>General English vocabulary</i>	93
	<i>Academic vocabulary</i>	94
	<i>Technical vocabulary</i>	94
4.4.2	What do the findings suggest about the nature of the high school sub corpora?.....	95
4.4.3	What do the findings suggest about the similarities and differences between the FIS corpus and corpora taken from universities?	95
4.4.4	What do the findings of this study reveal about the vocabulary levels of the FIS students and vocabulary levels they should have in order to understand the words spoken by their teachers?	96
	<i>The low vocabulary levels of the FIS students</i>	96

<i>Why knowledge of 6,000 words should be an immediate priority for the students at FIS</i>	96
4.4.5 What factors in addition to vocabulary levels help or hinder students' comprehension of spoken discourse?	97
4.5 Conclusion	98
Chapter 5: Conclusion	
5.1 Introduction	100
5.2 Review	100
5.2.1 Summary of findings	100
5.2.2 Contributions of the study	104
5.2.2.3 Pedagogical implications	105
<i>The extent of the learning task</i>	105
<i>Teaching approach to improving students' vocabulary levels</i>	105
<i>Materials development</i>	106
5.2.3 Limitations of the study	107
5.3 Recommendations for further research	108
5.4 Conclusion	109
References	111
Appendix A	117
Appendix B	127
Appendix C	135
Appendix D	139
Appendix E	143

LIST OF TABLES

Table 2.1	Vocabulary <i>types</i> in the FIS corpus at the BNC/COCA 7,000 word level	12
Table 2.2	<i>Word families</i> in the FIS corpus at the BNC/COCA 7,000 word level	13
Table 2.3	Sample questions from Laufer and Nation’s (1999) Productive Vocabulary Levels Test at the 2,000 word level	28
Table 2.4	Shared key verbs in business, linguistics and medicine (Granger and Paquot, 2010)	29
Table 2.5	Sample of new general service list in order of frequency	42
Table 3.1	PVLT reliability results as conducted by Laufer & Nation (1999)	51
Table 3.2	English L2 participants’ scores during Laufer & Nation’s testing of PVLTs’ validity (1999)	51
Table 3.3	Teacher backgrounds	53
Table 3.4	Student backgrounds	55
Table 3.5	Most frequently spoken words of FIS Science teachers for Pilot Study	60
Table 3.6	Number of words used across four subject areas taken from BAWE and MICASE corpora	64
Table 4.1	Presentation of results from PVLTs.....	68
Table 4.2	NGSL and NAWL frequency profile of FIS corpus.....	71
Table 4.3	BNC/COCA frequency profile of FIS corpus.....	73
Table 4.4	Off-list words (words not in the BNC/COCA list).....	75
Table 4.5	NGSL and NAWL frequency profile of FIS sub-corpora	78
Table 4.6	Word frequency comparisons across 4 corpora	81
Table 4.7	Frequency of common nouns across four corpora	84
Table 4.8	Comparison of lower frequency words in the FIS corpus and samples of the MICASE and BAWE corpora	86
Table 4.9	Profile chart comparison of FIS corpus and MICASE and BAWE samples	88
Table 4.10	Headwords of Coxhead’s Academic Word List word families in the FIS corpus	90
Table 4.11	The number of AWL word families in the FIS corpus from each AWL sublist	92

CHAPTER 1

INTRODUCTION

1.1 Introduction

One of my roles as an English as a second language (ESL) teacher at an international high school in Japan, was to support ESL students in some of their History and Science lessons. During these times I would sit alongside the ESL students so that I could give language support to them when needed and when appropriate without interrupting the lesson. In my one-to-one dealings with these students in the History and Science classroom and then during my own ESL lessons with small groups of these students, I came to recognise that many of them seemed to have a very poor vocabulary. While giving in-class support to these students I would listen to the History and Science teachers teaching their subject and get the sense that much of what they said was not comprehended by the ESL students. Later, when given the opportunity to ask students about what they had understood from their content teachers, it seemed clear that they had indeed failed to grasp much of the content.

I speculated about the factors that could be the cause of the ESL students not grasping what their teachers were saying. I speculated that it could be a difference between what I perceived to be their poor vocabulary levels and the potentially higher vocabulary levels used by the teachers. However, it could also have been any of the following causes or a combination thereof: the teachers' (unfamiliar) accents or dialects of English, the teachers' use of idiomatic language, the speed at which teachers talk, learners' lack of background knowledge in the subject, or the unstructured and unpredictable structure that often comes with unscripted, spontaneous speech. Rather than trying to account for all the possible causes of the problem, I set out to understand the nature of the vocabulary spoken by the teachers – with my main interest being how much high and low frequency vocabulary was being used, since as Schmitt (2010, 68) posits, 'frequency is an absolutely crucial factor to consider in vocabulary research'. I also wanted to know what the vocabulary levels of my ESL students were, to see what the discrepancy was, if any, between their vocabulary levels and the levels at which the teachers were speaking. I felt that if I knew what the potential vocabulary gap was, I could use my ESL class time to teach the 'missing' vocabulary, and thus bring my students' vocabulary up to that of what their teachers were using. These ideas formed the basis for this study. My study, therefore, aims to reveal what the nature of high school

teachers' spoken discourse is, in terms of what percentage of their discourse contains low, mid and high frequency vocabulary across four subject areas (English, History, Mathematics and Science), and in terms of what percentage of the teachers' speech comprises academic vocabulary. In addition to analysing the vocabulary frequency levels of teachers' spoken discourse, this study will assess the vocabulary size in terms of frequency levels of the high school students who were being taught by the teachers whose speech is analysed.

1.2 Context of the research problem

International schools, like the one where this study was conducted, are not language schools, where the focus is building students' general English language knowledge and where students may be placed at an appropriate level based on their language ability and their language skills systematically built up. Instead, international schools are like any 'traditional' school where a range of academic disciplines, such as humanities and the sciences, are taught and where teachers assume for the most part that their students understand the language of instruction. Many (but not all) teachers at international schools are sensitive to having non-native speakers in their classroom but their primary job is to teach the subject, not the language, and therefore these teachers, who for the most part are English mother-tongue speakers, often use vocabulary that is beyond the comprehension of many of their students.

FIS, as I will call the school where this study was done, has a large ESL population – approximately 33% of the students have Japanese and about 35% have Korean as their first language (L1). While some of these students had schooling at English medium schools in other countries, the vast majority came to FIS straight out of the Japanese or Korean school systems, where their schooling was in their L1. Since very little English is used in the city where the students of FIS live, there are few opportunities outside of school for students to develop their English skills.

Since the school has a large ESL population, the school's administration and ESL teachers try to provide a good ESL programme. A good ESL programme, according to Carder (2011, 53), is one where beginner ESL students start off their international school education in a beginner ESL class, where students receive language tuition (instead of being immediately mainstreamed), and then over the course of a year exit into 'controlled stages of sheltered instruction' in the content areas of Mathematics, Science and Humanities. Such classes would have teachers who are trained in second language tuition and who have sufficient knowledge of the content areas. In these classes students would be more likely to access subject matter,

because the language (sentence structure and vocabulary) used by the teachers is ‘at their level’ and therefore comprehensible. Once students show sufficient competence during the ‘sheltered instruction’ stage they become eligible to enter the mainstream classroom.

The unfortunate reality is that it is difficult to implement such a programme in a school setup. Firstly, it requires extra staffing and classrooms – a financial cost to the school. Secondly, there is pressure, particularly from parents, to rush ESL students into the mainstream as soon as possible and often before they are ready. Many students do not want the stigma of being in a class they see as ‘weaker than normal’ and parents want to see their money well-spent, with their children obtaining a recognisable and accredited qualification that would enable acceptance into English-medium universities. Carder (2011, 4) illustrates the problem of rushing ESL students into mainstream classes by stating that,

the weakness of many ESL programmes is that students are transferred to the mainstream before they have acquired enough SLIC (Second Language Instructional Competence) to do well in content classes. 5-8 years is the time shown by research for SLLs (second language learners) to score at the 50th percentile on tests of reading comprehension in English.

Unfortunately FIS is unable to provide a programme that can support beginner ESL students for their entire school day because there are only two ESL teachers who support the ESL students in Grades 6 through 10. During these grades there are six English lessons per week, four of which combine all the ESL and non-ESL students, leaving two lessons per week for the ESL teacher to work with the ESL students only. Grade 9 and 10 students have two elective classes per week where the ESL students are obliged to be in the ESL elective class. There are no electives in Grade 6-8, which means that there are no stand-alone ESL lessons for them (apart from the two ‘split’ classes that form part of English). Further ESL support comes from an ESL teacher joining one or two of the students’ Science or History classes per week, where the ESL teacher will be able to help ESL students one-to-one or in small groups during the lesson, while the content teacher will usually take this time to respond to the needs of the non-ESL students. Sometimes it may be deemed necessary to rather take one or more ESL students out of the content lesson so as to support them with any immediate English language needs. This often happens with new students or students with very little English knowledge. The problem with such an approach is that the next time that student has a Science or History lesson the ESL teacher may not be available for help, due to timetabling conflicts, in which case the student goes back into the mainstream class without support, and perhaps even more lost because he or she is one lesson behind the rest of the class.

With regards to problems related to vocabulary, I have already mentioned that difficulties for ESL students not only include having to comprehend the individual words spoken by their classroom teachers, but also having to make sense of those teachers' use of language patterns and formulaic language, such as idioms and fixed expressions. Moreover, it is difficult for students to make meaning of this lexis in the fluid, transitory context of a classroom where there is little time for them to process information before the teacher moves on to a new point.

In terms of the vocabulary used by teachers, it is often a mixture of general service vocabulary (words used in daily conversation), general academic vocabulary (words found across a variety of formal, academic disciplines), and technical vocabulary (core words specific to the subject being taught). The distinction between general service vocabulary and general academic vocabulary is often defined in terms of basic interpersonal communicative skills (BICS) and cognitive/academic language proficiency (CALP) (Cummins 1979 in Cummins 2008). BICS requires proficiency in general service vocabulary – words that are typically used in spoken language and that tend to be the 2,000 most common words used in English. In contrast, CALP (which is typical of written language) requires competence in general academic vocabulary, in addition to high frequency words up to the 2,000-3,000 word level, and words that occur in both spoken and written discourse, and thus has a wider vocabulary base. Since teacher discourse probably contains both the general service vocabulary common to spontaneously spoken discourse as well as academic words common to the formal learning environment of a classroom, it is necessary for students to be proficient in both BICS and CALP. Unlike a textbook, which may contain mainly language required for CALP, teachers may use features of both oral and written language, including a lot of general service vocabulary in order to get their points across; therefore, students also need BICS. It is difficult for an ESL student to prepare for a teacher's discourse because rarely do language learning materials accommodate both BICS and CALP; instead they tend to display features of one or the other. This complex nature of teachers' discourse, along with other factors such as unstructured script, unfamiliar pronunciation and lack of time allowed to process speech, may make listening comprehension in the classroom very challenging for ESL students.

1.3 Research problem

The main problems that this study will investigate are the vocabulary size, in terms of frequency levels, of the Grade 9 and 10 high school students, particularly those of the ESL students, at the research school, and the profile of their teachers' spoken classroom discourse.

It is important to identify what the students' vocabulary levels are as well as the levels at which their teachers are speaking, because then one can determine what the vocabulary gap is and assess to what extent the students can comprehend the discourse of the teachers.

Waring and Nation (1997, 10-11) state that it is vital for English language learners to prioritise learning the most frequently used 3,000 word families (high frequency words) since knowledge of 2,000-3,000 word families is adequate for 'productive use in speaking and writing'. However, knowledge of only 3,000 word families is not enough for students to comprehend texts that they might encounter in high school, such as novels, textbooks and historical articles they might have to read, and of course the discourse spoken by their teachers in class. Nation (2006) found that knowledge of 98% of the words in a variety of written and spoken texts is required to comprehend those texts. Nation (2006, 59) found that 8,000-9,000 word families make up 98% of a written text, and 6,000-7,000 words constitute 98% of a spoken text.

This study aims to identify the vocabulary profile used by teachers, and how much of their vocabulary covers 98% of their discourse, after which I will be able to identify how much vocabulary students need in order to comprehend their teachers. As part of this research problem, I will assess the vocabulary levels of the students to see the gap, if any, between student vocabulary levels and the vocabulary levels used by the teachers as they teach their subjects.

The second research problem that this study addresses is what seems to be a gap in the current research literature. Nation's (2006) finding that 6,000-7,000 word families are required for comprehension of spoken discourse is based on the generalised Wellington Spoken corpus, which included speech acts such as talk-back radio and interviews (Nation 2006, 77). In my investigation of the current literature, I was unable to find any answers as to what vocabulary frequency level is required for high school students to comprehend the vocabulary spoken by Grade 9 and 10 high school teachers. This gap is acknowledged by Schmitt (2010, 38) when he states, 'there is a big gap in the field's understanding of spoken discourse and vocabulary'. Academic corpora (databases of texts that are found in an academic context, e.g. essays written by university students) exist, and as a result research has been conducted into the nature of academic vocabulary. However, many of these corpora have been compiled from **tertiary** institutions and/or **written** texts, and thus may not reflect the nature of the vocabulary **spoken** by **high school** teachers. Therefore, the lack of spoken corpora from the

specialised context of a high school academic environment, that can be used to inform the vocabulary profile of high school teacher discourse, is one of the motivations for creating the small specialised corpus of the spoken English of high school teachers compiled for this study.

1.4 Research aims

This study has two aims. The first is **to assess the vocabulary levels of Grade 9 and 10 students at an international school**, where English is the language of teaching and learning. This was done through testing the students using Nation and Laufer's (1999) Productive Vocabulary Levels Test (PVLTL). The second aim is **to explore the vocabulary profile of the spoken discourse of Grade 9 and 10 high school teachers** at the research school. In order to do understand the teachers' vocabulary profile, I recorded at least one Grade 9 or 10 teacher from the subjects of English, History, Mathematics and Science while they were teaching. I made the recordings over a period of a few months. Most of the recordings took place in the second half of the students' academic year, mainly from February to April 2014, but a few were done in May and June of the same year and then again in September after the (northern hemisphere) summer break and at the start of the new academic year. During that same time-frame I transcribed the recordings and used corpus linguistics software programs to analyse the data.

1.5 Research questions

Two main research questions drive this study, with the second having a number of sub-questions:

- 1) What are the productive vocabulary levels of the Grade 9 and 10 high school students at an international high school?
- 2) What is the vocabulary profile of the spoken discourse of Grade 9 and 10 high school teachers at the research school?
 - a) How much of the high school teachers' spoken discourse is made up of general and academic English?
 - b) What is the nature of the academic vocabulary spoken by the high school teachers?
 - c) How much of the high school spoken corpus is made up of high, mid and low-frequency vocabulary?

- d) How do the high school spoken sub-corpora compare with one another in terms of their coverage of general and academic English?
- e) How does the nature of the words spoken by the high school teachers compare with the vocabulary profile found in other corpora?

1.6 Overview of the study

In order to assess the vocabulary levels of the students at FIS, I tested the students using Laufer and Nation's (1999) Productive Vocabulary Levels Test (PVLТ). All available Grade 9 and 10 students in the ESL programme were administered the test, and many of the non-ESL students were also tested. A total of 23 students participated in this component of the study. The results of these tests will be presented in section 4.1 of Chapter 4.

In order to explore the nature of the vocabulary spoken by the FIS teachers I created a corpus of their speech by firstly recording them while they were teaching, and then transcribing the recordings. I then used corpus software tools to analyse their discourse. Because of the aforementioned procedures (see Chapter 3 for more detail), much of my literature review (Chapter 2) discusses existing research conducted in the fields of vocabulary and corpus linguistics. In Chapter 2, I begin with a discussion on vocabulary, and then move to defining what a corpus is and discussing different types of corpora. For my study I created a spoken corpus (referred to as the 'FIS corpus'), which is small and specialised; in my literature review, I discuss existing corpora in terms of being generalised, specialised, written and/or spoken.

My literature review reveals the following gap in knowledge, on which my study attempts to shed light: Most existing corpora tend to be written, and those corpora which are spoken and academic in nature tend to come from a university context. There seems to be very little existing research, or available corpora, which have the nature of high school teachers' spoken discourse as the focus. Since my corpus involves the relatively unstudied area of the spoken discourse of high school teachers, I hope that my study will add new value to the field of corpus linguistics.

Since my experience with the ESL students at FIS suggested that one of their problems with understanding their teachers was their low vocabulary levels, I devoted a substantial part of my literature review to the field of vocabulary. Chapter 2 identifies two corpus software tools – *VocabProfile* and *WordSmith Tools* – which are commonly used to analyse the vocabulary

contained in corpora. Both these software tools are what I used to analyse my corpus. *WordSmith Tools* is used to create frequency word lists of all the words in a corpus in order to identify which words are used more regularly than others. My literature review discusses some of these word lists, such as the General Service List (GSL) and the Academic Word List (AWL). For my study I used *WordSmith Tools* to create a list in order of frequency of all the words in the FIS corpus, which I used to compare with other word lists. In so doing I was able to compare the vocabulary profile of the high school teachers with the vocabulary profile of other English contexts. These findings are presented in Chapter 4.

VocabProfile is used to analyse corpora in terms of bands of frequency levels. By running a corpus through *VocabProfile*, one is able to see what percentage of a corpus uses vocabulary at the most common 1,000, 2,000, 3,000 (and so on) frequency bands or one can see what percentage of the corpus is made up of academic vocabulary. *VocabProfile* uses a variety of corpora, namely the combined British National Corpus and Corpus of American English (BNC/COCA) list, GSL, New General Service List (NGSL), the AWL, and the New Academic Word List (NAWL) as points of reference for the analysis of whatever text one wishes to analyse. All of the above-mentioned frequency word lists are discussed in Chapter 2, and the way *VocabProfile* analysed the FIS corpus will become clearer in Chapter 4.

Much of the discussion on vocabulary frequency in Chapter 2 centres on research that was helped along with programs like *VocabProfile*. I mentioned in section 1.3 that knowledge of 98% of the words in a text is required in order to comprehend a text unassisted (Nation 2006, 1). *VocabProfile* is able to assist in identifying how many words of a text make up that 98%. For example, by using *VocabProfile*, Laufer (2010) was able to discover that knowledge of roughly 8000 words would give one 98% coverage of a reading text used for a university entrance test. For this study I ran the FIS corpus through *VocabProfile* in order to reveal how many words make up high school teachers' discourse, in so doing it became possible to evaluate whether or not students require more or less vocabulary in order to comprehend high school teachers' speech in comparison with that required to comprehend a university level reading text.

1.7 Significance of the study

The gap in knowledge addressed by this study is identifying the nature of teacher classroom discourse by establishing a corpus for analysing the nature of the vocabulary spoken by high school teachers. This study is significant because by creating a corpus of teacher classroom

discourse, measuring learners' vocabulary levels and determining what the gap between the two is (if any), it has the potential to show how important it is to increase the amount of vocabulary instruction being conducted in schools, especially schools with a large number of second language speakers. Indeed, as Lesaux, Kieffer, Faller and Kelley (2010, 198) point out, only 'superficial attention' tends to be given to language development in the middle and high school grades. If a large gap is shown between the vocabulary levels of teachers and those of students, then teachers at international schools may see the necessity to redress the problem indicated by Lesaux et al. (2010). The study will identify the frequency vocabulary levels of high school teachers' spoken discourse and it will identify what percentage of teachers' vocabulary is made up of general and academic English. This is significant in that it adds to the literature about spoken corpora and teacher-spoken discourse. Having measurable data of teachers' discourse will enable teachers, particularly ESL teachers, to focus on the specific vocabulary levels (especially those that make up 98%) of the teachers' discourse when attempting to improve the vocabulary knowledge of their students.

1.8 Paramaters of the study

This study explores the nature of the spoken vocabulary of high school teachers at a single school and, except for the purposes of comparison, the aim of this study is not to explain vocabulary used in its written form as found in the high school or any other context. This study also does not aim to describe the nature of spoken vocabulary in any context outside of the high school classroom. Another delimitation of this study is that it analyses vocabulary as individual word items, and not how words are associated with others, as in collocations. Finally, this study does not attempt to analyse the meanings of the words as they are used in the context in which they are spoken.

1.9 Definition of terms

Throughout the study various terms are defined. Some of the key terms are first defined below but will be elaborated on in later sections of this paper.

Content words: Words that carry meaning, e.g. nouns, main verbs and adjectives. More explanation and examples will be provided in section 2.2.2.3.

Controlled productive vocabulary: Words used when compelled to do so by a teacher or researcher.

Corpus/corpora: A database of naturally occurring language. Corpora will be discussed in more detail in section 2.3.1.

Formulaic language: Two or more words that go together to form a multi-word unit, e.g. collocations, idioms and fixed expressions. More explanation and examples will be provided in section 2.2.2.2.

Function words: (also referred to as grammatical or form words). These words have a grammatical function, e.g. articles and prepositions. More explanation and examples will be provided in section 2.2.2.3.

Headword: The root word or dictionary (base) form of a word. More explanation and examples will be provided in section 2.2.1.

Lemma: The base form of a word and its inflected forms. More explanation and examples will be provided in section 2.2.1.

Word family: The base form of a word, its inflected forms and its derived forms. More explanation and examples will be provided in section 2.2.1.

Word frequency: Refers to how often (or frequent) a word is used in a text. More explanation will be provided in section 2.2.2.4.

1.10 Summary of the remainder of this dissertation

Chapter 2 begins by providing a review of relevant research conducted in the field of vocabulary, which is followed by a review of the field of corpus linguistics.

Chapter 3 provides detail about the research method, procedures used to conduct the research for this study and analyse the data, and it describes the pilot study undertaken before the main study.

Chapter 4 is the findings chapter, where the results of the main study are presented and discussed.

Chapter 5 is the concluding chapter where I summarise the major findings of the study, describe the contributions of the study and acknowledge some of its limitations, and point to avenues for further research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a discussion on issues related to vocabulary studies. It is divided into two sections: a review of vocabulary is followed by one on corpora. In addition to providing an overview of what corpora are, this chapter includes a discussion on the value of corpora in language teaching. Similarly, the discussion on vocabulary is contextualised in terms of its importance in the high school learning environment. The two main sections (vocabulary and corpora) of this chapter were used because of their central role in this study. As part of my study, I measured the vocabulary levels of high school students, which is why I paid attention to vocabulary research in this chapter. Much of my study involved the creation of a corpus of the spoken discourse of high school teachers, hence the reason why much of this chapter discusses research that has been conducted in the field of corpus linguistics. My analysis of the FIS corpus concerned the frequency levels of the high school teachers at FIS, which is why I devoted attention to discussing research into vocabulary levels in this chapter.

2.2 Vocabulary

The importance of learning vocabulary for ESL learners cannot be overstated; indeed it could be the most important part of mastering a language. Pikulski and Templeton (2004, 1) acknowledge the central role of vocabulary knowledge in society when they state, ‘our ability to function in today’s complex social and economic worlds is mightily affected by our language skills and word knowledge’. Schools reflect the important role of vocabulary knowledge in society by placing high vocabulary demands on students. Indeed, Anderson and Nagy (1993, 2) state that ‘squaring with teachers’ experience, one of the most consistent findings of educational research is that having a small vocabulary portends poor school performance and, conversely, that having a large vocabulary is associated with school success’.

Central to this study are the areas of word frequency, the nature of the vocabulary (in terms of what combination of general, academic and technical words) spoken by high school teachers, and the vocabulary needs of ESL students. Although the research done in this study focuses on individual words, an understanding of a language, in terms of vocabulary, also involves

knowledge of groups of words that go together. For convenience multi-word lexis is described as formulaic sequences in this chapter and includes such lexis as collocations and idiomatic expressions.

2.2.1 Some key definitions

Throughout the presentation of this study, a number of key words are repeated, namely *type*, *token*, *word family* and *lemma*.

‘Type’ and ‘token’

When one runs a corpus through software, such as *VocabProfile*, the presentation of the results will be a count of the words in terms of how many *types* and *tokens* there are in that corpus. The total number of tokens refers to the total number of words in the corpus. A word here means any group of ‘letters separated by spaces or punctuation’ (Hunston 2002, 17). If the same word is repeated ten times, these ten occurrences are counted as ten tokens. The total number of types, however, refers to the total number of different words. If a word is repeated ten times it is only counted as one type. To illustrate the meaning of these words I have taken a screen shot of a portion of the analysis produced by *VocabProfile* of the corpus created for this study. Table 2.1 presents all the vocabulary *types* in the FIS corpus, compiled for this study, that are in the range of the 6,000-7,000 most frequent English words used according to the BNC/COCA corpus.

Table 2.1 Vocabulary types in the FIS corpus at the BNC/COCA 7,000 word level.

<p>BNC-COCA-7,000 types: [fams 24 : types 29 : tokens 88] extract</p> <p>allegory_[4] arithmetic_[4] barbarian_[1] barbarians_[3] caste_[1] combustion_[3] conscription_[1] detractor_[1] extinguisher_[1] feudal_[6] feudalism_[2] gradient_[9] innate_[3] menopause_[4] menstrual_[2] menstruation_[2] neon_[4] pees_[1] potassium_[1] ransom_[11] sandals_[1] snoring_[2] staunch_[2] subtract_[5] subtracted_[1] subtracting_[5] subtraction_[1] tar_[4] treason_[2] violet_[1]</p>
--

As one can see *barbarian* and *barbarians* are considered different types. The number in parenthesis alongside each *type* refers to the *tokens* (or the number of times each *type* was used). Thus the type, *barbarian* has one token, while *barbarians* has three tokens.

‘Word family’ and ‘lemma’

The term *lemma* ‘is used to mean the base form of a word, disregarding grammatical changes such as tense and plurality’ (Biber, Conrad and Reppen 1998, 29). A lemma includes a word

(such as *walk* – this would be the headword or base or root form of the word) as well as its inflected forms (such as *walks*, *walked* and *walking*). Thus, one lemma includes a number of different forms. A *word family* is not dissimilar to a lemma in that it also includes different word forms. The difference is that a word family includes more word forms than lemmas. In addition to the headword (base or root form), and its inflected forms, word families include the word's derived forms. Taking the same headword *walk*, these derived forms refer to the use of that word in its noun forms, namely *walk* (as in 'go for a walk') and *walker* (the person who walks).

Discussions about word frequency and word lists are often presented differently. Sometimes all forms of a word are counted separately (these would be indicated as *types*), while at other times words are presented as either lemmas or the headwords of word families. Table 2.2 shows all the *word families* in the FIS corpus, compiled for this study, that are in the range of the 6,000-7,000 most frequent English words used according to the BNC-COCA corpus.

Table 2.2 Word families in the FIS corpus at the BNC/COCA 7,000 word level.

<p>BNC-COCA-7,000 Families: [fams 24 : types 29 : tokens 88]</p> <p>VP-negative: bnc_coca-7</p> <p>allegory_[4] arithmetic_[4] barbarian_[4] caste_[1] combustion_[3] conscript_[1] detract_[1] extinguish_[1] feudal_[8] gradient_[9] innate_[3] menopause_[4] menstruate_[4] neon_[4] pee_[1] potassium_[1] ransom_[11] sandal_[1] snore_[2] staunch_[2] subtract_[12] tar_[4] treason_[2] violet_[1]</p>
--

While Table 2.1 presented *barbarian* and *barbarians* as different types, Table 2.2 shows that they are part of the same word family, that fall under the headword *barbarian*. The number in parenthesis alongside the word refers to the total number of times words within that word family were used. Thus, taken together, *barbarian* and *barbarians* were spoken a total of four times.

2.2.2 Types of vocabulary

In discussing vocabulary it is helpful to distinguish between different types of vocabulary. Schmitt (2010, 75) presents the following common classifications of vocabulary: *Word class*, *Formulaic sequences*, *Content and Function words*, *Frequency*, *Written and Spoken vocabulary* and *General, Academic and Technical vocabulary*. Another vocabulary distinction that will be discussed in this chapter is that of *Receptive* and *Productive vocabulary*. These categories are not distinct from each other, instead there is considerable

overlap. Thus we find, for example, that *general vocabulary* and *function words* tend to be *high frequency vocabulary*. In the following discussion I will attempt to limit repetition as far as possible, despite the above-mentioned crossover.

2.2.2.1 Word class

Word class involves viewing words grammatically, in terms of what part of speech they are, in other words whether they are nouns, verbs, adjectives, etc.

2.2.2.2 Formulaic sequences

Although this study focuses on words as individual items, a large percentage of the English language involves formulaic sequences, or groups of words that go together. Formulaic sequences include idiomatic or fixed expressions, such as *kick the bucket* and collocations, such as *free time*. Knowledge of individual words does not necessarily mean that someone knows a text or speaker's meaning when some of those words are put together. In the expression *kick the bucket*, an understanding of the three words as individual items *kick*, *the* and *bucket* will not in any way help someone understand that the expression *kick the bucket* means "die" (Conrad & Biber 2005, 57). Formulaic sequences, especially idiomatic expressions, very often need to be learned as single lexical units with their meanings, since these are not necessarily clear from the individual words that constitute them. Furthermore, it would be helpful for language learners to learn these formulaic expressions as single lexical units because it will 'reduce cognitive effort (and) save processing time' (Shin and Nation 2008, 340). Thus it will be much quicker and easier for ESL speakers to produce one set phrase resulting in them being more communicatively fluent.

Shin and Nation (2008) identified the most frequent collocations in spoken discourse by analysing the spoken section of the BNC. Their study confirms that there is a high use of collocations in spoken English. They found that as many as 5,894 collocations exist if one uses just the most frequent 1,000 content words as pivots (Shin and Nation (2008, 343). The most frequently spoken collocation was *you know* and the second was *I think (that)*. Both of these collocations were spoken at least three times more than the third most frequent collocation which was *a bit*. Due to the high frequency of these collocations in spoken discourse it would be useful to include them when teaching ESL learners, especially if the goal of the learning is proficiency in speaking for general purposes. Having said that, Shin and Nation (2008, 345) warn against using frequency as the only criterion for selecting

vocabulary for teaching. Collocations, like *Good morning* and *How are you?*, which are clearly very important items to learn, do not feature in the top 100 of Shin and Nation's frequency list of collocations. Nevertheless, Shin and Nation (2008, 345) posit that 'having a list of the most frequent collocations in spoken English to choose from is a useful starting point for syllabus design'.

2.2.2.3 Content and function words

Content words are words that carry meaning; they, therefore, tend to be nouns (e.g. *boy*), main verbs (e.g. *run*), adjectives (e.g. *happy*) and adverbs (e.g. *quickly*). On the other hand, function words (also referred to as form or grammatical words) do not carry much information about a topic; instead they are important for providing a grammatical function. Examples of function words are articles (e.g. *the*), pronouns (e.g. *he*), conjunctions (e.g. *so*) and prepositions (e.g. *in*). Frequency word lists consistently reveal function words as being the most frequently used words in English (Schmitt 2010, 54). Indeed the five most common words in English according to the New General Service List (NGSL) as shown in Table 2.5 in section 2.3.3.2 are all function words. Function words are more commonly used than content words because they are needed when writing or talking about any topic, whether that topic is for every-day communicative purposes or a highly technical one in an academic context. Schmitt (2010, 54) claims that one has to look beyond the most common 100 word forms in the BNC before one regularly encounters content words. The more genre-specific or specialised a corpus, the more content words are seen to have higher frequencies than generalised corpora (Lee 2001, 252). Kennedy (1998, 102) found that content words made up 36% of the most frequent 50 words in an economics corpus, while they made up only 6% of academic English as a whole and an even lower 2% of all of the English language.

2.2.2.4 Frequency

Waring and Nation (1997, 8) define word frequency as 'how often the word occurs in normal use of a language'. They also acknowledge that looking at frequency is a useful guide to determining how much vocabulary a person needs in order to do the things that language users are required to do. Due to the field of corpus linguistics and the software that allows for the creation of frequency word lists, it is possible to identify which words occur very frequently, and hence become a learning priority. Although word frequency is not the only factor in deciding what vocabulary to teach, it is clearly important (Aston 1997; Leech 1997 in Römer, 2008, 115). Schmitt (2010, 63) stresses the importance of prioritising high

frequency vocabulary when he states that word frequency is ‘the single most important characteristic of lexis that researchers must address’. Laufer and Nation’s (1999, 35) statement that ‘words should be learned roughly in order of their frequency of occurrence, with high frequency words being the first’ is supportive of the notion that learning high frequency words before lower frequency words should be an approach to learning a language. The difference between high and lower frequency words will be examined more closely in the sections that follow.

2.2.2.5 How much vocabulary does a native English speaker have?

In order to contextualise a discussion on frequency it is helpful to know beforehand how many words native English speakers typically have. However, Waring and Nation (1997, 7) note that measuring vocabulary size is problematic, as it is difficult to define what a word is and also it is difficult to test whether a word is known or not. As a result of these challenges, a number of studies in the area of vocabulary size ‘give very diverse and misleading results’ (Waring & Nation 1997, 7). Nevertheless Waring and Nation (1997, 7-8) maintain that a conservative estimate of the number of word families that a native English speaker, who has graduated from university, knows is 20,000 word families excluding proper nouns, foreign words, abbreviations, and compound words. Goulden, Nation and Read (1990 in Schmitt 2010, 6) found that university undergraduate native speakers of English had a vocabulary size of 17,000 word families. Thus it is fair to say that adult, university educated native English speakers have a vocabulary size of 16,000-20,000 word families, (a word family being the headword or base form of a word as well as its inflected and derived forms). Since there are a number of different word forms that fall under the same word family, it is probable that if one has a vocabulary size of 16,000 word families, then one might actually know more than four times that number (i.e. 64,000) in terms of separate words (or types). For somebody embarking on a study of English, whether it be for communicative or academic purposes, it would be unhelpful to suggest to them to learn all of these words, and certainly not in a random order. Adolescents – first language (L1) or ESL students – do not have to have a vocabulary of 16,000 word families to function in a high school setting. Zechmeister, Chronis, Cull, D’Anna, and Healy (1995, in Schmitt 2010, 6) found that L1 junior high school students (Grades 7 - 9) had an average vocabulary of 11,836 headwords. If native English speakers add roughly 1,000 words per year (Waring & Nation 1997, 7) then it is fair to say that L1 students working through Grade 10 to the first years of university will increase their vocabulary from approximately 12,000 to 15,000 word families. Since these are the

vocabulary levels of educated L1 speakers, it is probably unrealistic to expect ESL learners to match them (certainly if they are not immersed in an environment of using English all the time).

2.2.2.6 What constitutes high frequency vocabulary?

The first vocabulary target for ESL learners should be knowledge of high frequency words. High frequency vocabulary has traditionally been used to refer to the most common 2,000 word families. This is largely due to the General Service List (GSL) which is a list of high frequency words published by West in 1953. It contains 2,000 of the ‘most widely useful word families in English’ (Coxhead 2000, 213) and is based on a 5,000,000 word written corpus. Various studies show that these words make up between 80 and 90% of written texts. Browne, Culligan and Phillips (2013) found that the GSL covered 84% of the Cambridge English Corpus (which is a 2 billion word corpus made up of 90% written texts and 10% spoken texts). When it comes to purely spoken texts, 2,000 words have more coverage. Schonell, Meddleton and Shaw (1956 in Waring and Nation 1997, 9) discovered that as much as 96% of an informal spoken text is made up of 2,000 words, but in more recent research Nation (2006, 77) found that words at the 2,000 level account for a lower coverage of 89% of English conversation. Spoken language, therefore, requires a relatively small number of high frequency words when compared with written English. Written English tends to comprise richer and wider ranging words, words associated with CALP.

On the surface these seem like high percentages, but research suggests that much higher vocabulary levels are required for English comprehension, thereby making the gap between knowledge of 2,000 word families and text comprehension too large. Hu and Nation (2000, 403) discovered that none of the participants in their study could gain sufficient comprehension of a reading text when only 80% of the words were known and only some of the participants gained comprehension when 90-95% of the words were known. Hu and Nation (2000, 403) concluded that 98% coverage of the vocabulary in a fiction text was needed for learners to be able to gain comprehension. Because knowledge of 2,000 words covers fewer than 90% of written texts, Schmitt and Schmitt (2014) argue that 3,000 word families (not 2,000) should be considered high frequency words. Waring and Nation (1997, 11) are adamant about the importance of ESL learners knowing 3,000 high frequency words when they state that ‘these are an immediate priority and there is little sense in focusing on other vocabulary until these are well learned’. Lending weight to the importance of knowing

3,000 words comes from Schmitt (2010, 7) when he states, ‘current evidence suggests that it requires knowledge of between 2,000-3,000 word families to be conversant in English. This means that, although knowledge of 6,000 word families is ideal for comprehension of a spoken text, i.e. listening (refer to section 1.3), a much smaller vocabulary is required for making one understood when speaking. Adding support to defining high frequency words as those being the most common 3,000 English words is Laufer’s (2010) study into the relationship between ESL learners’ vocabulary sizes and lexical text coverage. She found that 3,000 words covered almost 91% of the British National Corpus (Laufer 2010, 22). If adding knowledge of proper nouns (mainly personal and geographic names) is added to this percentage then approximately 93% coverage of a fiction text can be achieved with knowledge of the 3,000 most frequent words. These figures fall between the 90-95% coverage that Hu and Nation (2000, 403) found was enough for some (although, not all) readers to have in order to comprehend a text.

Since most learners do not gain sufficient comprehension of texts with a vocabulary of fewer than 3,000 words, it may be necessary for these high frequency words to be explicitly taught. Coady (1993, 16) is clear about this when he states, ‘the ability to automatically recognise the highly frequent words in a language is absolutely crucial to success in L1 and L2 reading, and therefore these words should be taught’. Coady (1993, 16) goes on to state that ‘the less frequent words are to be learned through incidental contact in context (with the help of some strategic training) via extensive reading, but only after a critical level of automaticity has been achieved with the high-frequency or core vocabulary’. Explicit teaching and incidental learning of words will be discussed in more detail in section 2.2.5.

2.2.2.7 Mid-low frequency words

The ‘less frequent words’ to which Coady (1993) refers have been divided into two categories by Schmitt (2010, 70): mid-frequency vocabulary and low-frequency vocabulary. Mid-frequency vocabulary refers to words between the 3,000 and 8,000-9,000 range and low-frequency vocabulary refers to words beyond the 9,000 word threshold (Schmitt and Schmitt, 2014). The 9000 word threshold would be the safe highest estimate for the mid-frequency band. One reason for the distinction is that low-frequency words are traditionally not needed to be explicitly taught (as suggested by the earlier quote from Coady (1993)) because learners can often work out the meaning through context or be equipped with a learning strategy to deal with the words, as explicit teaching of those low-frequency words would be too time-

consuming. However, since research shows that learners need higher vocabulary levels than originally thought (from 6,000-9000 word families) to gain comprehension, it is unrealistic for learners to access the vocabulary between the 2,000 and 6,000 word range without help (Schmitt 2010, 70). Mid-frequency vocabulary – words between the 3,000 and 9,000 range – are important too for comprehension, and because students have been found to require up to 9,000 words before comprehension is generally gained, these words may need to be explicitly taught and hence, for pedagogic purposes, require attention. This view is supported by Coady (1997, 288) when he states:

If someone needs to achieve only small to moderate proficiency in the language without any time pressure, then the natural approach, whereby acquisition occurs through contextual use alone, would seem a reasonable suggestion.

If, however, students need to use a language for challenging academic purposes...then we can see a different type of approach being suggested in the literature. The proponents of the strategy approach are quite convincing in their claim that academically oriented students need help in order to improve their acquisition skills.

One frequency word list that includes high, mid and low frequency words comes from the British National Corpus and The Corpus of Contemporary American English (BNC/COCA) combined list of 25,000 word families. In order to develop an international frequency list and account for both North American and British varieties of English, Mark Davies (the researcher behind the COCA corpus) and Paul Nation combined the BNC and COCA lists into an integrated one and expanded it to the 25,000 word family level (Cobb, 2014). Since such a large number of words are included in the list, high, mid and low-frequency words are all included. All of the words are presented in bands of 1,000 words, so in actuality there are 25 small lists of 1,000 words each in the order of frequency. One advantage of the BNC/COCA list is that the risk of a word not being included is small. This is a limitation of other established lists such as the GSL and the academic word list (AWL). Because those lists are relatively small, and are of either a general or academic nature, it is possible that words students may need may not be included. Also when using software like *VocabProfile* to analyse a text in terms of how many GSL and AWL words are in that text, there is the possibility of the presence of many ‘off list’ words. ‘Off list’ words are words which do not feature in whichever list(s) that one has chosen to analyse one’s corpus. There is no way of knowing then whether those words are mid (hence useful to students) or low (and perhaps not useful) frequency words. Cobb (2014) notes that one reason for the BNC/COCA list having 25,000 word families is to reduce the number of these ‘off list’ words. Since it would be rare

to find a word that is of such low frequency that it would not feature in the most frequently used 25,000 word families, the words in the ‘off list’ category would be limited to proper nouns.

2.2.2.8 How much vocabulary does an ESL high school student need?

The research findings in this section point towards ESL learners who are studying English for academic purposes and who are using ESL as their language of learning needing 6,000-9,000 words. The bare minimum would be 3,000 words and that would be in order to function reasonably in a general English environment, not an academic one. Nation (2006, 77) found that 3,000 word families covered 95% of unscripted spoken English, which as stated in section 2.2.2.6, was enough of a percentage for some (not most) participants in his study to gain comprehension of a text, albeit a written text. Also 3,000 words might be adequate for speaking, but not necessarily for listening and reading. Laufer (1998, 1) states that Israeli ESL high school students are expected to have 3,500 – 4,000 word families (approximately 20,000 separate words or types) by the time they graduate. Another useful finding is that of Staehr (2009, 7) which is that advanced Danish listeners who knew 5,000 word families were able to comprehend a listening text on the Cambridge-ESOL Certificate in English listening exam. However, Nation (2006, 79) claims that ‘a 8,000-9,000 word-family vocabulary is needed for dealing with written text, and 6,000-7,000 families for dealing with spoken text’, that is if 98% is the ideal coverage. This means that ESL students should strive for a vocabulary of at least 6,000 word families.

2.2.2.9 General, academic and technical vocabulary

The distinction between general, academic and technical vocabulary is important in this study; however, it is a difficult distinction to make. Indeed, what constitutes ‘academic’ vocabulary may include (in addition to academic words) general words that are acceptable in an academic context as well as technical words which are specific to a particular subject area. Nevertheless when general service vocabulary is referred to in this study, it is those words found in the general service list (GSL), new general service list (NGSL) and the first 3,000 word families in the BNC/COCA-25 list, while academic vocabulary refer to those words found in the academic word list (AWL) and the new academic word list (NAWL).

2.2.2.9.1 General vocabulary

The General Service List (GSL)

One of the most well-known frequency lists is the GSL, which was compiled by West in 1953. From a corpus of 5 million words West managed to develop a list of the 2,000 most widely written words in English. The list was compiled in order to address the needs of ESL learners. Over the years it has been used as a basis for creating graded readers and other language learning materials which have been used to improve the English ability of ESL students around the world. Waring and Nation (1997, 14) note the usefulness of the list because ‘it contains words within the (word) family each with its own frequency’. To illustrate this point they provide the examples of *excited*, *excites*, *exciting* and *excitement* which fall under the headword *excite*. According to Coxhead (2000, 213), the criteria used by West for inclusion in the list include ‘frequency, ease of learning, coverage of useful concepts, and stylistic level’. One of the reasons for the GSL’s use in the development of teaching material is its high word frequency in written texts. Waring and Nation (1997, 13) acknowledge some limitations of the GSL, namely ‘its age, some errors and its solely written base’, but they argued at the time that ‘it still remains the best of the available lists because of its information about the frequency of meanings, and West’s careful application of criteria other than frequency and range’ (Waring and Nation 1997, 13). The GSL has aged even more since Waring and Nation’s (1997) comments, which was a motivating factor in the New General Service List (NGSL) being compiled in 2013. Other criticisms of the GSL (and hence the need for an updated version) are that it was developed by a corpus that was too small and that it did not clearly define what a ‘word’ was (Browne et al., 2013).

The New General Service List (NGSL)

The NGSL, which was published in 2013, is a list of 2800 high frequency words ‘carefully selected from a 273 million-word subsection of the 2 billion word Cambridge English Corpus (CEC)’ (Browne et al., 2013). In developing the list from such a large corpus Browne and his colleagues addressed one of the criticisms of the GSL, namely that it was drawn from a corpus that was too small; also the corpus used by Browne et al. (2013) was an updated one. There is an interesting difference between the GSL and NGSL in terms of the word families and lemmas the lists contain. The NGSL consists of 2,368 word families which is more than the 1,964 word families contained in the GSL, whereas the 2,818 lemmas of the NGSL is considerably less than the 3,623 lemmas of the GSL. The NGSL is actually presented in terms

of lemmas, as opposed to word families (which is the case with the GSL). One advantage of compiling a lemma based list is that the ranking and organisation of frequency bands more precisely reflects the words actual occurrence in a text (Brezina & Gablasova 2015, 1). The GSL covers 84% of the CEC corpus, while the NGSL covers 90% of the corpus. It can therefore be argued that because of its higher coverage, ESL students will have greater vocabulary coverage if they learn words from the NGSL as opposed to the GSL. It should be noted here that this understanding of English is for the general use of English - the English found in fiction and everyday conversation. An ESL student who requires English for academic purposes would need an additional type of vocabulary, which is why more specialised academic word lists like the Academic Word List (AWL) have been created.

2.2.2.9.2 Academic vocabulary

Word lists like the GSL may cover a high percentage of the vocabulary used in daily conversation and fiction stories; however, in the academic environment of the upper grades of high school or university, lower frequency words of a more academic nature are prevalent. Academic vocabulary is ‘vocabulary used across all academic disciplines but is not the technical vocabulary of a particular academic discipline’ (Zhou 2010, 14). Zhou (2010) goes on to state that academic words tend to have more Latin and Greek roots than everyday spoken words and they cover approximately 10% of the words in an academic text. If students are going to find success in an academic environment then these words need to be learned.

Corson (1997, 671) notes the difference between general service vocabulary (words associated with BICS) and academic vocabulary (CALP words). He argues that in order to achieve academic success in English one must be proficient in Graeco-Latin academic vocabulary (as opposed to words from Anglo-Saxon origin, which largely make up every day conversational English). In a comparison of high frequency general service words from the Collins Birmingham University International Language Database (COBUILD) and the academic vocabulary in the University Word List (UWL), Corson (1997) notes that only two of the 150 most frequent words in COBUILD are of Graeco-Latin origin, those being *very* and *because*. Examples of commonly used words not from a Graeco-Latin origin are *some*, *and*, *also* and *do* (Corson 1997, 677). Corson (1997) further illustrates the difference between the two lists by pointing out that all but five words in the UWL are of Graeco-Latin origin. Examples of commonly used academic words from Graeco-Latin origin in the UWL are

analyse, contrast, precede and *element*. Corson's (1997) study highlights the different nature of general service and general academic vocabulary. Having a firm grasp of general service vocabulary only will not ensure an individual's success in more formal, academic environments. If we learn what we are exposed to, then it will be very difficult for students to do well academically if they are only exposed to the vocabulary used in daily conversation. Students, therefore, need access to the academic discourse and texts of formal education in order to become proficient in the essentially 'Latinated' academic vocabulary used in such environments. These findings also illustrate how important reading is for vocabulary development.

The Academic Word List (AWL)

To help students access vocabulary of an academic nature, Coxhead (2000) created the AWL. This high frequency academic word list, was compiled so that university students studying English for academic purposes (EAP) would have a good idea of the academic vocabulary that will be of most use to them. The list contains 570 word families, all of which come from a written academic corpus comprising the four domains of arts, commerce, law and science. Coxhead (2000) made use of three criteria on which selection of the words was based. Firstly, the words could not be included in the GSL. Secondly a word had to have a wide range. A member of a word family, therefore, 'had to occur 10 times in each of the four main sections of the corpus and in 15 or more of the 28 subject areas' (Coxhead 2000, 221). Finally a member of a word family needed to have occurred 'at least 100 times in the Academic Corpus' (Coxhead 2000, 221). Coxhead (2000, 213) notes that the words in the list accounted for approximately 10% of the total words (tokens) in academic texts whereas they only made up 1.4% of the total words in a collection of fiction texts of the same size. The much higher prevalence of the AWL words in academic texts is evidence that they are indeed academic words.

In developing the AWL, Coxhead (2000) drew on the work done in compiling the university word list (UWL) (Xue & Nation, 1984) and tried to address some of its weaknesses. Waring and Nation (1997, 16) suggested that after learning the 2,000 words on the GSL, EAP students should study the UWL. Since the AWL is an improvement on the UWL, a similar suggestion to EAP students can be made, except that they study the AWL, rather than the UWL. The value of knowing both the GSL and the AWL can be seen in their combined coverage of English texts. Coxhead (2000, 225) found that the first 1,000 words of the GSL

covered 71.4% of the academic corpus, the second 1,000 words of the GSL covered 4.7% and the AWL covered 10%. Thus, taken together, the GSL and the AWL cover 86.1% of the academic corpus. Although this takes students closer to the 98% coverage of a text which Nation, as mentioned earlier in the chapter, maintained as necessary for comprehension, it still falls short of that target.

The New Academic Word List (NAWL)

The NAWL is very similar to the AWL. The first version was compiled by Browne et al. in 2013 (and an updated one in 2014), not so much because they found great fault with the AWL, but rather because, just like the AWL was meant to complement the GSL, Browne and his colleagues felt they needed an academic word list to fit well with their newly created NGSL (Browne et al. 2014). The NAWL contains 963 words which were selected from a 288 million word academic corpus. Much of the corpus (86.3%) comes from academic texts used in the Cambridge English Corpus. Approximately 12% of the corpus is derived from top selling textbooks across the four academic categories: arts and humanities, life sciences, physical sciences and social sciences. The final component of the corpus, which only comprises about 1% of the corpus comes from oral academic discourse. These words also come from the four categories mentioned above. The oral component of the corpus comes from a roughly equal distribution (about 1.5 million words each) of two academic spoken corpora, namely the British Academic Spoken English (BASE) corpus and the Michigan Corpus of Academic Spoken English (MICASE) (Browne et al., 2014). While the GSL and AWL cover 87% of this corpus, the NGSL and NAWL cover 92% of it, giving EAP students more coverage of academic texts.

2.2.2.9.3 Technical vocabulary

Technical words are words that are particular to a certain subject area and consequently they differ from subject to subject. Zhou (2010, 14) states that technical words tend to make up 5% of the words in a text. Flowerdew (1993, 236) claims that many words used in academic environments do not fall neatly into either a technical, general or academic vocabulary category. These words could be described as sub-technical in that they make up a 'middle ground between specialised and general' (Baker 1988, 92). Sub-technical words are words that are in general usage; however, in a technical or specific subject area they have a special or particular meaning. Baker (1988, 92) provides, as an example of sub-technical vocabulary, the word *solution*. Its meaning in general English is different to its technical meaning in the

specialised field of chemistry. Zhang (2013) also identifies that technical meanings of words may come from combinations of general and academic words in the form of collocations (Zhang 2013, 165). These collocations may be regarded as sub-technical vocabulary.

Since technical words have specific meanings pertaining to their subject area, they may be completely unknown to students, whether the students are L1 or ESL students. Therefore, these words tend to be taught by the content teacher. Nation (1990, 19) states, ‘learning the subject involves learning the (technical) vocabulary. Subject teachers can deal with the (technical) vocabulary but the English teacher can help with learning strategies’. Of course ESL teachers, knowing the importance of understanding technical vocabulary to understanding the subject, will tend to also spend time on teaching technical vocabulary. The ESL teacher, however, can probably fulfil a greater role by teaching sub-technical vocabulary. Since content teachers may not devote any time to clarifying the meaning of these words as they are used in their subject area, this becomes the task of the ESL teacher (Flowerdew 1993, 236). The importance of the ESL teacher helping students with the learning of high frequency technical words is backed up by Sutarsyah, Nation & Kennedy (1994, 48). They argue that teachers can draw attention to these words’ “narrower meaning and pointing out the parts of their meaning that are important for their use in (a) specialised text”.

2.2.2.10 Written and spoken vocabulary

Due to their different communicative purposes the vocabulary used in speaking and writing often differs. Situations involving writing tend to be more formal than those requiring speaking. This view is supported by Pikulski and Templeton (2004, 2) when they state that ‘written language is more formal, more complex and more sophisticated than spoken language’. One could argue that this is not always the case if one compares the informality of a written text message between friends to that of a planned state of the union speech by a president of a country, but in the academic context of a high school from which this study is drawn, Pikulski and Templeton’s (2004) view has weight. Lee (2001, 250) agrees when he comments that ‘the lexis of spoken language is rather different from that of written language’.

Nevertheless, Lee’s (2001) view is that the written and spoken vocabulary ‘divide’ should be viewed in terms of a continuum, one where there is a core of high-frequency vocabulary in the middle that is common to both written and spoken language, but where there is also vocabulary towards the ‘extremes’ of either end of the continuum that are more common to either spoken or written language (Lee 2001, 274).

Pikulski and Templeton (2004, 2) mention that ‘we tend to have a larger group of words that we use in reading and writing’. Support for this claim comes from Lee (2001, 271) who compared the spoken and written subcorpora of the BNC and discovered that written texts indeed make use of a much greater variety of words than spoken texts. One reason for this could be that many speech acts are spontaneous in nature, not allowing for complex planning, while writing often allows for ‘repeated reading and close analysis’ (Crystal 1995, 6). By analysing one’s writing before having it read, one has more time to add a greater variety of words. Also from a stylistic point of view writers are encouraged to use synonyms if an idea is to be repeated. In contrast, since speech is so often spontaneous, a speaker often does not have time to think of vocabulary that will add variety to his/her discourse. In fact repeating the same word multiple times may be helpful to listeners, especially in a classroom where there are a number of ESL students. Support for this comes from Oxford and Scarcella (1994, 234) when they state that ‘the frequency with which learners are exposed to lexical items affects their acquisition of these items’. Nation (1990 in Sokmen 1997, 241-242) states that a range of 5-16 ‘encounters with a word’ is needed for a student to acquire it, while a much greater number – ‘well over 20 or even 30 meetings’ (Waring and Takaki, 2003, 151) could be required to learn new words. Certainly, repeating the same vocabulary could be a deliberate strategy of teachers to help their students learn new vocabulary. The topic of frequency of occurrence will be discussed in more detail in section 2.2.4.

To conclude this section, it is worthwhile going back to the research done by Nation in terms of frequency in spoken and written texts. To deal with unscripted spoken English (the kind used in conversation) knowledge of 6,000-7,000 word families is needed to gain 98% coverage, while more words (8,000-9,000 word families) is needed for the same coverage of written texts (Nation 2006, 79).

2.2.2.11 Receptive and productive vocabulary

It was mentioned earlier that ESL students should have knowledge of a certain number of words in order to function in various English environments. However, what constitutes knowledge of a word and how to measure this is very difficult. One way of measuring word-knowledge is through the distinction between receptive and productive mastery (Schmitt 2010, 79-80).

Knowledge of receptive (also referred to as *passive*) vocabulary is knowing a word when it is seen or heard, whereas productive (or *active*) vocabulary knowledge is the ability to produce

that word in speaking or in writing. Generally learners have more receptive than productive vocabulary knowledge (Schmitt 2013, 80). In a study conducted by Laufer (1998, 265) it was found that receptive vocabulary increased by 1,600 word families in one year of high school, however only half that (approximately 850 word families) could be used productively. This productive use was measured by Laufer and Nation's (1999) Productive Vocabulary Levels Test (PVLТ), which measures *controlled production* of vocabulary through the use of a modified cloze test. Despite the high increase in passive vocabulary, and reasonably high increase in *controlled productive vocabulary*, Laufer (1998, 263) found that in her study there was hardly any progress in *free productive vocabulary* (this was measured by having students freely write from a writing prompt). From Laufer's (1998) study one can deduce that it is easier for students to gain knowledge of receptive vocabulary than it is to gain productive knowledge. It is possible then for students to understand what their teachers are saying in class (*receptive knowledge*), but then not be able to express themselves using the same vocabulary in speaking and writing (*productive knowledge*).

Laufer and Nation's (1999) Productive Vocabulary Levels Test (PVLТ)

The PVLТ used in Laufer's (1998) study is the same test which I used to test the vocabulary levels of the students in this study. It was mentioned earlier that the PVLТ is a *controlled productive test* that tests students' 'ability to use a word when compelled to do so by a teacher' (Laufer and Nation 1999, 37). The PVLТ tests five frequency word levels, specifically the 2,000, 3,000, 5,000, University Word List (UWL), and the 10,000 word levels. There are three versions of each levels test. Each test contains 18 test items. In each sentence there is one target word (at the vocabulary level pertaining to that particular test) which is incomplete. The first letters (usually two or three) of that word are provided, which the test taker is required to complete in order to show knowledge of and the ability to produce that word. Table 2.3 shows a sample of some of the test items in version A of the 2,000 word level test:

Table 2.3 Sample questions from Laufer and Nation's (1999) productive vocabulary levels test at the 2,000 word level

1. They will restore the house to its orig_____ state.
2. My favourite spo_____ is football.
3. Each room has its own priv_____ bath and WC.
4. The tot_____ number of students at the university is 12,347.

The examples in Table 2.3 come from Version A of the 2,000 word level test, which was used to test the FIS students in this study. Each of the 18 target words in each level's test are words in the same word frequency band. Therefore, all the target words in the 2,000 word PVLT are in the most frequent 1,000 – 2,000 English words. Laufer and Nation (1999, 41) admit to the subjective nature of judging a test taker's mastery of a vocabulary level through the use of the test, but suggest that a score of about 15 out of 18 (or 83%) shows adequate mastery of a level. (Refer to Appendix A for the PVLTS used in this study.)

2.2.3 What vocabulary does a high school student need and its implication for teaching?

High school students need to know general service vocabulary, general academic vocabulary as well as technical vocabulary specific to each of the subjects they are studying. Firstly, a high school student needs to be proficient in general service (or common core) vocabulary since these words make up most of what is said and written across a variety of fields. Sutarsyah et al. (1994, 46) found that the 2,000 general service words in the GSL made up 78% of a general academic corpus and 82% of an English for specific purposes (ESP) text (specifically an economics text). Of course, in English classes novels are often studied; here general service vocabulary plays even a greater role. Hirsh (1992 in Waring and Nation 1997, 15) found that the GSL made up almost 91% of the vocabulary in short novels.

Since high school students are using English for academic purposes it will therefore be useful for them to know academic vocabulary. The 570 words in the AWL constitute approximately 10% of academic texts (Coxhead 2000, 213), and therefore would be worthwhile for students to know. Having said that, the texts used to compile the AWL were taken from a university context – and therefore at a level that could be too advanced or unnecessary for high school

students. Also, Coxhead found that the AWL only made up 1.4% of fiction texts (Coxhead 2000, 213). One could argue whether it is worth it for high school students to learn the 570 academic words if they only make up 1% of the texts they are studying in an English literature class. Since there is no academic word list available specifically for high school students, it may be a consideration for high school ESL teachers and students to go through the AWL and pick out words that will be most useful to students at high school level. In search of a common core of academic words (Granger and Paquot 2010) compiled a list of 106 key verbs that are commonly used across a range of academic texts. These words, which are reprinted in Table 2.4, are of a general academic nature and could be useful for high school as well as university students.

Table 2.4 Shared key verbs in business, linguistics and medicine (Granger and Paquot, 2010)

account, achieve, affect, analyze, apply, assess, assign, associate, base, calculate, characterize, classify, compare, comprise, compute, conduct, consider, consist, construct, contribute, correlate, correspond, define, demonstrate, denote, depict, derive, describe, determine, develop, differ, differentiate, document, eliminate, enhance, establish, estimate, evaluate, examine, exclude, exhibit, explore, facilitate, favor, focus, form, generate, highlight, hypothesize, identify, illustrate, improve, include, increase, indicate, influence, initiate, interact, investigate, involve, lack, limit, link, maintain, manifest, measure, minimize, model, modify, note, obtain, occur, participate, perform, predict, present, process, produce, promote, provide, range, reduce, refer, reflect, relate, report, represent, require, result, reveal, review, score, select, show, signal, study, suggest, summarize, support, target, test, underlie, use, utilize, vary, yield

Two difficulties with teaching a common core of academic vocabulary that Granger and Paquot (2010, 7-8) acknowledge are that the words may be used differently across academic disciplines and also that corpus-based studies suggest that the specificity of academic disciplines should be emphasised. Nevertheless Granger and Paquot's (2010) list shows that there is a core of verbs that are used to organise and structure academic texts in general. In order for ESL teachers to teach these words effectively it may be worthwhile to impart their general usage or similarities across disciplines as well as how they are specifically used in each discipline. Instead of studying all the words in the AWL, it may be worthwhile for high school students and teachers to begin with focusing on the 106 common core academic verbs, picked out from the AWL by Granger and Paquot (2010). Alternatively, since the AWL is presented in the form of ten sublists, teachers and students could simply work through some

of the high frequency sublists only. Teachers could explicitly teach the words and identify how they are used in the students' various subjects in order to help them produce the words in speaking and writing.

Finally, high school students need knowledge of technical and sub-technical vocabulary which are specific to the subjects that they are studying. These words differ from subject to subject and must be known if the student is going to understand each of his or her subjects. Technical words are typically found in the glossary section of each chapter of a subject's textbook and may be highlighted in the text of the chapters. With regards to understanding sub-technical vocabulary ESL teachers play an important role. Dictionary definitions may not give students the precise meaning of a word as it is used in the context of a particular subject, consequently Zhang (2013) argues that a corpus is an 'efficient and effective tool to observe the behaviours of individual words and groups of words'. ESL teachers can therefore use corpora, specific to a certain subject area, in order to explicitly teach how words are used in such contexts. It is hoped that the corpus as a whole, as well as the sub-corpora, compiled for this study can be an effective tool for ESL teachers, which they can use to teach the meanings of the technical (in addition to academic and general) vocabulary spoken in the specific contexts of various high school classrooms.

2.2.4 How many repetitions do learners need exposure to for the uptake of new words?

This section is relevant to my study because it deals with the potential learning of new vocabulary for students through teacher talk. It should be said at the outset that frequency of occurrence or repetition of words is important for learning new words, whether the input mode is reading or listening, and specifically 'weaker learners are indeed more dependent on frequency than stronger ones' (Zahar, Cobb and Spada, 2001).

Current research shows that it tends to be more difficult for learners to learn new words from listening than it is from reading. The results of a study of Japanese university students of English literature conducted by Brown, Waring and Donkaewbua (2008) show that learners gained more new words from the reading-while-listening and reading-only modes than from the listening-only mode. Two reasons that Brown et al. (2008, 148) offer as to why learning new words from listening-only is difficult are: Firstly, the speed of stream of speech is not under the control of the ESL listener, and because of this it is possible for those listeners not to recognise the spoken form of words that they might actually know in their written forms.

Secondly, in speech many words flow seamlessly into one another, making it difficult to identify where one word begins and ends. This is certainly not ever a problem with language in its written form.

Since it is more difficult to learn new words from listening than reading, one can infer that more repetitions of words are needed during listening in order for them to be learned than if the mode of input is reading. Indeed, Brown et al. (2008, 152) found this to be the case. Specifically it was found that 45% of participants in their study required 7-9 repetitions of a word before it would be learned in the reading-only mode and 46% of participants required 7-9 repetitions of a word in the reading-while-listening mode. However, with the listening-only mode, even if a word was repeated 10-13 times, only up to 36% of participants were able to gain uptake (Brown et al., 2008, 153). Whereas in written language 6-16 exposures (Rott 1999, 609; Zahar et al. 2001) might be needed for there to be uptake of new words, Brown et al. (2008) found that more than 30 repetitions of a word might be needed for uptake to occur if the mode of input is listening.

To sum up, whether the mode is reading or listening, repetition of words is important for uptake of new words to be possible, and more repetitions are required for uptake to occur through listening than through reading.

2.2.5 Incidental vs intentional learning of vocabulary

Incidental learning refers to 'learning without an intent to learn' (Laufer & Hulstijn 2001, 10). This means that during an activity, such as reading, words are learned even when the original intent of doing the activity is not to learn vocabulary, instead the intention might be overall text comprehension or reading for pleasure. Intentional learning of vocabulary, on the other hand, takes place when the intention of taking part in an activity is to learn vocabulary.

Typically studies on incidental learning tend to determine how much incidental learning of vocabulary takes place through extensive reading. Grabe and Stoller (2002, 259) define extensive reading as learners reading 'large quantities of material within their linguistic competence'. The key to extensive reading programs is that learners read a lot from material that is within their grasp; most of the vocabulary contained in the material should be understandable, so that learners' reading can be pleasurable. Reading for pleasure is the 'intent' suggested in the earlier quote by Laufer and Hulstijn. One advantage of extensive reading is that new words that are met have the potential to be learned incidentally because

the words and context around the new word allow for meaning to be inferred. Nevertheless, Waring and Takaki's (2003, 130) research reveals that while extensive reading can 'develop and enrich already known vocabulary', it is not an effective method for learning lots of new vocabulary. Research shows that new vocabulary is not learned much at all through incidental learning. Waring and Takaki (2003, 131) found that, on average, only one out of 25 new words was remembered after a three month period.

If little learning of new vocabulary is learned incidentally while reading, it is even less so during listening. Horst (2010) researched how much teacher talk supported incidental vocabulary acquisition and concluded that 'attending to teacher speech is an inefficient method for acquiring knowledge of the many frequent words learners need to know' (Horst 2010, 161). One reason for this is because the teacher discourse used for the study showed that words were not repeated enough in order for uptake to take place. It has already been mentioned that as many as 30 or more exposures of a word are required for words to be learned during listening, but Horst (2010, 174) found that most of the words that were used that many times were already familiar to students. Many of the unfamiliar or new words that students were exposed to were only repeated 6 times – which is not enough for uptake in a listening context to occur. Van Zeeland and Schmitt's (2013) research revealed similar findings. They explored incidental vocabulary acquisition through L2 (second language) listening and found that only 8.5% of the meaning of new words could be recalled, and this was on a test immediately after the listening. However, the researchers did find that participants in the study were able to acquire knowledge of word form relatively easily through L2 listening (Van Zeeland and Schmitt. 2013, 615).

Since it is difficult to learn new vocabulary incidentally, it is beneficial for students to employ intentional learning. Bereiter and Scardamalia (1989, 363) define intentional learning as 'cognitive processes that have learning as a goal rather than an incidental outcome'. Teachers can encourage intentional learning by including explicit exercises for students to do as they read or listen. Yali (2010, 85) found that 'reading plus explicit exercises' instruction results in better retention (of new words) than incidental learning instruction'.

Depth of knowledge

Yali found that doing vocabulary exercises leads to students acquiring a 'greater numbers of words as well as greater depth of knowledge' (Yali 2010, 85). Depth of knowledge means knowing a word in its various forms, how it is used with different grammatical patterns, its

collocations and how it can be used in certain situations. By creating explicit vocabulary exercises related to a written or spoken text, teachers will be helping their students increase their vocabulary levels and give more depth to their vocabulary knowledge. This view is backed up by Van Zeeland and Schmitt (2013, 622) when they state, ‘the low acquisition rate of word meaning found here, as well as in other incidental learning studies, emphasises once more the importance of combining incidental learning with some sort of explicit focus’.

Pedagogical implications

Although it has been shown that learning words incidentally requires a number of encounters with a word for it to be learned, acquiring words incidentally should not be discouraged. Classroom time might not be enough to provide students with an adequate number of encounters with a word for it to be learned, so students need to be encouraged to meet new words in a variety of contexts outside of class. This can come from extensive reading, listening to the radio, watching television, taking part in hobbies that involve a lot of language use and the internet. Indeed, Ebner and Ehri (2013, 480) state that ‘The Internet may accelerate movement toward acquiring comprehensive knowledge about word meanings by providing students with immediate access to multisensory and varied experiences with words’. The multi-sensory nature of the internet, with the user being able to see, read and listen, allows students to encounter words in different contexts. McLaughlin and Rasinski (2015) also promote the important role that incidental learning plays in vocabulary learning. They state that teachers must create an environment that promotes word consciousness. In other words teachers should motivate students to enjoy learning new words and have an interest in building their vocabularies. This requires teachers to develop students’ curiosity in words (McLaughlin & Rasinski 2015, 63).

Despite the importance of incidental learning, the pedagogical implication of the findings discussed in this chapter is that teachers need to also make learning of vocabulary intentional. Since subject teachers might not repeat new words enough for uptake to occur, ESL teachers should provide their students with supplementary material such as vocabulary exercises to help students learn these words. McLaughlin and Rasinski (2015, 68) provide some examples of how intentional learning can be encouraged such as the use of graphic organisers and the explicit teaching of root words, prefixes and suffixes.

Supplementary material could also come in the form of a teacher corpus like the one created for this study. If ESL teachers knew the vocabulary levels of their students (this could come from testing them using the PVLIT), then by going through the teacher corpus they could pick out vocabulary that the students should know (but do not) and direct their teaching to the learning of these words. Since using corpora could prove valuable for the learning of new vocabulary, the topic of corpus linguistics is discussed in the remainder of this literature review.

2.3 Corpus linguistics

Corpus linguistics is ‘the study of language in use through corpora’ (Bennett 2010, 2). In other words it is an attempt to understand language by analyzing the language contained in corpora. To add to this, Bennett (2010, 3) maintains that corpus linguistics does not explain why a language is used, nor does it attempt to provide all the possible uses of language at a given time, and it also does not answer the questions of what language is possible, impossible, correct or incorrect. Instead, corpus linguistics tells us for what and how (not why) language is used.

Cook (2003, 73) states that corpora can reveal ‘the patterns and regularities of language use’. Researchers tend to attempt to understand these ‘patterns and regularities’ through either a corpus-based or a corpus-driven approach. A corpus-based approach to research, according to Tognini-Bonelli (2001, 65), is used to ‘expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study’ whereas corpus-driven research is where ‘the linguistic constructs themselves emerge from analysis of a corpus’ (Biber, 2012). To illustrate this difference, if, for example, one had studied that the ‘correct’ grammar of a sentence was, *I have been sitting here for ages* as opposed to *I have been sat here for ages* then a corpus-based approach would involve looking for these two different sentence patterns in a corpus or a number of corpora to see which pattern is used when and by whom in order to verify one’s view. However, if one strictly followed a corpus-driven approach then one would not have any pre-conceived notions of which sentence pattern was ‘correct’ before looking at a corpus. Instead, in a corpus-driven approach, one’s first port of call is the corpus, and then any descriptions made about the use of language would come only from the way the language is used in that corpus, rather than any pre-existing knowledge.

2.3.1 What is a corpus?

According to Flowerdew (2013, 160), a corpus is a ‘collection of language, usually held electronically, which can be used for the purposes of linguistic analysis’. Corpora derive from written or spoken language and can come from a variety of sources, such as novels, newspapers, cell phone conversations, academic essays and classroom lectures. A key feature of corpora is that the language contained in them occurs naturally and this is noted in Bennet’s (2010, 2) definition of a corpus, which he defines as ‘a large, principled collection of naturally occurring examples of language stored electronically’. Indeed the corpus compiled in this study, as is the case in other corpora, includes authentic language as it is used in the real-life situation of a high school classroom.

Early corpora were compiled by hand but since the advent of computerised technology in the twentieth century corpora tend to be stored electronically. One of the earliest corpora to be stored electronically is the Brown corpus, developed at Brown University, USA, in the 1960s, and which consists of one million words. At the time the Brown corpus would have been considered large, but not in today’s terms; the British National Corpus (BNC), for example, has a total of over 100 million words.

Some corpora have been created with a specific academic purpose in mind, such as the Michigan Corpus of Academic Spoken English (MICASE). Others, like the Corpus of Contemporary American English (COCA) may be more general and have a variety of uses, such as the creation of dictionaries and in COCA’s case, to track changes in the language over time (Davies, 2008).

2.3.2 Types of corpora

There are several types of corpora but many of them can fit into two broad categories, namely generalised corpora and specialised corpora. The majority of corpora, whether they are generalised or specialised, tend to come from written sources. However, corpora, like the one created in this study, can come from solely spoken texts or a combination of spoken and written texts.

2.3.2.1 Generalised corpora

Generalised corpora tend to be large and can contain words reaching into the millions. According to Aston (1997), they aim at a broad general coverage of language production.

Although no corpus, large or small, can account for all possible language, Bennett (2010) maintains that generalised corpora can give users a general picture of how a language is used as a whole. Two well-known generalised corpora that are relevant for this study are the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA).

The BNC is a sample of some 100 million words of present-day spoken and written British English (Leech, Rayson & Wilson, 2001). It is ‘present-day’ in that a large majority of the texts used for the corpus come from the period 1985-1994. Approximately 90% of the data comes from written texts and 10% from spoken texts. One can see the general nature of the BNC when one looks at the variety of the texts chosen to make up the corpus. These include fictional written texts such as novels, plays and poems and non-fictional expository texts that come from a variety of domains, such as commerce, world affairs and natural science. Not only does the spoken component incorporate a variety of texts, for example sports commentaries, classroom interactions and sermons, but it is also derived from the language spoken by people of different ages and social groupings in order to account for English variations. The BNC plays a role in this study specifically as it is used along with COCA in the analysis of the vocabulary profile of the teachers recorded for this study.

The Corpus of Contemporary American English (COCA) is a corpus of American English created at Brigham Young University, and whose main researcher is Mark Davies. On the COCA website it is claimed that this corpus is ‘the largest freely-available corpus of English’ (Davies, 2008). The corpus consists of over 450 million words which come from an equal distribution of spoken, fiction, magazine, newspaper and academic texts. COCA was compiled by including 20 million words from each of the years from 1990 to 2012. This design, as asserted on the website, makes the COCA a suitable corpus to track changes in American English through the 22 years of its compilation.

2.3.2.2 Specialised corpora

Bennett (2010, 13) describes a specialised corpus as one that ‘contains texts of a certain type and aims to be representative of the language of this type’. Specialised corpora in English are most likely to be used in English for Specific Purposes (ESP). Bennet (2010) notes that the Academic Word List (AWL) is an example of a word list that was generated from specialised corpora for ESP, namely that of academic purposes. Because specialised corpora are derived from specific text types only, so as to capture the particular language of a certain field, they tend to be smaller than generalised corpora. Three specialised corpora which feature in the

findings chapter of this study are The Michigan Corpus of Academic Spoken English (MICASE), the British Academic Written English Corpus (BAWE) and one created from the language used in the Science class at Sultan Qaboos University, Sultanate of Oman (SQU).

Although the MICASE is considerably smaller than the BNC and COCA, it is larger than many specialised corpora in that it contains approximately 1.8 million words. It is specialised in that it is comprised of only academic speech spoken in a university setting (Nesselhauf 2005, 3). The corpus was compiled from transcribed speech recorded during academic events such as lectures, seminars, lab meetings and office hours at the University of Michigan. One of the questions that drove the project was: *‘What are the characteristics of contemporary academic speech — its grammar, its vocabulary, its functions and purposes, its fluencies and dysfluencies?’* (Simpson, Briggs, Ovens, and Swales, 2002). The MICASE team has used this corpus to help teachers create better teaching and testing materials for ESL students. The team argues that spoken English in an academic setting differs from that which is written and also from everyday spoken discourse and, since grammar and vocabulary books are mainly based on written language, there is a need for developing alternative teaching materials, which are based on spoken content. In the same way that MICASE is a corpus of speech taken from an academic context, the FIS corpus compiled for this study is an academic one taken from a high school setting.

The BAWE corpus is considered a specialised corpus because it is comprised of written texts taken from a university academic setting in the United Kingdom. There are approximately 6,5 million words in the corpus which come from academic writing composed by students at three tertiary institutions: University of Reading, University of Warwick and Oxford Brookes University. The texts were collected from a number of subject areas across four disciplines: Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences. The truly specialised nature of the BAWE corpus is revealed when one focuses on only one of the subject areas, rather than looking at the language used in the corpus as a whole. The majority of the texts in the corpus are essays, but other types of writing, such as recounts, reports and case studies are also included. Approximately 1,000 students from all university years submitted a total of roughly 2800 texts to make up the corpus (<http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf>).

The small, specialised Science corpus from Sultan Qaboos University, Sultanate of Oman (SQU), a corpus of academic discourse, was created in order to help the students of the

university come to terms with the English needed for their academic studies. The corpus specifically includes language used in the science (or biology) class. The corpus is made up of a total of 104,483 words which come from both selected readings and transcriptions of verbal lectures from several lecturers. Once that was done, selections of the corpus were used as input for the English course. This was done so that the material taught in the English course would correspond with the actual language used in the science course (Flowerdew 1993, 232). This corpus is relevant to my study because the corpus which I have created (the FIS corpus) is also comprised of transcriptions of verbal lectures from a classroom setting, and, just like Flowerdew's (1993) corpus, it includes discourse from the science classroom.

2.3.2.3 The value of a small specialised corpus

Large generalised corpora like the Collins Birmingham University International Language Database (COBUILD), which has approximately 4.5 billion words, have been effective in providing information on general English usage from which general ESL resources have been created. However, analysing a small specialised corpus can be a 'valuable activity ... as a means of discovering the characteristics of a particular area of language use' (Aston 1997). Flowerdew (1993) illustrates the value of a small, specialised corpus with his analysis of the specialised science corpus created at SQU. Like the generalized corpus, COBUILD, the most frequent words in the specialised corpus were grammatical (or form) words (such as *the*, *and*, *of* and *is*). However, when one looks beyond the initial high-frequency grammatical words, one begins to see differences, for example the conjunction *so* was used much more frequently in the science course when compared to COBUILD, suggesting that cause and effect relationships play a greater role in academic environments than in everyday general English environments.

A further difference between the two corpora is in the frequency of certain types of nouns. Flowerdew (1993, 236-237) notes that the twenty most frequent nouns in COBUILD are all different to the twenty most frequent nouns in the science corpus. Unlike the nouns in COBUILD (examples of which are *people*, *world*, *day* and *house*) the common nouns in the specialised corpus are technical or sub-technical words. Examples of words that occur frequently in the science corpus are *cytoplasm*, *membrane*, *structure* and *concentration*. These are words that one would encounter more regularly in the domain of science or other similar academic environments than in general, everyday usage. This kind of comparison is similar to the kind of comparison I have done in this study, except instead of using

COBUILD, I compared words from the FIS corpus to that of words from the New General Service List (NGSL, which was discussed earlier in the chapter), MICASE and the BAWE corpus.

Finally, Flowerdew (1993, 235) found that the past tense form of verbs occur more frequently in the generalised corpus than in the biology corpus. Flowerdew (1993, 235) cites the example of *was* which is the tenth most frequent word in COBUILD but only the fiftieth most frequent word in the specialised corpus. Flowerdew (1993) contends that this finding is probably due to the relatively common use of past narratives in general English usage, while in a science class the present tense is more often used for descriptive purposes. A similar finding comes from an analysis of MICASE. Alejo, Adel, Kruis and Swales' (2007) study into the use of phrasal verbs in MICASE reveals that verbs like *ends up* tend to be used in the present tense and not in the past. *Ends up*, for example, was used in its present simple form 73% of the time, while it was used only 25% of the time in its past form and 2% in the progressive.

Biber et al. (1998, 161) also found a much higher frequency of past tense verbs in generalized corpora than in specialised corpora. They compared two ESP corpora (one was comprised of ecology research articles and the other history research articles) to a general one comprising English fiction. Biber et al. (1998) acknowledge differences between all three corpora, even between the two specialised ones, indicating that if one is to use a specialised corpus to inform language teaching, then one must create teaching materials for a particular subject from a sub-corpus taken from the same discipline. Biber et al. (1998) found that the history texts were more narrative in nature (and hence made use of more past tense verbs) than the ecology texts, but they did not make use of nearly as much past narrative as the general fiction texts. Like Flowerdew's (1993) argument, mentioned in the previous paragraph, Biber et al. (1998, 161) claim that the ecology texts make use of a high frequency of present tense verbs and a large number of 'generic statements' because they deal with 'general processes and findings about plants, animals, and their environments'. Furthermore, since people do not play much of a role in ecology texts (unlike in history texts), fewer third person pronouns are used in ecology texts than in history ones. ESP teachers can make use of such findings, to inform what specific language structures and vocabulary to include in their syllabi, as well as how best their ESL students can communicate that language in speaking and in writing.

2.3.2.4 Written vs spoken corpora

This section shows that words are used with different frequency depending on whether they are written or spoken. Schmitt (2010, 14) states that ‘the frequencies of lexical items differ considerably between spoken and written discourse’. For this study, I made a conscious decision to create a corpus based on speaking, and not writing, because of the lack of spoken corpora created from academic contexts. I also thought it would be useful for students to have access to a spoken discourse corpus because of the limited opportunities they had to listen to authentic spoken discourse outside of the classroom. Support for the need for a corpus based on speech comes again from an analysis of MICASE. In the same study mentioned previously by Alejo et al. (2007) it was discovered that the five most frequent phrasal verbs in MICASE are *go on*, *go through*, *go back*, *ends up* and *figure out*. Alejo et al.’s. (2007) paper focuses on the phrasal verb *ends up*. The researchers found that *ends up* was much more frequently spoken in academic contexts than in written or general speaking environments. They reasoned that this was due to the phrase having a ‘causal meaning’, which is not its common use in general English. The example they cite is:

they secrete these enzymes and stimulate the stroma to produce more enzymes, which ends up digesting a path through the surrounding tissues.

Although this study does not cover phrasal verbs, Alejo et al.’s (2007) findings support the need for a spoken corpus, such as the one created in this study. More evidence pointing to the value of a spoken corpus can be seen by looking at an analysis of Flowerdew’s (1993) specialised corpus from SQU. Some lexis (words and phrases) that one might expect to encounter often in academic environments was shown by the corpus to be relatively infrequent (Flowerdew 1993, 236-237). Vocabulary, such as *therefore* and *as a result*, which one might predict to be used frequently because of their academic nature, were not used much by the science teachers recorded in the experiment. Instead, the synonym *so* was used very frequently. This could be that in academic writing one is encouraged to use a variety of more formal lexis, whereas in a spoken context the repetition of the same more basic words could make it easier for students to process information.

It would be appropriate to comment on the distinction between basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP) here. As mentioned in section 1.2, BICS refers to the ability to communicate in a conversational manner (as friends in a social context would do), while CALP refers to the more formal,

academic language (both oral and written) encountered in a (school) teaching and learning environment (Cummins 2008, 71). While CALP is important for success in school, Flowerdew's (1993) findings indicate that BICS is also important for understanding teachers' spoken discourse, as is evidenced by the teachers' preference for the use of *so* (a word associated with BICS), over more CALP associated words, such as *therefore* and *as a result*.

2.3.3 Analysing a corpus

In order to make analysis of language possible there are a number of corpus software tools available, such as the free program *VocabProfile* (<http://www.lextutor.ca/vp/comp/>) and *WordSmith Tools* (Scott, 2012: <http://www.lexically.net>), which requires registration and payment. There are numerous features of language that these tools can analyse, two of which are *word frequency* and *collocation*. *Word frequency* involves analysing vocabulary in terms of what words are used more often than others and *collocations* refers to two or more words that regularly co-occur. An example of a *collocation* is 'make mistakes'. The verb *make* (as opposed to another verb, such as *do*) tends to be used with the noun, *mistakes*. Both these concepts will be discussed in more detail later in the chapter.

2.3.3.1 VocabProfile

VocabProfile can measure a corpus's word frequency by analysing the vocabulary according to a number of established word lists. Some of these established word lists are: the General Service List (GSL), the Academic Word List (AWL), The British National Corpus and Corpus of Contemporary American English frequency lists (BNC/COCA) and updated versions of the GSL and AWL in the form of the New General Service List (NGSL) and New Academic Word List (NAWL).

2.3.3.2 WordSmith tools

Two common ways to use *WordSmith Tools* (Scott, 2012) are in the creation of *word lists* and to see the context in which selected words are used through a *concordancer* tool. With the wordlist tool one can see a list of all the words in a text. The list can be set to show the words in alphabetical or frequency order. Table 2.5 shows the kind of information that software programs like *WordSmith Tools* can generate.

Table 2.5 Sample of new general service list in order of frequency

RANK	LEMMA
1	THE
2	BE
3	AND
4	OF
5	TO

Table 2.5 shows the five most common lemmas (a lemma being a word and its inflected forms) in the new general service list. These words are listed in frequency order. Therefore, the word *the* is the most frequently used word in the corpus.

Programs like *WordSmith Tools* also have the capacity to compare word lists created from different texts. Flowerdew (1993, 235-236) illustrates how, in terms of word frequency, corpus software can be used to analyse language. He compared frequently used nouns in COBUILD to frequently used nouns in the small specialised science corpus, SQU (refer to section 2.3.2.2) and found that there was a marked difference between the two noun frequency lists: ‘None of the top twenty nouns in COBUILD occurs among the top twenty nouns of our specialist corpus’ (Flowerdew 1993, 235). In this case corpus software helps us understand that the vocabulary, in terms of content, found in a specialised field, such as Science, can differ vastly from the content vocabulary found in general English usage. Consequently if ESL teachers need to prepare their students for the vocabulary demands of specific subject areas, then Flowerdew’s (1993) and Biber et al’s. (1998) findings suggest that they incorporate vocabulary from those specialised areas, as opposed to focusing only on general service vocabulary.

A second way to use *WordSmith Tools* is with the *concordancer* tool. With a *concordancer* one can choose a word, the node, and then set the *concordancer* to show for example six words, or however many one wants, on either side of the node word. (The more words one selects the better chance one has of observing the context in which the word is used.) The *concordancer* will display a list of all the node words found in the corpus situated in the centre of the screen along with the words found on either side of it. One can then see the immediate context in which that word is used which helps students in understanding its practical meaning. In addition, by looking at the words on either side of the node word one can identify which words typically collocate, or go together with, the node word. Such

information (as found in Hunston 2002, 69) would look like this, with *gaze* being the node word:

by Philippe Halsman, in which the **gaze** of the elderly Duke of Windsor
muscular head and penetrating **gaze**, Licks is a big puppy
one below the other and if my **gaze** happens to flick downwards at nearly
sat down in a chair to wait, her **gaze** moving round the shop. Rose had
our way and men and women would **gaze** closely upon me, too. Would they
a break for it. The other nurse's **gaze** switched to something over my

Since *WordSmith Tools* can inform what vocabulary is used frequently and hence what vocabulary is useful (although frequency is not the only indicator of usefulness), and how words are used through concordancing, corpora with the aid of *WordSmith Tools* can be used to improve language teaching.

2.3.4 Corpora and language teaching

It was mentioned in the previous section that corpora can be used in the design of language teaching material as they are a useful source of language because of their authentic nature. From a corpus, we can see 'how language is used today and how that language is used in different contexts, enabling (teachers) to teach language more effectively' (Bennett 2010, 7). Indeed, large, general corpora have proved invaluable in the creation of teaching syllabi (Römer 2008, 114). Thus a syllabus where the focus is communicative competence for general purposes (BICS), can make use of a corpus composed from conversations among friends to highlight lexis that ESL students can expect to encounter in informal situations, while academic corpora can be used to inform students and teachers what academic vocabulary is used in academic environments for improving their CALP-based language use.

Bennett (2010, 8) identifies the language teaching areas of phraseology, lexicogrammar, ESP, and appropriate syllabus design that corpora can inform. The AWL is one example of how corpora can 'address ESP, in this case, academic purposes' (Bennett 2010, 11). Coxhead (2000, 227) states that 'the AWL might be used to set vocabulary goals for EAP (English for Academic Purposes) courses, construct relevant teaching materials, and help students focus on useful items'. It is hoped that the corpus compiled for this study will be used similarly, that is to say high school ESL teachers will be able to use the corpus to identify the vocabulary

used by teachers, and then be able to help ESL students understand and use those words. This should assist ESL students in better understanding their teachers when they hear them speak those words, as well as be able to speak and write those words themselves.

Biber et al. (1998, 79-83) comment on how ESL textbooks could be improved by taking into account information gained from corpus-based research. They use an example of how students' use of lexico-grammar can be strengthened through the use of corpora. (Lexico-grammar is the intertwining of grammar and vocabulary to the extent that they cannot be studied separately.) When it comes to subject *that*-clauses – a subject *that*-clause being the words in square brackets in the following sentence: [*The fact that in many insect groups the newly emerged adults show a slow gliding type of flight linked to dispersal*] suggests, however, that the earliest winged insects may have evolved in temporary habitats of small erect plants, Biber et al. (1998) state that although text books provide grammatical explanations of subject *that*-clauses, they tend not to explain when these clauses are appropriately used. According to Biber et al., (1998) a study of corpora will show that subject *that*-clauses are 'almost never used by native speakers in spoken discourse' and therefore 'there is little to be gained by having students practice (them) orally'. By including authentic examples of when to use various lexico-grammatical structures, like subject *that*-clauses, in language textbooks and teaching, students will be better informed as to their functional use and thus be better able to produce those sentence patterns appropriately in various contexts.

Another way that corpora can be used to make language teaching more effective centres on the area of word frequency and how this is important for the selection of appropriate teaching materials. As indicated earlier, computer software programs are able to analyse corpora in a way that reveals the percentages of high- and low-frequency vocabulary. Traditionally, high-frequency vocabulary (words used often in a language) includes words up to the 2,000 word level (Nation 2001, 32). If a teacher runs a text through a program like *VocabProfile* to find that a very high percentage, for example 98%, of the words are high-frequency words, then it will be safe to say that that text, or lexis from that text, will be suitable for beginner students of a language. If students are at an intermediate or advanced level then it will be lower frequency vocabulary that will be difficult to comprehend. The data generated from such software can inform teachers what percentage of low frequency words are contained in a text as well as what that vocabulary is. This will help teachers decide whether or not that text is suitable for the students (Nation 2001, 33). If a teacher knows the vocabulary levels of the students as well as the vocabulary load of the text then the teacher can identify what lexis he

or she should be teaching the students. Indeed, in this study, Grade 9 and 10 students' vocabulary levels were tested, and the vocabulary load of their teachers' spoken discourse was analysed using corpus software. The aim is to use the insights gained from measuring the students' vocabulary and the teacher discourse to help inform ESL teachers at international schools what vocabulary they should teach their students to better prepare them for understanding what their content teachers are saying in class.

Corpora can also be used in language teaching in what Hunston (2002, 185) refers to as 'consciousness-raising activities', which can be brought about with the aid of concordance lines. Hunston (2002, 184) acknowledges that corpora do not 'teach' language; rather they act as evidence for how language is used. Here teachers can provide students with an extract of authentic text and then draw students' attention to lexis which they think may be useful for students to learn. However, since that lexis may only occur once or twice in the selected extract – and hence does not give students the full extent of that item's meaning and usage – a number of concordance lines involving that lexis can be selected from a corpus. By looking at a number of concordance lines in which that lexis is used, teachers and students will be able to gain a better understanding of that lexis and be able to recognise how it is used in one or more sentence patterns (Hunston 2002, 185-187).

Concordance lines also help in regard to the study of phraseology (or formulaic sequences), which refers to the meanings of words as they relate to the words around them. The identification of formulaic sequences (of which collocations is a part) has been made easier by the field of corpus linguistics. Indeed corpus linguistics has shown that 'a far greater proportion of language use is composed of collocations than was previously imagined' (Cook 2003, 73). Lewis (1997, 20) agrees when he states that 'much of our mental lexicon is stored as prefabricated multi-word chunks'. As mentioned in section 2.3.3.2, concordance lines created from corpora can be used to identify formulaic sequences and Bennet (2010, 9) goes as far as to say that 'only through corpus study can we find the details of phraseology – collocations, lexical bundles, and language occurring in preferred sequences'. Wei and Huo (2011, 709) state that formulaic sequences 'are stored and extracted as a whole' and thus 'play an important role in saving language processing efforts'. (Wei and Huo 2011, 712) go on to mention that knowing formulaic sequences helps in terms of listening comprehension because if a formulaic sequence involves, for example, four words, listeners do not need to devote a long processing time to decode each word separately in order to comprehend meaning. If they know the 'chunk', their 'cognitive load in language processing' is reduced

and their ability to achieve the ‘language proficiency’ of native speakers is improved (Wei and Huo 2011, 712). These advantages of knowing formulaic sequences show that it would be helpful for students if language teachers included them in their teaching syllabi.

2.3.5 Issues surrounding the building of a spoken corpus

Corpora are not ‘simply a collection of texts’ but rather they aim to *represent* a language as a whole or part of that language (Biber et al. 1998, 246). The design of a corpus, therefore, needs to take into account what that corpus is supposed to represent. Biber et al. (1998) illustrate their point about *representivity* by providing the example that if one’s corpus is made up solely of conversations between teenagers, then the language in that corpus cannot be said to represent all conversations. Consequently, a corpus should be compiled from a variety of sources or texts in order to be fully representative of the whole. Hunston (2002, 28) provides the example that if one seeks to understand the language of newspapers, then a corpus of that language should be compiled from a range of newspaper types, such as broadsheet and tabloid, as well as a variety of newspaper article types, such as sports, editorials and hard news.

Diversity is another issue in corpus design. Biber et al. (1998, 248) state that ‘a well-designed corpus must ... represent the different registers of the language’. Hunston (2002, 29) suggests that if one wants to research spoken language then one can account for different registers by making a list of variables that would take into account the backgrounds of the speakers (their age, social class, gender and where the speakers live) in addition to the language settings, for example radio broadcast, casual conversation and classroom talk. A *balanced* corpus can be created by collecting similar amounts of data from each heading. Since the corpus created for this study involves classroom discourse for specific subjects, the issue of diversity may not be as central a criterion as it may be for a more generalised corpus. The issue of *balance*, however, was an important consideration, which is why I wanted to compile a corpus that comprised a similar number of words across four subject areas.

Both Biber et al. (1998, 248-249) and Hunston (2002, 25-26) comment on *size* as being an important issue in corpus design and allude to the idea that a corpus should be large enough to provide reliable data. Too few texts and too few words could mean that only a small number of all the possible variations of particular lexis are accounted for. Biber et al. (1998, 249) state that many common grammatical structures can be observed in a 1,000 word sample of a corpus; however, for much less common grammatical features a larger selection of a corpus

would be required. In support of Biber et al.'s (1998) argument, Carter and McCarthy (1995 in Hunston 2002, 26) claim that a small corpus is sufficient for observing grammar in spoken language. The process of collecting spoken data is more difficult than collecting written data due to the time-consuming task of transcribing recordings. Therefore, some researchers who are looking to build a balanced corpus may have to restrict the amount of written data included in their corpora in order to match the amount of spoken data they have (Hunston 2002, 26). Hunston (2002, 26) concludes her discussion of *size* as an issue in corpus design by arguing that any debate on the optimum size of a corpus is academic, because most researchers simply use as much data as is available without being too concerned with what is not at their disposal.

2.4 Conclusion

The purpose of this literature review was to provide an overview of vocabulary studies and corpus linguistics. It included insights into different types of vocabulary and what constitutes high, mid and low frequency words. Through the discussion on vocabulary this chapter also sought to reveal what kind of vocabulary, as well as how much vocabulary high school students need in order to succeed in an academic environment. In terms of corpus linguistics, this review provided a discussion on different types of corpora, how corpora can be analysed using computer software tools, some issues in building a corpus as well as how corpora can be used in language teaching.

CHAPTER 3

RESEARCH METHOD

3.1 Introduction

This chapter describes the research method used to answer the study's two main research questions which involve identifying both (i) the productive vocabulary levels of Grade 9 and 10 high school students at an international school where English is the language of teaching and learning and (ii) the vocabulary profile of the spoken discourse of Grade 9 and 10 high school teachers at the research school. The method used to assess the students' vocabulary levels will be discussed in this chapter. To answer the second research question a small corpus of recorded speech of high school teachers was created in order to inform the vocabulary levels used by teachers while they were teaching. The corpus was also created so that one could determine what percentage of the teachers' speech comprised academic and general usage vocabulary. This chapter describes the research method used in creating the aforementioned corpus. This chapter also presents background information on the participating teachers and students involved in the study, as well as the location of the study. Details of the type of research, the research instruments, pilot study, and the analysis of the data are also discussed.

3.2 Type of research

This study falls on the quantitative side of the qualitative-quantitative research continuum since it involves the relatively objective collection of measurable learner data as well as data from which generalisations can be made regarding the vocabulary nature of teachers' discourse. The assessment of students' vocabulary levels through a written test and the empirical data which that produced also makes this study a quantitative one. However, unlike much quantitative research, which uses hypotheses and hence are deductive (that is to say they test a theory), this study was more exploratory and inductive in nature. Since the data were collected from naturally occurring speech, this study has a non-experimental design.

3.2.1 Approach

The study is analytic because once the data were collected, computer software was used to reveal what percentage of the teachers' speech was general service or general academic vocabulary. The analysis performed by the software also provided information about what

percentage of the teachers' speech was at various frequency levels. I used an analytical approach when I compared the academic words spoken by the teachers to the academic words in Coxhead's (2000) Academic Word List, and when I made comparisons between my frequency list and other lists, namely the NGSL and samples from MICASE and BAWE. I also used an analytical approach with regards to dealing with the results of the vocabulary levels tests written by the students.

A corpus-driven, as opposed to a corpus-based, approach was used in this study. Whereas a corpus-based approach makes use of a corpus as supporting evidence for a pre-existing assumption or hypothesis, a corpus-driven approach involves little to no assumptions of how or what language is used prior to the analysis of the corpus. The corpus itself is where the data are obtained or, as Tognini-Bonelli (2001, 84-85) put it, 'the corpus itself should be the sole source of our hypotheses about language'. Because the vocabulary in speaking tends to be less formal than in writing, I suspected that the vocabulary used by the teachers recorded in this study might not include as much academic vocabulary as academic written texts. However, the teacher discourse may be more academic in nature than general conversation (because the speaking context is an academic one). Consequently, I did not have any strong convictions or hypotheses that I wanted to test. My study is a corpus-driven one, because I wanted to create and analyse the corpus first, before analysing the data.

3.2.1.1 Size, diversity, representativeness and balance

In section 2.3.5 of Chapter 2 I discussed the importance of representativeness and balance in creating a corpus. With regards to the issue of *size*, the corpus which I compiled is small in comparison to others. It is composed of a total of approximately 37,000 words. The issue of *diversity* was addressed by attempting to collect a similar number of words from each sub-corpus. The sub-corpora of English, Science and Mathematics are over 8,000 words, while the History sub-corpus is made up of over 11,000 words.

In order for my corpus to be representative of the language spoken by teachers at FIS, only words that teachers, not students, spoke were used. I also attempted to make my corpus *representative* of teacher spoken discourse by recording different teachers with different ages and backgrounds. Naturally, because of its small size, the corpus is not representative of teacher talk in all contexts. Furthermore, all recordings of teachers were conducted while teachers were delivering lessons in the classroom. Words read from a textbook or any other written source were not included, in order to keep the corpus representative of spoken

discourse. It helped that I was present in most of the classrooms, and that I administered the recordings. In this way I was able to maintain scientific rigor. I also maintained rigor by personally doing the transcriptions without the aid of voice recognition software tools. I was confident that I would be able to transcribe accurately what the teachers had said. Firstly, since I was present most of the time when the recordings were done, I was fully aware of the context in which words were spoken, so if there was a word that perhaps was not completely audible, I could accurately tell what the word was. Furthermore, since the teachers who were recorded were colleagues whom I talked to on a regular basis, I was comfortable with their accents and their speech patterns, and thereby felt confident that I could transcribe their words accurately. Lastly, there were times when I could not decipher a word. On these occasions I simply moved on to the next word I could decipher (without attempting to transcribe the indecipherable word) and transcribed again from there. I also maintained rigor by having a teacher colleague (not part of the research) check a random selection of recorded excerpts and the corresponding transcriptions to verify that I had transcribed accurately. (Refer to Appendix E)

In terms of maintaining rigor when administering the PVLTs, I did this by being present while the participants wrote the tests, and by keeping a test condition atmosphere.

3.2.1.2 Reliability and validity of the productive vocabulary levels test (PVLt)

Laufer and Nation (1999, 38) conducted two studies to check the reliability and validity of the PVLt. Initially native speakers were required to answer the tested items in four versions of the test, in the presence of a researcher. If there was difficulty in retrieving an item, the researcher made an amendment to the target item, such as the inclusion of an additional letter or the sentence context was modified. Later the modified tests were administered to four groups of subjects, all of whom were English L2 high school students. In terms of scoring the tests, each item was graded as to whether each item was correct or not. If there were minor spelling or grammar mistakes the subjects' answers were marked correct. The reliability indices for Version A of the test are given in Table 3.1, where the Kuder-Richardson formula KR21 was used.

Table 3.1 PVLТ reliability results as conducted by Laufer & Nation (1999, 39)

Level	Reliability
2000 level	.77
3000 level	.81
UWL	.84
5000 level	.84
10000 level	.90

A reliability index of between .70 and .90 is considered highly acceptable (Calkins, 2005). As one can see from Table 3.1, of the five levels tested by the PVLТ the 2,000 word level test had the least reliability of .77 (albeit still reliable) and the most reliable was the 10,000 level. Taken together, version A of the PVLТ had an ‘internal consistency of .86’ (Laufer & Nation 1999, 39), making the PVLТ a reliable test.

In addition to being a reliable test, the PVLТ was shown to be a valid measure of vocabulary growth as Table 3.2 below shows, using mean raw scores.

Table 3.2 English L2 participants’ scores during Laufer & Nation’s testing of PVLТs’ validity (1999, 39)

	10th grade (n = 24)	11th grade (n = 23)	12th grade (n = 18)	University (n = 14)
2,000 level	11.8	15.0	16.2	17.0
3,000 level	6.3	9.3	10.8	14.9
UWL level	2.6	5.3	7.4	12.6
5,000 level	1.0	3.9	4.7	7.4
10,000 level	0.0	0.0	0.9	3.8
TOTAL	21.7	33.4	40.1	55.8

Table 3.2 shows the results of the L2 participants who were administered the test. The mean score which is presented in the table is the average score out of 18 (the number of test items at each level). Thus 24 students in Grade 10 took the 2,000 word level test and the mean score of all the students was 11.8. As one can see, the higher the grade level (starting with Grade 10 and ending with university students), the better the scores at each of the five frequency levels. This, coupled with the fact that with each increase in vocabulary level (hence an increase in

the difficulty of the tests) the participants' mean scores become lower, indicates that as students become more proficient in the language, their scores on the tests become better. From these results Laufer and Nation (1999, 41) concluded that the PVLTs are, in terms of developmental validity, a 'valid measure of vocabulary growth' and thereby in terms of construct validity, the test measures what it is supposed to, i.e. the vocabulary levels of participants.

3.2.2 Theoretical purpose

The theoretical purpose of the study is more heuristic-inductive than deductive. This is because no strong preconceived ideas existed about the data prior to its collection. The research questions focused more on **what** the lexical profile of teachers' speech was, rather than analysing how or **why** those teachers were using that vocabulary.

3.3 School context

The school where this study was conducted and where I worked as an ESL teacher for three years, is an international school in Japan. It is an English medium school which follows the International Baccalaureate (IB) diploma curriculum in Grades 11 and 12 and the IB primary years program until Grade 5. In the middle years (Grades 6-10) the school follows its own 'home-grown' curriculum. Teachers in the middle years draw on their experience working in private and public schools around the world as well as current international teaching trends in order to inform the middle year's curriculum. Of course, since the IB diploma is studied from Grade 11, much of the skills and content taught in the middle years (particularly in Grades 9 and 10) are intended to prepare students for the rigorous IB diploma.

Since FIS is an international school with affiliations to international school bodies, the majority of teachers are hired from abroad. For many teachers, moving to Japan and teaching at FIS is their first experience of living in Japan and many tend to stay for two to four years only before moving on to teaching in another country.

Even though FIS is an English-medium international school, it does not have the numbers of L1 English users that one might find in major international centres such as Tokyo. Approximately one third of the students at FIS are Japanese citizens, one third are Korean, and the remaining third come from other countries (with sizeable numbers from the UK and USA). That said, the majority of those whose citizenship is neither Japanese nor Korean actually have one parent who is Japanese. Many of these students have spent most of their

lives in Japan, meaning that even though they may be registered in the school as British or American (or any other country where English is the national language), their L1 could well be Japanese, and not English.

Even though the school caters for students from wide age ranges (K3 - Grade 12) it is a small school, with a total of 220 students. In the secondary school (Grades 6 -12) there is only one class per grade. When I started this study in the 2013-14 academic year, there were twelve students in the Grade 10 class and eleven students in Grade 9.

3.4 Participants

The participants in this study were teachers whose voices were recorded while they were teaching, and Grade 9 and 10 students who were tested in order to assess their vocabulary levels.

3.4.1 Teachers

Six secondary school teachers participated in the study. Table 3.3 provides information about each teacher:

Table 3.3 Teacher backgrounds

	Subject	Grade	Gender	Nationality	English level
Teacher A	English	9	Female	Romanian	Fluent, L2
Teacher B	English	9	Male	Canadian	L1
Teacher C	English	10	Male	South African	L1
Teacher D	History	10	Male	Australian	L1
Teacher E	Mathematics	9 &10	Male	Costa Rican	Fluent, L2
Teacher F	Science	9	Male	Mauritian	Fluent, L2

Three teachers of English were recorded, while only one teacher of History, Mathematics and Science respectively was recorded. The English and Mathematics recordings came from both Grades 9 and 10, whereas the History recordings only came from Grade 10 and the Science recordings came from Grade 9 only. I would have liked to have had recordings of at least two

teachers for each subject as well as both grade levels for each subject; however, with me being the main recorder as well as being a full-time teacher at the school, it was not feasible to get the range of recordings I wanted.

Half of the teachers recorded speak English as a first language, while the other half are fluent in English although it is not their first language. Teacher E speaks English as a second language but was schooled in a British school in Costa Rica and all his teaching experience has been in English-medium schools. Although teacher A was schooled in a Romanian school, she has spent most of her adult life either living or teaching in an English environment. Both teachers A and F speak English at a mother-tongue level. Teacher F also speaks English as a second language, and which I, as a native English speaker, judged to be fluent. However, his command of general English is not quite as good as that of teachers A and E. He speaks with a relatively heavy accent and makes fairly common subject-verb agreement errors. Nevertheless, he has spent many years teaching and living in English-speaking environments and he is a specialist in the field of Science, therefore, he communicates confidently in the Science classroom.

3.4.2 Students

A total of 23 students wrote the PVLTs. Table 3.4, which follows, provides some background information on each of the students.

Table 3.4 Student backgrounds

	Grade	Gender	Nationality	ESL Programme
Student A	Grade 9	Female	Japanese	Yes
Student B	Grade 9	Male	Korean	Yes
Student C	Grade 9	Female	Brazilian	Yes
Student D	Grade 9	Female	Korean	Yes
Student E	Grade 9	Male	Japanese	Yes
Student F	Grade 9	Male	Korean	Yes
Student G	Grade 10	Male	Japanese/Chinese	Yes
Student H	Grade 10	Male	Japanese	Yes
Student I	Grade 10	Female	Korean	Yes
Student J	Grade 10	Female	Korean	Yes
Student K	Grade 10	Female	Japanese/Chinese	Yes
Student L	Grade 10	Female	Korean	Yes
Student M	Grade 9	Male	American	No
Student N	Grade 9	Female	Indonesian	No
Student O	Grade 9	Male	Bangladeshi	No
Student P	Grade 9	Male	Korean	No
Student Q	Grade 9	Male	Korean	No
Student R	Grade 10	Female	Korean	No
Student S	Grade 10	Female	South African	No
Student T	Grade 10	Female	Korean	No
Student U	Grade 10	Female	Japanese	No
Student V	Grade 10	Female	Japanese	No
Student W	Grade 10	Female	Korean	No

As one can see from Table 3.4, half of the students (students A-L) were in the ESL programme at the time of the first PVLTA administration when the students were either in Grade 9 or 10. The ESL programme involved 2 lessons per week with an ESL teacher, as well as from 2 to 4 lessons per week with an ESL teacher giving in class support to the ESL students during Science and History lessons. The remaining 11 students who were tested (students M-W) did not receive ESL support in either Grade 9 or 10, although it is possible that some of those students had received ESL support in the past, and had exited the ESL

programme prior to entering Grade 9. Students' ages ranged from 14-17 through the duration of the study. Of the 23 students who were tested 14 were girls and 9 were boys. The majority of students hold passports of countries whose official language is not English, which means that some students who are not in the ESL programme are also not totally proficient in English. Having never lived in an English speaking country, nor having had prior schooling where English is the language of learning and teaching could partly account for low vocabulary levels (relative to L1 English speakers in English speaking countries) from the majority of the students tested. These results will be presented in Chapter 4.

None of the 12 ESL students, prior to this study, had ever lived in a country where English is widely spoken. Students A and E came straight to FIS from the Japanese school system; similarly, students B, F, I, J and L came to FIS from the Korean school system. Students G, H and K all had some schooling at international schools in China.

The majority of the non-ESL students (students M, N, P, S, T, U, V and W) had had some schooling in English speaking countries or had been schooled abroad in international schools which have English as their language of learning and teaching. Despite both students O and R not having Japanese passports most of their schooling prior to entering FIS had been in the Japanese school system. Student Q had only ever been schooled in the Korean school system.

3.5 Research instruments

3.5.1 Voice recorder

A voice recorder was needed to record teachers while they were speaking in the classroom. All recordings were done on the inbuilt recorder on an iPhone and uploaded and stored on a password-protected computer.

3.5.2 VocabProfile

VocabProfile was one of two corpus linguistics software tools that were used to analyse the nature of the teachers' discourse. *VocabProfile* was specifically used to identify what percentage of words in the whole FIS corpus and the sub-corpora were at various general service and academic vocabulary frequency levels. After entering the *VocabProfile* website at <http://www.lex Tutor.ca/vp/comp/> you paste your text in the space provided, select one of the word lists or combination of word lists mentioned above, and then click on the 'submit window' button in order for the program to perform the analysis. Depending on which word

list you selected, the program can identify what percentage of the words in the corpus are at various 1,000 word frequency levels. This is very useful to see how many words in that corpus can be said to be high, mid or low frequency words.

3.5.3 WordSmith tools

WordSmith Tools was the second corpus linguistics software tool used to analyse the data. It was used in order to create five word frequency lists. One list was a frequency list of all the words in the FIS corpus, while the other four lists were created from the sub-corpora of English, History, Science and Mathematics. *WordSmith Tools* presents the words in order of frequency and provides the number of occurrences of each word. Refer back to section 2.3.3.2 for a description of how *WordSmith Tools* can be used to analyse data.

3.5.4 Productive vocabulary levels test (PVLТ)

All available Grade 9 and 10 students who were present in the classrooms of the recorded teachers had their vocabulary levels tested by taking versions A and B of Laufer and Nation's (1999) Productive Vocabulary Levels Tests (PVLТ) (see Appendix A). The PVLТ is a test of controlled productive vocabulary or active vocabulary (words that someone is able to use when prompted). The productive version of the test was used in favour of a test of receptive vocabulary that might come in the form of a multiple choice set of questions and a test of free productive vocabulary which would involve test takers' composing a piece of writing. I selected the PVLТ because it is a modified cloze test (the first two to three letters of the target word are presented), which forces the test taker to show knowledge of that specific target word, rather than a substitute word that the test taker may know. It also reduces the chance that students would use a word from a different frequency level, thereby ensuring that the test is a true test of the students' competence in vocabulary at that level.

Versions of the test assess word knowledge at the 2,000, 3,000, 5,000, University Word List (UWL) and 10,000 word levels. Two versions of each test (versions A and B) are available for free download on Nation's website (<http://www.victoria.ac.nz/lals/about/staff/paul-nation>). Each level includes 18 test items. If one scores 15 (83%) or more on a level then it is fair to say one has sound knowledge of the vocabulary at that level (Laufer and Nation 1999, 41). The 5,000 word level test requires a larger vocabulary than the UWL, and since none of the students scored showed proficiency on the 5,000 level test (i.e. none scored higher than 15 out of 18), I did not administer the 10,000 level test to any of the students at FIS. More

details on how I administered the PVLTs in this study will be provided in section 3.7.1 of this chapter.

3.6 Pilot study

For the pilot study I trialled the recording of one teacher and transcribed the recorded data in order to determine the most effective method and tools. I also did a preliminary analysis of the transcriptions. I did not pilot the vocabulary testing of the students, since the test is a widely used one with a high reliability index (as shown in section 3.2.1.2). In terms of research instruments there were no changes between the pilot and the actual study. The tools I used in the pilot, namely the iPhone voice recorder, *Voice Walker*, *VocabProfile* and *WordSmith Tools* all worked adequately for what I wanted to achieve, hence the decision to use the same tools in the main study.

During my first recording I kept the voice recorder in my shirt pocket, which I avoided doing in subsequent recordings. I found that every time I moved, the movement of my shirt made a noise and this rendered the recording inaudible. For subsequent recordings I kept the iPhone stationary on a desk.

When I set out to do my pilot I intended to gather transcriptions of two teachers from each subject, since the more teachers I could record the greater the likelihood of accounting for more diversity in vocabulary use. Register variation, according to Biber et al. (1998, 248), ‘is central to language’ and ‘a fully representative corpus of a language would include (regional or social dialects)’. Unfortunately, as stated in section 3.4.1, it was not feasible for me to get as many as eight teachers to be recorded so for my actual study I had to settle on three for English and one each for History, Mathematics and Science. For my pilot study I only recorded the Science teacher.

My first recording during the pilot study lasted 18 minutes, and took more than two hours to transcribe. Since personally transcribing the recordings would be time-consuming, I considered purchasing voice recognition software. However, after researching the industry, and noting the high costs of such software, coupled with the risks of a program misinterpreting the teachers’ discourse, I decided to continue doing the transcriptions myself. Because I knew the teachers personally – their accents and registers – and I was for the most part the one doing the recordings and therefore in the class at the time, I was able to transcribe accurately. The accuracy of my transcriptions was also verified by a colleague of mine who

compared a random selection of recordings with their corresponding transcriptions. This colleague is a high school teacher, but not at the research school. She randomly selected recordings from different subjects recorded for this study and then listened whilst reading my transcriptions. She found no inconsistencies in my transcriptions.

During the pilot study recordings, to ensure that my data contained only teacher oral discourse, I paused the recorder whenever students spoke or when the teacher read from written script, and this was a procedure I continued during the actual study. While transcribing the recordings, if any word was indecipherable or if any students' discourse was recorded I simply moved on and transcribed the next decipherable word from the teacher. Again, this was an approach I continued during the actual study.

As previously stated, I ran my transcription through *VocabProfile* as one part of the data analysis. I used the 'classic' option, which analyses the data according to the General Service List (GSL) and the Academic Word List (AWL). The pilot study showed the teacher to be using 84% general service vocabulary (high frequency words up to the 2,000 word level). A little over 3% of the words were academic words as found in the AWL. As many as 12% of the words were 'off-list words', in other words, words that are neither in the GSL nor the AWL. Because of such a high percentage of 'off-list words', I did not solely use the 'classic' option for my actual study because the GSL accounts for only high frequency words up to the 2,000 word level, and it is the GSL which is used by the 'classic' option for analyses. In order to reduce the number of off-list words by accounting for words in the 2,000-3,000 word frequency band, I used the 'VP compleat' option (<http://www.lex Tutor.ca/vp/comp/>), which accounts for 3,000 high frequency words as found on the New General Service List (NGSL) and the New Academic Word List (NAWL). The inclusion of the 2,000-3,000 frequency band indeed reduced the number of 'off-list' words and it had the added advantage of aligning more closely to the PVL. In the actual study I also used the BNC/COCA 1-25k (refer to Table 4.3 in section 4.3.2 for a presentation of these results) option for analysis, which reduced the number of 'off-list' words even more.

In addition to using *VocabProfile*, I ran both the pilot study and main study transcriptions through *WordSmith Tools* after which I was able to create a ranked scale from the data which showed all the words transcribed for the pilot in order of frequency. Table 3.5 is a reduced table of the results from the pilot study:

Table 3.5 Most frequently spoken words of FIS Science teachers for pilot study

Rank Order	Type	Frequency
1	IT	62
2	YOU	61
3	THE	55
4	SO	50
5	TO	38
6	IS	37
7	AND	31
8	OKAY	29
9	A	22
10	HAVE	21
11	THIS	18
12	OF	17
13	THAT	17
14	CARBON	16

Table 3.5 reveals that the most frequently spoken word in the pilot was ‘it’, and spoken a total of 62 times during the 18-minute recording. I followed the same approach to analysing word frequency in my main study, except for my actual study I included additional word frequency lists for comparison with the FIS list.

3.7 Research procedures for main study

In terms of the procedures put in place to obtain data in the main study there were two main components, namely the testing of students’ vocabulary levels and the collection of teachers’ spoken discourse.

3.7.1 The administration of the PVLТ

The PVLТ was first administered to the students at the same time that I started recording the teachers for the main study. This was in February 2014. All Grade 9 and 10 students who

were present on the days that I tested them consented to being tested (refer to section 3.9). I made copies of version A of the PVLТ and personally administered the tests during class hours. Students in the ESL programme started with the 2,000 word level test, and non-ESL students started with the 3,000 word level test. I scored the tests immediately, so that if a test taker achieved 83% or above (indicating competence) on a level, they could write the next level during the same session. For the most part students who did not achieve 83% on a level did not write the higher levels. Firstly, some students did not want to do so because knowing that they had not achieved a high score at one level, they did not want to see an even lower score at a higher level. Furthermore there was not enough time for some of the weaker students to write more tests. Those students who did well on a specific level tended to complete the tests quickly and therefore had time to complete tests at a higher level. From a pedagogical point of view it would have been informative had all students completed tests at all levels. In this way the English teachers could have gone through students' test answers during classroom, and helped the students learn the target words that they did not know while taking the test. However, that was beyond the scope of this study. One of this study's aims was to identify the vocabulary levels of the participants. If participants failed to achieve 83% on a test, which was the case with many of the ESL students on the 2,000 level test – with many of them achieving lower than 72% - then their vocabulary level can be identified at that level or lower, and the aim of the study was achieved.

I scored the tests according to memoranda, (see Appendix B) which I created for both versions A and B of the PVLТ. I did this by doing the tests myself and by checking with the GSL, BNC and UWL word lists that my answers fell within the correct frequency range. I also sent my memorandums to my lecturers for checking.

Like Laufer and Nation (1999), I accepted spelling errors, subject/verb agreement errors and errors of tense as correct answers; however, I did not accept words if they fell outside of the word family (refer to section 2.2.1 for elaboration of what constitutes a word family). An example of a student response that I did not accept was *elective* in response to the item: *They met to ele_____ a president.*

Towards the end of the study, version B of the PVLТ was administered to all available students who were in Grade 9 and 10 at the start of the study. By the time of the second testing, approximately 14 months later, the students were in Grades 10 and 11. Unfortunately the school is small, particularly in those grade levels, so only 23 students were tested at least

once. Also, the student body is quite transitory, so five of the 23 students who had been tested at the beginning of the study were not at the school by the end of the study and were therefore not tested a second time. My approach to the second round of testing was slightly different to the first, in that I allowed students to write higher levels of the PVLТ even if they did not achieve 83% on a test. The reasons for this were, firstly because I had more time with one group of the students and secondly because by the time of the second testing I had come to know the students better and suspected that their vocabulary levels were higher than the first round of testing indicated, particularly with lower frequency vocabulary. Since there is very little exposure to English outside of the school context for many of the students and most of their exposure to English is in the classroom, I suspected that the students might have more knowledge of academic or technical vocabulary, which would be more likely to be represented in the lower frequency levels tests than for example the 2,000 word levels test. Table 4.1 in Chapter 4 of this dissertation presents the participants' scores on the PVLТ.

3.7.2 Data capturing and analysis of the vocabulary data

After administering and scoring the PVLТ, I captured the data using Microsoft Word. The sum of each student's correct answers at each level of versions A and B of the PVLТ was recorded in table form, along with the means and standard deviations for each level of each test. These data are presented in Table 4.1 in chapter 4.

3.7.3 Procedures for the administration of the recordings

The recording and transcribing of teachers' discourse was a time-consuming yet critical factor in the study, as it helped me achieve one of the aims of this study, which is: to identify the vocabulary profile of the spoken discourse of Grade 9 and 10 high school teachers. All teachers were recorded in their classroom while they were in 'lecture mode' and all recordings were done using a voice recorder on an iPhone. I made the recordings while sitting in on classes, except for the Mathematics recordings which were done by the Mathematics teacher. I would pause the recording as often as possible during times when students were speaking, since the focus of the study was on teachers' discourse. I also paused recordings as much as possible if teachers were reading from a text, since my study is on analysing teachers' spoken discourse, rather than on what is written in text books. Consequently, I did not make any recordings of lectures from the start to the finish of a lesson. This did not negatively impact on the achievement of the aims of this study, since the analysis of the teachers' discourse comes in the form of words as individual items. Although I acknowledge the

important role of formulaic sequences (of which collocations are a part) in the make-up of English, I considered my method of recording and transcribing to be adequate for focusing on individual word use.

I made an attempt to have recordings of a similar number of words from each of the four recorded subjects. The recordings were uploaded onto a computer after which I personally transcribed all of them (see Appendix C for samples of a transcribed extract from each subject). A random selection of transcriptions was checked by a colleague of mine, and no inconsistencies were found. Transcribing the recordings was a time-consuming task since I had to pause and repeat recordings many times in order to type up what was said. I did make use of a free software program called *VoiceWalker*. One can set the program to slowly ‘walk one through’ a recording by repeating a group of words a set number of times, before moving onto the next group of words. These segments overlap which assists in capturing all words. The use of *VoiceWalker* reduced the number of times I would have to pause and manually repeat what was said. However, if a teacher suddenly increased their talking speed or if I could not catch immediately what the teacher had said, I would have to stop the program and repeat the recording manually. If, after a few attempts, I could not decipher some of what was said or was not confident I heard the speaker correctly, I simply ignored that utterance and continued from the next word I could decipher. In this way I ensured that the words that make up the corpus are a true reflection of the words spoken by the teachers. However, since at various points in the corpus a word that immediately follows another word may not have been the word that was immediately spoken by the teacher, one must be cautious of using concordance lines in order to identify formulaic language from the FIS corpus.

In order to capture a similar number of words across the four subjects, there were differences in the number of recordings made for each subject. For example, the nature of the Science lessons was such that the teacher did not speak for long periods of time. This was due mainly to the teacher’s preferred teaching style. He preferred to avoid lecturing for long periods of time, in favour of the students working through material in the text book. Thus the Science sub-corpus was created from recordings that were done from many lessons over a period of months. On the other hand, the Mathematics teacher chose specific lessons to do his recording, where he knew he was going to be lecturing for long periods of time, which meant I managed to obtain the requisite number of words from Mathematics over a relatively short period of time. These differences had an impact on the nature of the sub-corpora. For example,

by recording Science over a period of months, I was able to gain data from a variety of teaching units, resulting in a greater variety of technical vocabulary being obtained. In contrast, all of my History recordings came from one long unit of work which meant that there was considerable repetition of the same technical vocabulary in that sub-corpus.

The transcriptions were stored in five separate folders on my computer. Four of the folders contained the transcriptions of each of the four subjects: English, Mathematics, Science and History. These separate folders allowed comparisons to be made between the vocabularies of the various sub-corpora. The fifth folder contained the transcriptions of all of the subjects. In order to aid my analysis of the FIS corpus, I made use of frequency word lists from other corpora, such as the NGSL, which contains 2818 words, and which was easily downloaded from the NGSL website: (<http://www.newgeneralservicelist.org/>). Since the FIS corpus was compiled from an academic context I wanted to make comparisons with other academic word lists. To do this I created two frequency word lists from samples from the MICASE corpus and the BAWE corpus (refer to section 2.3.2.2). I selected texts from the same subject areas as those used for my corpus, and I made my samples a similar number of words to the FIS corpus. Table 3.6 shows the number of words from each of the subjects taken from the BAWE and MICASE corpora.

Table 3.6 Number of words used across four subject areas taken from BAWE and MICASE corpora

	BAWE sample	MICASE sample
History	8,666	8,755
Science	8,134	10,964
Mathematics	6,737	7,236
English	9,319	12,682
TOTAL WORDS	32,856	39,637

As one can see from table 3.6 the BAWE sample included 32,856 words, of which 8,666 words came from written texts from the History department, 8,134 words came from Science texts, 6,737 from Mathematics essays and 9,319 words came from English essays.

The MICASE sample comprised 39,637 words. This was composed from 8,755 words involving 10 speakers in an ‘African History Lecture’, 10,964 words spoken by two speakers

in a ‘Biology of Cancer’ lecture, 7,236 words involving three speakers in a ‘public math colloquium’ and 12,682 words spoken by seven people in a ‘Fantasy in Literature Lecture’.

3.8 Data processing and analysis of corpus

After all transcriptions were completed I was ready to use the computer software *VocabProfile* and *WordSmith Tools* to analyse the corpus and compare sub-corpora.

I used *VocabProfile* to analyse the percentage coverage of words from various established general and academic word lists, namely the GSL, NGSL, AWL, NAWL and BNC/COCA in the FIS corpus, as well as its sub-corpora. *VocabProfile* revealed how much of the FIS corpus and the sub-corpora could be found at various word frequency levels of general and academic English. Refer to tables 4.2, 4.3 and 4.5 in chapter 4 for a presentation of these results.

I used *WordSmith Tools* in order to create a ranked scale that shows all of the words in the FIS corpus in order of frequency. This allowed me to see which words were more frequently spoken by the high school teachers than others. I was also able to use *WordSmith Tools* to compare the FIS list with other word frequency lists, namely the NGSL, and the two lists which came from the MICASE and BAWE samples. This data is presented in tables 4.6 and 4.7 in chapter 4. By comparing these frequency lists to each other I was able to identify to what extent the nature of the words spoken by the high school teachers was similar or different to the nature of the vocabulary contained in the other lists.

3.9 Ethical considerations

Prior to recording lessons and testing students I informed all parties concerned (the school principal, teachers and students) what I would be doing and what my research was about. The principal signed a consent form giving me permission to conduct the research in her school, and the teachers signed consent forms to allow me to record their lessons. Students and their parents signed consent forms informing them that lessons in which the students were sitting would be recorded and that their vocabulary levels would be tested. See Appendix D for examples of signed consent forms.

3.10 Conclusion

In this chapter the research method used in creating the small corpus for this study was discussed, as was the procedure involved in assessing the vocabulary levels of high school students. In addition to the students, background information was also provided of the

teachers whose speech was recorded and transcribed for the corpus. The method of analysing the corpus and the students' vocabularies was also discussed. The results of these analyses will be presented in the next chapter.

CHAPTER 4

FINDINGS

4.1 Introduction

This chapter begins by looking at the vocabulary levels of a number of the Grade 9 and 10 students at FIS, the school where the research into the vocabulary profile of teachers took place, followed by this study's findings regarding the nature of the vocabulary spoken by the high school teachers.

Data obtained from the transcriptions of recorded speech of the Grade 9 and 10 teachers at FIS is presented, with a view to analysing the vocabulary profile of those teachers. To recap, approximately 37,000 words spoken across four subject areas, namely English, History, Mathematics and Science were transcribed and analysed using two corpus linguistics software tools. *VocabProfile* was used to measure what percentages of the FIS corpus fell under various high to low frequency vocabulary categories, while *Wordsmith Tools* was used to create a frequency word list of the FIS corpus. In order to understand the vocabulary profile of the FIS high school teachers, this chapter presents a number of comparisons of the FIS corpus with other established word lists.

This chapter will reveal what proportion of the teachers' discourse is made up of general English and what proportion is academic. In addition, this chapter will show how much general and academic vocabulary students need in order to comprehend the words that the teachers are using. In so doing, this study provides ESL teachers with a vocabulary goal for their students in order to bridge the gap between students' current vocabulary levels and the vocabulary levels needed in order to comprehend the vocabulary spoken by teachers.

4.2 Research question 1

What are the productive vocabulary levels of Grade 9 and 10 high school students at the research school?

All available Grade 9 and 10 students at FIS wrote Laufer and Nation's (1999) Productive Vocabulary Levels Tests (PVLТ) at least once. The test items were at the 2,000, 3,000, 5,000 and University Word List (UWL) word levels. The students were first tested in February 2014, around the same time as the recording of teachers' classroom talk was started. The students were tested again between January and May 2015; by the time of the second testing, roughly

one year later, the students were in Grades 10 and 11. Unfortunately some of the students who took the first test were no longer at the school by the time of the second testing. Table 4.1 presents the results of all the tests taken by the students.

Table 4.1 Presentation of results from PVLTs

Students in the ESL programme					Students not in the ESL programme						
	2,000 word level		3,000 word level			3,000 word level		UWL		5,000 word level	
	2014	2015	2014	2015		2014	2015	2014	2015	2014	2015
A	56%				M		100%		83%		83%
B	56%				N	67%	100%				67%
C	72%				O	89%			83%	67%	67%
D	28%	56%			P	89%			94%	83%	89%
E	61%	61%			Q	83%					
F	72%	78%		61%	R	78%	78%				28%
G	94%		72%		G		94%		67%		44%
H	50%	50%		39%	S	89%		56%			
I	72%	67%		50%	T	72%		61%	72%		67%
J	67%				U		94%		44%		44%
K	56%	50%			V		94%		61%	61%	56%
L	67%				W				89%	67%	78%
Mean (%) Raw (18)	62.5	60.4		50		82	93	58.5	74	69.5	62.3
Standard Deviation (SD)	14.6	9.2		7.8		8.2	7.4	2.5	14.6	8.2	18.3
Range (min – max)	28- 94	50-78		39-61		67-89	78- 100	56-61	44-94	61-83	44-89

As one can see from Table 4.1, I grouped the results of the participants who were in the ESL programme on the left side of the table and the non-ESL students on the right side. I did this in order to identify separate mean scores for the two groups. All the ESL students present on the day (a total of 12 ESL students) took the 2,000 level test at the beginning of the study in February 2014. Only one participant, student G showed mastery of the 2,000 word level with a score of 94%. Laufer and Nation (1999, 41) put mastery at ‘around 15 or 16 out of 18 (83%

to 90%)'. By the time this student was administered the PVLТ in 2015 he had exited the ESL programme, which is why you will notice his 2015 scores reflected in the non-ESL side of the table. By placing his 2015 scores with the non-ESL students his scores could contribute to the overall mean of the UWL and 5,000 level scores for the non-ESL participants. Since no ESL students completed the UWL or 5,000 level tests, it was best to have his scores with the other (non-ESL) participants. The remaining 11 ESL students did not show mastery at the 2,000 level in 2014, and due to their low scores they did not take the 3,000 level test or higher that year. The mean score of the ESL students' results on the 2,000 level test was 62.5%, which is approximately 20% short of what can be considered mastery of the words at that level. The few students who did the 3,000 level test in 2015 (despite again not showing mastery of the 2,000 level test when they took it in 2015) had an overall mean of just 50%. The low mean scores on the tests taken by the ESL students indicate that their vocabulary levels are low. However, with the small sample size, one must be cautious in drawing generalisations from the results, and I would not suggest that these low scores are representative of ESL students in all international schools.

It is worthwhile to note that for the most part the ESL students who were in Grade 9 in 2014 showed improvement in their 2,000 level test scores in 2015, while the ESL students who were in Grade 10 in 2014 did not show improvement in their scores in 2015. Students D, E and F were in Grade 9 in 2014 and like all the participants were administered version A of the PVLТ. When they were administered version B of the PVLТ in 2015, students D and F improved their scores while student E's score remained the same. Students H, I and K were in Grade 10 in 2014. When they were administered version B of the PVLТ in 2015, student H's score was the same and students I and K's scores were actually lower. One potential factor that might account for this difference is that the ESL programme at FIS only provides support for students from Grades 6-10. It is possible then that the Grade 10 students in 2015 benefitted from being in the ESL programme, while the Grade 11 students in 2015 did not have the ESL support.

The scores of the students who were not in the ESL programme are presented on the right side of Table 4.1. Due to time constraints none of the non-ESL students were administered the 2,000 word level test and since the mean of their scores on the 3,000 word level test in 2015 (82%) showed mastery of that level, it is reasonable to suggest that they would have shown mastery at the 2,000 word level had they taken that test. The table shows an encouraging higher mean for the 2015 scores than the 2014 scores at the 3,000 level and

UWL level (although some participants in 2015 did not take the tests in 2014 and vice versa). The mean score for the 5,000 level test is lower in 2015 than that of the 2014 scores; however, this is probably due to the fact that most participants who took the test in 2015 did not take it in 2014. Of the four participants who were administered the 5,000 word level test in both 2014 and 2015, one participant's score (Student O) remained the same in both years. This student showed mastery of the UWL vocabulary and scored 67% on the 5,000 word level test both times she wrote the test. One participant (Student V) had a lower score in 2015 (56%) than in 2014 (61%) and two participants improved their scores from 2014 to 2015, by showing mastery at or just short of the 5,000 word level: Student P improved from 83% to 89% and student W improved from 67% to 78%.

In conclusion to this section, the results of the PVLTS show that overall the ESL students at the research school have on average low productive vocabulary levels at less than 2,000 words, while the non-ESL students have productive vocabulary levels of from 3,000 and 5,000 words.

4.3 Research question 2

What is the vocabulary profile of the spoken discourse of Grade 9 and 10 high school teachers at an international high school?

In order to understand the vocabulary profile of the high school teachers at FIS, the transcriptions of each of the four sub-corpora of English, History, Mathematics and Science were combined and analysed with the aid of *VocabProfile* and *Wordsmith Tools*. There are a variety of vocabulary lists or combinations of lists that one can use in order to understand the vocabulary profile of a corpus, as the following section will reveal.

4.3.1 Research question 2(a)

How much of the high school teachers' spoken discourse is made up of general and academic English?

Table 4.2 shows the data based on using the New General Service List (NGSL 1 - 3) (Browne et al, 2013) and the New Academic Word List (NAWL) (Browne et al., 2013) in order to analyse the vocabulary profile of the FIS corpus. The NGSL is made up of high frequency words in general English usage, while the NAWL is made up of words common to academic English.

Table 4.2 NGSL and NAWL frequency profile of FIS corpus

Frequency Level	Families (%)	Types (%)	Tokens (%)	Cumulative. Token %
NGSL 1 [1,000 lemmas]	738 (49.76)	1119 (45.10)	31004 (<u>84.48</u>)	84.48
NGSL 2 [1,000 lemmas]	379 (25.56)	459 (18.50)	1699 (<u>4.63</u>)	89.11
NGSL 3 [801 lemmas]	200 (13.49)	227 (9.15)	1231 (<u>3.35</u>)	92.46
NAWL [963 lemmas]	166 (11.19)	197 (7.94)	750 (<u>2.04</u>)	94.50
Off-List:		592 (23.86)	2014 (<u>5.49</u>)	99.99
Total (unrounded)	1483	2481 (100)	36698 (100)	100.00

Table 4.2 shows that 2481 types (or different words) make up the 36,698 tokens (or total number of words) in the FIS corpus. NGSL 1 refers to the most common 1,000 lemmas (a lemma being a word and its inflected forms) as found in the Cambridge English Corpus (CEC), NGSL 2 is the second most common 1,000 lemmas and NGSL 3 is the third most common 801 lemmas. The final column in Table 4.2 indicates the cumulative token percentage at each vocabulary level. Thus the most common 1,000 words in general English usage (NGSL 1) covers roughly 85% of the FIS corpus. This is in keeping with what Nation (2006, 79) states is common for spoken text. The fourth column in Table 4.2 reveals that 4.63% of the words spoken by the FIS teachers are words in the second 1,000 most common English usage words (NGSL 2). The cumulative token of the most common 2,000 words therefore covers a little under 90% of the FIS corpus. A little over 3% of the FIS corpus is made up of NGSL 3 words, bringing the percentage coverage of the most common 3,000 English words to roughly 92.5%. All of these percentages conform to what Nation (2006,79) claims to be the expected coverage of high frequency words in spoken texts, although the approximate 92.5% coverage is slightly higher than what Browne et al. (2013) discovered

when they compiled the NGSL word lists. They found that taken together NGSL 1, NGSL 2 and NGSL 3 accounted for 90% of the CEC corpus.

Another finding indicated by Table 4.2 is that the FIS spoken corpus includes many more words than the NGSL than occurs in academic written texts. Coxhead (2000) found that 76% of the General Service List (GSL) covered the written academic corpus which she used to compile the Academic Word List (AWL) (Coxhead 2000, 213). The GSL is made up of 2,000 word families and is not too dissimilar from the combined NGSL 1 and NGSL 2 lists. In all, 89% of the FIS spoken corpus is covered by the most common 2,000 words (the combined NGSL 1 and NGSL 2 lists).

From this data it is clear that knowledge of high frequency general English vocabulary is very important for students, as is evidenced by up to 93% of teachers' discourse made up of the most common 3,000 words in English. It seems to play a far greater role in high school teachers' discourse than it does in university geared academic written texts.

4.3.2 Research question 2(b)

How much of the high school corpus is made up of high, mid and low-frequency vocabulary?

Table 4.2 showed that five percent of the tokens in the FIS spoken corpus are off-list words. Since the three NGSL levels only account for high-frequency vocabulary, it is difficult to tell whether the off-list words are mid-frequency (and hence worthwhile to include in a language learning curriculum) or mainly made up of low-frequency words or words that can legitimately be left out of a general English teaching curriculum. Proper nouns and numbers also fall in the off-list category, which adds to the high percentage. In order to gain a better understanding of these off-list words, it is helpful to use the combined British National Corpus and Corpus of Contemporary American English 25,000 word list (BNC/COCA) to analyse the FIS corpus (see Table 4.3), because the BNC/COCA lists account for words with lower frequencies than the most common 3,000 words. (Refer back to section 2.2.2.4.3 for a discussion of the BNC/COCA lists.)

Table 4.3 BNC/COCA frequency profile of FIS corpus

Frequency Level	Families (%)	Types (%)	Tokens (%)	Cumulative token %
K-1 Words :	687 (41.11)	1147 (46.23)	31358 (85.45)	85.45
K-2 Words :	400 (23.94)	578 (23.30)	2104 (5.73)	91.18
K-3 Words :	233 (13.94)	298 (12.01)	1162 (3.17)	94.35
K-4 Words :	109 (6.52)	129 (5.20)	461 (1.26)	95.61
K-5 Words :	58 (3.47)	63 (2.54)	164 (0.45)	96.06
K-6 Words :	51 (3.05)	56 (2.26)	112 (0.31)	96.37
K-7 Words :	24 (1.44)	29 (1.17)	88 (0.24)	96.61
K-8 Words :	16 (0.96)	17 (0.69)	60 (0.16)	96.77
K-9 Words :	18 (1.08)	19 (0.77)	61 (0.17)	96.94
K-10 Words :	17 (1.02)	19 (0.77)	29 (0.08)	97.02
K-11 Words :	10 (0.60)	10 (0.40)	55 (0.15)	97.17
K-12 Words :	8 (0.48)	8 (0.32)	24 (0.07)	97.24
K-13 Words :	9 (0.54)	9 (0.36)	28 (0.08)	97.32
K-14 Words :	6 (0.36)	6 (0.24)	15 (0.04)	97.36
K-15 Words :	7 (0.42)	9 (0.36)	47 (0.13)	97.49
K-16 Words :	6 (0.36)	6 (0.24)	13 (0.04)	97.53
K-17 Words :	3 (0.18)	3 (0.12)	12 (0.03)	97.56
K-18 Words :	4 (0.24)	4 (0.16)	8 (0.02)	97.58
K-19 Words :	1 (0.06)	1 (0.04)	2 (0.01)	97.59
K-20 Words :				
K-21 Words :	1 (0.06)	1 (0.04)	3 (0.01)	97.60
K-22 Words :	1 (0.06)	1 (0.04)	1 (0.00)	
K-23 Words :	1 (0.06)	1 (0.04)	3 (0.01)	97.61
K-24 Words :				
K-25 Words :	1 (0.06)	1 (0.04)	2 (0.01)	97.62
Off-List:		156 (6.29)	886 (2.41)	100.00
Total (unrounded)	1671+?	2481 (100)	36698 (100)	100.00

Table 4.3 shows that high-frequency words, words up to the 3,000 word level (K1-3 words), account for 94.3% of the words spoken by the FIS teachers. This finding is similar to that shown in Table 4.2 where the FIS corpus was analysed according to the NGSL. There is, to a

large extent, an overlap of the K1-3 words with the NGSL 1-3 words. One advantage of using the BNC/COCA list over the NGSL is that it enables one to examine the use of mid-frequency vocabulary. Mid-frequency vocabulary includes words from the 3,000-9,000 word levels (Schmitt and Schmitt, 2014). Approximately 2.5% of the words in the FIS corpus are in this mid-frequency range according to Table 4.3. Taken together, knowledge of mid and high-frequency vocabulary (words up to the 9,000 word level) would give one almost 97% coverage of the vocabulary teachers use. It was stated in section 2.2.2.4.4 that knowledge of 98% of the words in a text is adequate for understanding that text. Nation (2006) argues that 98% coverage, or knowledge of 6,000-7,000 word families plus proper nouns, ‘may be needed to cope with the transitory nature of spoken language’ (Nation 2006, 79). Nation (2006, 77) found that approximately 1% of spoken texts are proper nouns. These words are not included in the word lists. Thus knowledge of 6,000-7,000 words actually gives one 97% coverage and the 98% coverage is reached with the inclusion of the proper nouns. The proper nouns are added because they are considered a ‘minimal learning burden’ (Nation 2006, 70).

It is important to look at the nature of the ‘off-list’ words spoken by the FIS teachers in order to evaluate whether these can be considered a ‘minimal learning burden’ and therefore be included as words that high school students would know. Table 4.4 contains all the off-list words spoken by the FIS teachers.

Table 4.4 Off-list words (words not in the BNC/COCA list)

<p>ab_[2] adolf_[1] africa_[1] african_[1] ah_[1] america_[9] american_[4] americans_[8] anglo_[1] animalism_[6] ashtray_[1] asia_[2] asian_[3] australia_[13] australian_[8] b_[26] bakumatsu_[2] battleships_[2] bisector_[4] bitmas_[1] blake_[4] bodyguards_[1] bohr_[2] breakdown_[1] britain_[9] british_[13] buddhism_[7] buddhist_[2] c_[7] china_[8] choshu_[3] christian_[2] christianity_[13] christians_[5] churchill_[3] classmates_[3] classroom_[6] coastline_[1] colorado_[1] confucian_[1] confucianism_[3] countryside_[1] craftspeople_[2] craftsperson_[1] daimyo_[12] deshima_[4] diagrammatical_[1] dutch_[14] e_[4] edo_[13] endometriums_[1] england_[3] english_[9] epididymis_[1] europe_[7] european_[22] europeans_[4] f_[1] figurehead_[3] fiji_[3] forever_[6] france_[2] freezed_[2] french_[3] garibaldi_[2] gasses_[1] german_[3] germans_[1] germany_[1] gestapo_[4] golding_[2] guevara_[3] h_[3] henry_[1] hitler_[8] holland_[1] homework_[2] honshu_[1] hulled_[1] humankind_[1] hydro_[1] ib_[2] india_[1] indonesia_[1] indonesian_[1] italian_[1] jack_[35] japan_[82] japanese_[31] john_[1] joi_[2] jones_[2] jpop_[1] judaism_[2] kagoshima_[4] kanagawa_[6] kokugaku_[3] korea_[5] korean_[4] kyoto_[3] kyushu_[1] latin_[2] lcd_[1] lifetime_[4] manmade_[3] mao_[3] masterless_[2] meiji_[4] micronesia_[1] motorcycle_[1] muslim_[1] mutsuhito_[1] mx_[2] n_[18] napoleon_[3] nazi_[1] neil_[1] nichrome_[4] notebook_[1] ns_[1] oh_[3] oneninenumbertwo_[1] overland_[1] p_[3] pacific_[2] pdf_[1] pedmas_[2] penotine_[1] perry_[1] philippines_[1] quadratics_[2] r_[4] ralph_[17] rangiku_[1] renraku_[2] richardson_[1] roger_[6] ronan_[1] ronin_[1] russia_[12] russians_[1] russo_[2] rutherford_[1] sacho_[2] sakoku_[8] samoa_[6] samoan_[3] sangoku_[1] seppuku_[2] shimabara_[5] shimonoseki_[4] shipwrecked_[1] shogun_[14] shogunate_[17] shoguns_[2] signboards_[1] simon_[6] skyscrapers_[1] smith_[1] snowball_[2] sonno_[2] spectromedium_[1] springtime_[1] sundial_[2] textbook_[5] textbooks_[2] tokugawa_[18] tonga_[3] triangled_[1] underground_[1] unfertilised_[1] x_[86] y_[19] yokohama_[1] z_[3] zealand_[4] zen_[5]</p>
--

Table 4.4 shows a list of words not accounted for by the BNC/COCA word list. The numbers in parenthesis alongside each word are the tokens (or number of times each word was spoken). There are 176 words in the off-list. A number of these items are letters, such as *b*, *x* and *y* which would have been spoken by the Mathematics teacher. Letters, numbers and fillers make up 9% of the off list words. Other words are proper nouns, such as *Churchill*, *Henry*, *America* and *Kagoshima*. Proper nouns make up a substantial 53% of the words in Table 4.4. Because these are off-list words, no account of word families has been made. Thus *Europe*, *European* and *Europeans* have not been shown to be associated with each other and thus are written as separate words. Other proper nouns include religions, for example *Christianity* and *Buddhism*. Other types of words in the off-list category that would be useful for learners are semantic compounds, examples of which are *shipwrecked*, *textbook* and *signboards*, and they

make up 16% of this group of words. Of course some words, such as *textbook* would be familiar in the educational context. The semantic compounds vary in their level of difficulty. The meaning of the word *shipwrecked* is not obvious, and therefore would probably require some active learning from ESL students; however, other semantic compounds would be relatively easy to learn due to the transparency of each member of the compound, such as *sign* and *board* in *signboard*. If one knew the meanings of *sign* and *board*, it should be easy to recognise the meaning of *signboard*. Technical words make up 9% of the off-list and are very subject specific. These technical words include: *Animalism*, *bisector*, *BITMAS*, *diagrammatical*, *epididymis*, *nichrome*, *PEDMAS*, *penotine* and *quadratics*. These words may be as difficult for native speakers of English to learn as they would be for ESL students, and, therefore, it can be expected that content teachers will explain these words, rather than they being solely the domain of the ESL teacher. Words from other languages such as *rangiku* and *seppuku* make up 9% of the off-list words.

One can split the off-list words into two groups: words that would probably need to be learned by any high school student; these are the semantic compounds and the technical words, which together comprise about 23% of the off-list words. The second group includes the proper nouns, words from other languages, numbers, letters and fillers. These words make up about 72% of the off-list words. This second group of words is more in line with the ‘proper nouns’ or ‘minimal learning burden’ which Nation (2006) discusses. Although the words from other languages may be difficult to learn, I included them in this second group because they do not fit into a group that students would expect to learn in order to improve their general, or even academic, English.

Table 4.3 revealed that 2.4% of the FIS spoken corpus was made up of off-list words. Since the nature of these off-list words is quite diverse we cannot treat them all like Nation’s ‘proper nouns’ and simply add the 2.4% to whatever vocabulary level a student has. The grouping that makes up 72% of the off- list words, however, is more in line with Nation’s ‘proper nouns’. These words make up 1.7% of the off list words.

We can now return to the findings displayed in Table 4.3 and determine with more accuracy the vocabulary level students require in order to achieve the 98% ideal coverage required to comprehend teacher discourse. That level can be put at 6,000 words, which is in line with Nation’s (2006) finding that 6,000 words is necessary for understanding unscripted spoken discourse. Table 4.3 indicates that 96.37% of the FIS corpus is covered by 6,000 words (K-6

words). If one adds the 1.7% of the group of off-list words made up of numbers, letters, fillers, words from different languages and proper nouns then one would reach the 98% coverage required for comprehension.

4.3.3 Research question 2(c)

How do the high school spoken sub-corpora compare with one another in terms of their coverage of general and academic English?

So far the vocabulary profile of the FIS corpus as a whole has been analysed, but when one compares the four sub-corpora a few differences between them can be observed. Table 4.5 below was created after analysing each sub-corpus using *VocabProfile*.

Just as in Table 4.1, the sub-corpora were analysed in terms of their coverage of the NGSL and the NAWL. The ‘combined’ column is a reproduction of Table 4.1 and shows the profile of the whole FIS corpus, which has already been discussed. What is of interest here is the vocabulary profile of the sub-corpora, from the four different subject areas, and in particular the percentage of words spoken at the 1,000 word level (NGSL 1). As one can see, the English and History teachers made more use of higher frequency vocabulary, 87% and 86% respectively, than the Mathematics and Science teachers. Both the Mathematics and Science teachers made use of the most frequent 1,000 words approximately 82% of the time. Whereas the English and History teachers used more vocabulary at the 1,000 word level, both Mathematics and Science teachers made more use of words at the 2,000 and 3,000 word levels, compared to their English and History peers. In fact the Mathematics lessons involved as much as four times more words than History and more than double that of English at the 3,000 word level.

Table 4.5 NGSL and NAWL frequency profile of FIS sub-corpora

	Combined	English	History	Mathematics	Science
Words in text (tokens)	36698	8367	11568	8636	8127
Different words (types)	2481	928	1294	658	971
NGSL 1:					
tokens	84.48%	87.16%	86%	81.83%	82.39%
types	45.10%	61.75%	58.27%	64.13%	53.86%
families	49.76%	65.90%	64.45%	66.87%	59.88%
NGSL 2:					
tokens	4.63%	4.09%	3.46%	5.57%	5.86%
types	18.5%	15.62%	14.37%	13.98%	15.45%
families	25.56%	19.88%	19.74%	15.53%	19.03%
NGSL 3:					
tokens	3.35%	2.72%	1.82%	6.26%	3.06%
types	9.15%	7.33%	6.49%	9.57%	9.06%
families	13.49%	9.94%	9.51%	10.35%	11.71%
NAWL:					
tokens	2.04%	1.06%	1.13%	2.63%	3.73%
types	7.94%	3.45%	4.56%	6.84%	7.72%
families	11.19%	4.28%	6.3%	7.25%	9.37%
Off List words:					
tokens	5.49%	4.96%	7.57%	3.71%	4.96%
types	23.86%	14.55%	19.94%	8.97%	17.61%

Another difference between the English and History and the Science and Mathematics sub-corpora is that both the Mathematics and Science teachers made use of more academic vocabulary (as shown by the NAWL row in Table 4.5). than English and History teachers. As one can see from Table 4.5, only a little over 1% of NAWL words were used in the English and History classes, while more than double that percentage covered the Science (3.73%) and Mathematics (2.63%) classes.

Nevertheless, all of the above NAWL percentages fall far short of what Coxhead (2000) found when she created the AWL. She noted at the time that the academic words cover approximately 10% of written academic texts (Coxhead 2000, 225). The fact that the FIS corpus is a spoken one, as opposed to written, and the context is that of a high school, as opposed to the more advanced university environment, would probably account for the lower frequency of academic vocabulary, and the higher frequency of less formal and more general English vocabulary.

It should be noted here that the high prevalence of off-list words in the History sub-corpus is not due to a high amount of mid-frequency vocabulary. Instead it is mainly due to the large number of proper nouns (for example, *Japan* and *Yokohama*) and words from other languages (for example, *rangiku* and *sakoku*). (Much of the History transcript that made up the sub-corpus was related to a unit on Japanese history).

4.3.4 Research question 2(d)

How does the nature of the words spoken by high school teachers compare with those found in other corpora?

Wordsmith Tools was used in order to investigate the most frequently spoken words in the FIS corpus. The results showing the top 50 words in FIS are displayed in the first column of Table 4.6. To the right of the FIS top 50 are three additional top 50 lists which have been included for the purpose of comparison. The NGSL, as has been mentioned, is a list of high-frequency general English vocabulary. The MICASE sample is a frequency list created from a sample which I took from the Michigan Corpus of Academic Spoken English (MICASE). (Refer to section 2.3.2.2 in Chapter 2 for a discussion of the MICASE corpus.) The BAWE sample is a list created from a sample taken from the British Academic Written English corpus. This corpus was also introduced in section 2.3.2.2. Both the original MICASE and BAWE corpora are much larger than the FIS corpus but samples (instead of the corpora in their entirety) were used for the comparison with the FIS corpus. The samples are of similar

sizes to the FIS corpus. The FIS corpus has 36,698 words, the MICASE sample has 41,366 words and the BAWE sample has 33,224 words. In choosing the samples I attempted to select similar content areas to those used in the FIS corpus. The MICASE sample included the following spoken texts: A Biology of Cancer lecture, a Public Math Colloquium, a lesson on the role of fantasy in Literature and an African History lecture. These speech texts correspond with the four subject areas that made up the FIS corpus, namely English, History, Mathematics and Science. Similarly, the BAWE sample comes from a selection of written essays, critiques and experiments from the same four subject areas written by university students. The comparison of the three corpora allows one to see the similarities and differences between the spoken vocabulary of high school teachers, general service vocabulary, the spoken English found in universities and academic written vocabulary.

Table 4.6 shows the top 50 words from the FIS corpus, NGSL and the MICASE and BAWE samples. Table 4.6 shows considerable similarity across the top fifty words of all four lists. Function, or grammatical words (as opposed to lexical, or content words) feature strongly. Indeed, *the* is the most frequently used word in all four corpora. Other grammatical words common to all four lists are: *of*, *to*, *a* and *in*. These words, though crucial in forming correct grammar, are very general and do not provide much meaning in terms of providing content.

Nevertheless, the few content words that do feature in the top 50 words of these corpora are found in the two specialised tertiary based corpora, MICASE and BAWE. *Cells* and *cancer* feature in the MICASE list and *war* and *atoms* appear in the BAWE list. Although the MICASE corpus is a spoken one and BAWE is written, both are taken from specialised academic contexts at tertiary level. The formal and specialised nature of these two contexts could be the reason for the relatively high frequency of the four content words mentioned. None of the words in either the FIS or the NGSL top 50 are obviously content words (except for possibly *right*, *going* and *know*). This is not surprising with regards to the NGSL because that list comes from a generalised corpus. However, since the FIS list comes from the fairly specialised and formal context of a high school teaching setting, one might have expected that there should be a higher frequency of content words. In fact it is true that content words feature with more frequency in the FIS list than the NGSL. It is just that they do not feature in the top 50. The word *cells*, for example, is ranked at 179 in the FIS list, while it features with less frequency on the NGSL list, where it is ranked at 502.

Table 4.6 Word frequency comparisons across 4 corpora

	FIS Corpus	NGSL	MICASE sample	BAWE sample
1	THE	THE	THE	THE
2	TO	BE	OF	OF
3	YOU	OF	AND	AND
4	IS	AND	TO	TO
5	THAT	TO	A	IN
6	SO	A	THAT	A
7	OF	IN	IN	IS
8	IT	HAVE	YOU	THIS
9	AND	IT	IS	WAS
10	A	YOU	THIS	THAT
11	THIS	FOR	I	AS
12	WE	NOT	IT	FOR
13	WHAT	THAT	UH	BE
14	IN	ON	SO	WITH
15	ARE	WITH	ONE	IT
16	HAVE	DO	WAS	BY
17	OKAY	AS	UM	WERE
18	ONE	HE	BUT	NOT
19	DO	WE	HAVE	ON
20	I	THIS	CELLS	ARE
21	NOT	AT	FOR	FROM
22	THEY	THEY	WITH	AT
23	IF	BUT	ON	AN
24	BE	FROM	THESE	HIS
25	IT`S	BY	ARE	WHICH
26	GOING	WILL	IT`S	WE
27	BUT	OR	ABOUT	HAVE
28	ABOUT	HIS	WE	HE
29	FOR	SAY	AS	HAD

Table 4.6 continued

	FIS corpus	NGSL	MICASE sample	BAWE sample
30	OR	GO	NOT	TWO
31	CAN	SHE	IF	ALSO
32	BECAUSE	SO	THEY	TIME
33	THERE	ALL	OR	MORE
34	THAT`S	ABOUT	WHAT	FORD
35	WITH	IF	THERE	THESE
36	JUST	ONE	KNOW	WAR
37	ON	MY	BE	CAN
38	WILL	KNOW	TWO	THEIR
39	TWO	THERE	AT	THERE
40	ALL	WHICH	CANCER	ATOMS
41	RIGHT	CAN	JUST	POINT
42	AT	GET	FROM	BETWEEN
43	WAS	HER	LIKE	ONLY
44	FROM	WOULD	ALL	PEOPLE
45	HE	THINK	AN	THEY
46	LIKE	LIKE	CAN	WILL
47	AN	MORE	WHICH	COULD
48	KNOW	THEIR	INTO	HOWEVER
49	HOW	YOUR	THINK	WOULD
50	NOW	WHEN	THAT`S	BUT

Since there are very few content words across the top 50 of each of the four corpora, I decided to identify the most frequent 20 common nouns in each corpus (with the view that common nouns would carry more content or lexical information). Table 4.7 shows these words.

It has already been mentioned that the MICASE and BAWE samples had common nouns represented in the most frequent 50 words of those corpora, whereas the FIS and NGSL lists did not. One could therefore expect that the most frequent 20 common nouns in the MICASE

and BAWE lists would have higher frequencies than those in the FIS and NGSL lists. This is true for the BAWE list. Table 4.7 shows that the 20th most frequent noun, *darkness*, is the 117th most frequent word in the corpus. Interestingly, the 20th most frequently spoken common noun, *light*, in the FIS corpus has a higher overall frequency than its counterpart in the MICASE corpus. Therefore, in spite of the MICASE corpus having two common nouns in the top 50 and the FIS corpus not having any, there are more common nouns in the most frequent 161 words in the FIS corpus, than there are in the most frequent 161 words in the MICASE list. Even though the most frequent 20 common nouns have an overall higher frequency than the MICASE and NGSL lists, there is not enough difference between them to suggest that high school teachers' speech is more content laden than the speech found at universities. Instead what is more likely is that high school students, many of whom are ESL students, may need more repetition of the same words than university students in order to clarify meaning, and this could result in the slightly higher frequency than the words in the MICASE list.

With that said, if Table 4.7 is anything to go by, the nature of the vocabulary spoken by the FIS high school teachers is not necessarily less advanced than that spoken in universities. One could assume then that, providing the FIS students can grasp the meaning of what their teachers are saying, they should be adequately prepared for a tertiary environment when they finally enter such institutions. Adding weight to this suggestion is the result of when I ran each of the words in Table 4.7 through *VocabProfile*. Coincidentally an equal 97.5% of the top 20 nouns from each of the FIS, MICASE and BAWE lists all came from words in the BNC/COCA most frequent 4,000 words.

Table 4.7 Frequency of common nouns across four corpora

	FIS corpus	NGSL	MICASE sample	BAWE sample
1	Minus (64)	Time (53)	Cells (20)	Time (32)
2	Question (71)	People (57)	Cancer (40)	War (36)
3	People (76)	Year (63)	Book (65)	Atoms (40)
4	Line (80)	Thing (91)	Problem (72)	People (44)
5	Point (82)	Day (97)	Time (85)	Population (52)
6	Answer (87)	Company (112)	Something (88)	Structure (64)
7	Example (92)	Child (115)	People (92)	Soldier (65)
8	Things (93)	Man (122)	Things (107)	System (69)
9	Time (109)	Life (125)	System (109)	Prey (73)
10	Something (111)	Place (129)	Metastasis (128)	Novel (77)
11	Thing (113)	Problem (136)	Fact (133)	Potential (82)
12	Word (125)	Something (139)	Growth (139)	Rate (83)
13	Equation (137)	Number (143)	World (141)	Data (87)
14	Term (143)	Country (152)	Tumor (146)	Solution (88)
15	Electrons (150)	School (156)	Page (150)	Story (89)
16	Number (156)	World (171)	Thing (158)	Program (101)
17	Samurai (157)	Report (178)	Process (162)	Ions (110)
18	Chemical (159)	House (180)	Body (163)	Model (111)
19	Distance (160)	Group (181)	Century (164)	Way (114)
20	Light (161)	Home (183)	Place (169)	Darkness (117)

Table 4.7 allows one to compare the use of more generalised content (words that are not easily categorised into a particular domain) to more subject specific or technical content. As

one might expect, the top 20 common nouns in the NGSL, the most general list of the four, are on the whole more general than the top 20 words in the three more specialised corpora. It is not easy to classify words into general or technical. Indeed some words may be seen as sub-technical. Words like *report* and *company* in the NGSL list could be argued to be sub-technical by nature. These are words that are ‘in general usage, but which have a special meaning within the technical area’ (Flowerdew 1993, 78). Being specialised corpora, it is not surprising to find technical words in the FIS, MICASE and BAWE lists. The FIS corpus has the following five words which could be classified as technical: *minus* and *equation* (which are clearly Mathematics words), *electron* and *chemical* (which are clearly Science words) and *samurai* (which would be a History word). In the MICASE list *cells* and *metastasis* are technical, while *cancer*, *tumor*, *system* and *process* may be sub-technical, bordering on technical. Both the FIS and MICASE lists are quite similar with roughly the same number of words being general, high frequency words (about 14 of the words in each of the two lists could be considered general) and the same number of words being technical or sub-technical. Indeed, the following high frequency general words are found in both lists: *people*, *thing*, *time*, *something*. Incidentally those same four words are in the NGSL list, which is not so surprising when one considers that a high percentage of the words in the FIS corpus (84%) and the MICASE corpus (79%) are words in NGSL 1.

The nature of the words in the BAWE list is quite different to the other three lists. Whereas there is a very high percentage of general words in the NGSL and a reasonably high percentage of general words in the FIS and MICASE lists, there is a lower percentage of general words and a high percentage of technical and sub-technical words in the BAWE list. Probably, the only words one would consider to be general are: *time*, *people*, *story*, *darkness* and *way*, while there are many that one might consider sub-technical. Words like *war*, *system*, *rate*, *solution*, *model* probably fall into this category. Technical words in the BAWE list would be: *ions*, *prey* and *atoms*.

Tables 4.6 and 4.7 revealed the nature of the higher frequency vocabulary in the FIS corpus. Table 4.8 looks at some of the lower frequency words in the FIS corpus and compares them to words with the same low frequency in the MICASE and BAWE lists.

Table 4.8 Comparison of lower frequency words in the FIS corpus and samples of the MICASE and BAWE corpora

Frequency	FIS CORPUS	Frequency	MICASE SAMPLE	Frequency	BAWE SAMPLE
2107	DISAGREE	2845	DISAPPEAR	3220	DISAGREEMENTS
2108	DISAPPEARED	2846	DISBARRED	3221	DISAPPEARANCE
2109	DISCHARGING	2847	DISCHARGED	3222	DISAPPEARED
2110	DISCOVERED	2848	DISCRETE	3223	DISAPPEARS
2111	DISCOVERY	2849	DISCS	3224	DISARM
2112	DISGRUNTLED	2850	DISCUSSING	3225	DISCIPLINE
2113	DISPERSING	2851	DISH	3226	DISCOLOURED
2114	DISPLAY	2852	DISHTOWELS	3227	DISCREPANCY
2115	DISSOLVE	2853	DISHWASHING	3228	DISCUSS
2116	DISTINCTIVE	2854	DISJOINTED	3229	DISEASE
2117	DISTRACT	2855	DISOBEY	3230	DISGUISES

All of the words in Table 4.8 were only used once in their respective corpora. Certainly, when comparing these three selections from the FIS corpus and the MICASE and BAWE samples, one would not think that the high school vocabulary is any less academic or advanced than the higher level MICASE and BAWE corpora. Indeed there is one word (or form of that word), *disappear* which can be found in all three lists. Few of the 10 words in the FIS selection would be considered common, everyday vocabulary, whereas *discs*, *dish*, *dishtowels* and *dishwashing* from the MICASE selection could be. At least three of the words in the FIS selection can be considered technical vocabulary: *discharging*, *dispersing* and *dissolve* which is roughly the same number, if not more, than what we find in the MICASE and BAWE selections. This is not to say that this selection of FIS corpus words is more formal or more academic than the MICASE and BAWE selections. Words like *disbarred* and *discrete* from the MICASE selection and *discrepancy* and *discuss*, while they may not be technical words, would be considered academic words. All of the words: *discharging*, *dispersing*, *dissolve*, *disbarred*, *discrete*, *discrepancy* and *discuss* are Graeco-Latinate words which are common to academic words. What we can take from Table 4.8's data is that despite the FIS corpus's relatively high coverage of high-frequency generalised vocabulary, it still contains a

percentage of academic vocabulary. This point is illustrated further by Table 4.9 on the next page, where the FIS corpus and MICASE and BAWE samples were analysed using *VocabProfile*.

Table 4.9 shows that a high percentage of roughly 85% of the words in the FIS corpus are from the most frequent 1,000 words in English, whereas these high-frequency words make up a lower percentage, roughly 80%, of the MICASE sample, and an even lower percentage of about 70 percent of BAWE. Nevertheless, the FIS corpus still has a percentage of academic words not too far from that of the MICASE and BAWE samples. The FIS corpus is made up of 2.04% of NAWL words with the MICASE sample comprising a slightly higher percentage of academic words, specifically 2.26%. As we would expect, since BAWE was compiled from academic written texts, the NAWL covers a greater percentage (3.41%) of its corpus than the NAWL in the FIS and MICASE corpora. With that said, the similar numbers of academic words across the three corpora reveal why the lower frequency words in Table 4.8 were of a similar academic or technical nature.

Table 4.9 Profile chart comparison of FIS corpus and MICASE and BAWE samples

	FIS corpus	MICASE sample	BAWE sample
Words in text (tokens)	36698	41366	33224
Different words (types)	2481	3515	4152
NGSL 1:			
tokens	84.48%	79.16%	70.35%
types	45.10%	36.47%	30.37%
families	49.76%	43.12%	39.51%
NGSL 2:			
tokens	4.63%	5.06%	7.29%
types	18.5%	17.70%	17.51%
families	25.56%	27.79%	28.69%
NGSL 3:			
tokens	3.35%	3.94%	3.92%
types	9.15%	10.13%	9.68%
families	13.49%	16.20%	17.76%
NAWL:			
tokens	2.04%	2.26%	3.41%
types	7.94%	7.48%	7.54%
families	11.19%	12.89%	14.03%

4.3.5 Research question 2(e)

What is the nature of the academic vocabulary spoken by the high school teachers?

In order to further understand the nature of the academic words spoken by the FIS teachers, the FIS corpus was compared to Coxhead's (2000) Academic Word List (AWL). Much of the analysis of the FIS corpus academic words up until now has looked at how the FIS list relates to the New Academic Word List. However the easy availability of Coxhead's (2000) AWL, and the user-friendly presentation of the words in the form of sublists meant that it was more feasible to analyse the academic words in the FIS corpus further by comparing those words to Coxhead's (2000) AWL. Table 4.10 shows the headwords of the word families in Coxhead's

(2000) AWL that are in the FIS list. The numbers in brackets in all sublists is the number of the sublist in which the word can be found in Coxhead's sublists. Thus, the headword *negate* (which is in FIS Sublist 1) is in Coxhead's Sublist 3. The number after the headword and before the bracketed number (55 in the case of *negate*) is the number of times that a word from that word family is used in the corpus.

As one can see the words in Table 4.10 are presented in the form of sublists. The inclusion of the words in this form is to correspond with the way Coxhead (2000) laid out her original lists. Like Coxhead's (2000) list, Table 4.10's sublist 1 contains the most frequently spoken AWL words in the FIS corpus and Sublist 6 and 7 have the least spoken words. All words in Sublists 6 and 7 were spoken only once. They were separated into two lists only to break up such a large number of words. All words in Sublists 4 and 5 were spoken three or two times respectively. With regards to Sublists 1-3 the number of times the word was spoken is written next to the word, for example the headword *negate* was used 55 times.

Table 4.10 shows that the most frequently spoken academic headword in the FIS corpus is *negate*. In the FIS corpus it is not actually used in its headword form. Instead it is mostly used in a few of its derivative forms, for example *negative* which was used 38 times. It may be important to note here that *minus*, which was spoken 90 times, is not part of the same word family as *negate*. Since *negate* is only in AWL Sublist 3, there are at least 120 academic words used with more frequency in the AWL corpus, making it a much more important headword in the FIS corpus. There are a number of other academic headwords that are used with a relatively higher frequency in the FIS corpus than in the AWL. These include *chemical* (FIS sublist 1 and AWL sublist 7), *theme* (FIS sublist 1 and AWL sublist 8) and *revolution* (FIS sublist 1 and AWL sublist 9). Differences in the two lists work the other way too. There are a number of academic words that were only spoken once by the FIS teachers, but were used relatively frequently in Coxhead's (2000) academic corpus. These headwords include: *approach*, *distribute*, *environment*, *indicate* and *interpret*. All of those words are in AWL sublist 1. Of course there are also words in the AWL that were not spoken at all by the FIS teachers. *Data*, for example, is a high frequency academic word in Coxhead's (2000) list (AWL sublist 1), however it was not used at all by the FIS teachers in the lessons recorded.

Table 4.10 Headwords of Coxhead’s academic word list word families in the FIS corpus

Sublist 1:					
negate 55 (3)	positive 50 (2)	equate 48 (2)	culture 35 (2)	chemical 35 (7)	
plus 32 (8)	economy 28 (1)	formula 26 (1)	react 24 (3)	sequence 22 (3)	
paragraph 22 (8)	physical 21 (3)	period 19 (1)	element 19 (2)	obvious 19 (4)	
specific 18 (1)	conflict 18 (5)	contact 17 (5)	technique 16 (3)	grade 14 (7)	
normal 13 (2)	factor 12 (1)	ratio 12 (5)	revolution 12 (9)	occur 11 (1)	
process 11 (1)	theme 11 (8)				
Sublist 2:					
area 10 (1)	function 10 (1)	communicate 10 (4)	create 9 (1)	policy 9 (1)	
previous 9 (2)	precede 9 (6)	define 8 (1)	chapter 8 (2)	stable 8 (5)	
fee 8 (6)	transport 8 (6)	release 8 (7)	unique 8 (7)	final 7 (2)	
maximise 7 (3)	decline 7 (5)	energy 7 (5)	intelligence 7 (6)	definite 7 (7)	
random 7 (8)	involve 6 (1)	major 6 (1)	percent 6 (1)	community 6 (2)	
link 6 (3)	locate 6 (3)	philosophy 6 (3)	whereas 6 (5)	reverse 6 (7)	
Sublist 3:					
concept 5 (1)	evident 5 (1)	identify 5 (1)	proceed 5 (1)	similar 5 (1)	
restrict 5 (2)	emerge 5 (4)	hence 5 (4)	status 5 (4)	image 5 (5)	
substitute 5 (5)	symbol 5 (5)	neutral 5 (6)	eventual 5 (8)	infrastructure 5 (8)	
authority 4 (1)	legal 4 (1)	respond 4 (1)	significant 4 (1)	vary 4 (1)	
feature 4 (2)	impact 4 (2)	resource 4 (2)	tradition 4 (2)	coordinate 4 (3)	
document 4 (3)	emphasis 4 (3)	task 4 (3)	ignorant 4 (6)	incidence 4 (6)	
identical 4 (7)	military 4 (9)				
Sublist 4 (all headwords spoken 3 times):					
benefit (1)	constitute (1)	individual (1)	require (1)	section (1)	aspect (2)
conclude (2)	focus (2)	institute (2)	potential (2)	transfer (2)	initial (3)
sex (3)	concentrate (4)	cycle (4)	job (4)	phase (4)	expand (5)
author (6)	index (6)	minimum (6)	adult (7)	finite (7)	automate (8)
restore (8)	portion (9)	undergo (10)			
Sublist 5 (all headwords spoken twice):					
accommodate (9)	analyse (1)	induce (8)	sphere (9)	series (4)	remove (3)
region (2)	technology (3)	domain (6)	mature (9)	consist (1)	assess (1)
initiate (6)	assume (6)	structure (1)	predict (4)	guarantee (7)	finance (1)
consequent (2)	contribute (3)	constant (3)	challenge (5)	conduct (2)	erode (9)
establish (1)	generation (5)	logic (5)	instruct (6)	exceed (6)	maintain (2)
issue (1)	code (4)				
Sublist 6 (all headwords spoken once):					
approach (1)	distribute (1)	environment (1)	indicate (1)	interpret (1)	
research (1)	role (1)	distinct (2)	administer (2)	design (2)	
achieve (2)	relevant (2)	site (2)	affect (2)	strategy (2)	
complex (2)	compute (2)	primary (2)	immigrate (3)	deduce (3)	
component (3)	partner (3)	layer (3)	rely (3)	proportion (3)	
technical (3)	access (4)	apparent (4)	dimension (4)	stress (4)	
principal (4)	summary (4)	investigate (4)	commit (4)	subsequent (4)	goal (4)
Sublist 7 (all words spoken once):					
alter (5)	aware (5)	network (5)	perspective (5)	prime (5)	reject (5)
accurate (6)	display (6)	explicit (6)	inhibit (6)	nevertheless (6)	overseas (6)
reveal (6)	aid (7)	convert (7)	differentiate (7)	eliminate (7)	file (7)
globe (7)	hierarchy (7)	infer (7)	innovate (7)	survive (7)	topic (7)
ultimate (7)	detect (8)	implicit (8)	revise (8)	vehicle (8)	device (9)
medium (9)	suspend (9)	team (9)	vision (9)	depress (10)	

There is probably more than one reason for these differences. Certainly one reason could be (and this is one of the limitations of this study) that the recordings of the FIS teachers did not come from the full curriculum of the subject areas used. Thus the headword *negate* has the

highest frequency because a number of the Science recordings were done during a unit on electricity where the words *negative* or *negatively charged* were used often. If the data had come from recordings done throughout multiple years of high school, then perhaps *negate* would lose its high frequency usage and the FIS sublists of academic words may come to resemble Coxhead's (2000) sublists more. Another reason for the differences could be due to the limited subject areas that make up the FIS corpus. Only four subject areas, namely History, Science, English and Mathematics were used, whereas the AWL came from many additional subject areas, such as philosophy, criminal law and finance.

Another point of interest as presented in Table 4.10 deals with the frequency of occurrence of the AWL headwords spoken by the FIS teachers. It was demonstrated in section 2.2.4 that more than 30 repetitions might be required for uptake of vocabulary to occur through incidental listening (Brown et al., 2008). With only six of the AWL headwords spoken by the FIS teachers being used more than 30 times (and many of those 'repetitions' might be different forms of the word), it is fair to say that most of the AWL words were not spoken with enough frequency for them to be acquired incidentally, that is, if they are words unfamiliar to the L2 students.

It may be thought that since the AWL came from mostly written texts (as opposed to the spoken FIS corpus) and tertiary environments, that another reason for the differences between the two lists is that the AWL headwords could be of a more advanced nature. The similarities between the two lists, however, suggest this not to be the case. A look at Table 4.11 below lends support to this view.

Table 4.11 **The number of AWL word families in the FIS corpus from each AWL sublist**

AWL sublists	Number of families in FIS corpus from AWL sublist
Sublist 1 (60)	44
Sublist 2 (60)	35
Sublist 3 (60)	27
Sublist 4 (60)	22
Sublist 5 (60)	20
Sublist 6 (60)	22
Sublist 7 (60)	22
Sublist 8 (60)	13
Sublist 9 (60)	12
Sublist 10 (30)	3

Table 4.11 shows that the largest number of AWL headwords from any of the AWL sublists spoken by the FIS teachers is AWL sublist 1, with a total of 44 headwords. Almost without fail, with each lower frequency AWL sublist, the number of academic words used by the FIS teachers becomes fewer and fewer. Thus we see that as many as 44 of the 60 word families in the highest frequency AWL sublist (sublist 1) are in the FIS corpus, while only 3 of the 30 word families in the lowest frequency AWL sublist (sublist 10) can be found in the FIS sublists. This similarity would be reassuring to the FIS high school teachers and students in that it shows that the higher frequency academic words at tertiary level are being used with greater frequency at FIS high school than the lower frequency academic words. If the FIS corpus were made larger by making more recordings, the data in Table 4.11 suggests that the frequency of the academic words used would continue to resemble even more closely that of the AWL.

4.4 Discussion

In this section I discuss the findings as presented in the first part of this chapter.

4.4.1 What do the findings suggest about the type of vocabulary students need in order to comprehend the words spoken by their teachers?

General English vocabulary

The findings of this study (as presented in Table 4.3) revealed that approximately 85% of teachers' speech is made up of the most common 1,000 English words. The most common 2,000 words make up about 91% of teachers' speech and the most common 3,000 make up 94%. Although these are quite high percentages of high frequency general English words in the somewhat formal and specialised setting of a high school classroom, it is still not necessarily enough to give most learners the amount required for comprehension of both spoken and written texts in an academic context. This is especially so when one considers that the mean score on the PVLТ of the ESL students at FIS was only 62.5% (refer to Table 4.1). (This mis-match between the ESL students and teacher discourse is discussed in more detail in section 4.4.4.). It is critical that students have thorough knowledge of general English vocabulary up to at least the 3,000 word level. If, to this level, one adds off-list words with a 'minimal learning burden', such as proper nouns, one will then reach 95% coverage of a text which is enough coverage for some people to be able to comprehend a text (Nation, 2006). Furthermore, it was discussed in section 2.2.2.6 that a vocabulary at the 3,000 word level enables one to be 'conversant in English' (Schmitt 2010, 7). Only one of the ESL students at FIS met this level (in the second round of testing in 2015), while all except one of the non-ESL students showed mastery of the 3,000 word level.

Table 4.2 also showed that the NGSЛ covers many more words in the FIS spoken corpus than in academic written texts. It was mentioned in section 2.2.2.9.2 that 2,000 high frequency English words cover 76% of the AWL. This is considerably lower than the 89% coverage in the FIS corpus. The 13% greater coverage could be a result of the differences between written and spoken discourse. Although the high school context is a formal one, a lot of general vocabulary would be used in order to convey a message more clearly, and unlike written texts (such as textbooks), 'space' is not an issue. Thus teachers can repeat what they say, backtrack and rephrase their message using general English vocabulary many times. Furthermore, Coxhead's (2000) AWL was designed to benefit university-going students; thus the corpus she used could be of a more advanced nature to Grade 9 and 10 teacher discourse.

Academic vocabulary

This study shows that high school students do require an understanding of academic vocabulary, just not to the extent required in understanding written academic texts. Coxhead (2000, 225) discovered that up to 10% of academic written texts comprise academic words. The FIS teachers on average only made use of about 2% of academic words in their discourse, although this depended on the subject area; History and English teachers made use of less academic vocabulary than Mathematics and Science teachers. Table 4.11 showed that most of the academic vocabulary spoken by teachers was ‘high-frequency’ academic vocabulary – words found in the AWL sublists 1-3. It is fair to say then that learners could benefit from learning higher frequency academic vocabulary, while learning the lower frequency academic words would perhaps not be as helpful. The reasonably high mean score (74%) of the non-ESL students on the UWL PVLIT, taken in 2015, shows that most of the non-ESL students probably meet the requirement of knowing the higher frequency academic vocabulary. Since many of these academic words would be covered by the mid-frequency words (K3-9) in the BNC/COCA list, it is suggested that if vocabulary from word lists is to be used by learners, then the BNC/COCA list (which covers high to low frequency vocabulary) would perhaps be better than the AWL. Alternatively the 106 common core academic verbs from the AWL as picked out by Granger and Paquot (2010), and discussed in section 2.2.3, could be worthwhile for students to learn, since the findings in Table 4.11 showed that academic words in the higher frequency sublists of the AWL, were more frequently used by the high school teachers than the words in the lower frequency sublists.

Technical Vocabulary

As is to be expected, in addition to general and academic vocabulary, teachers also use a proportion of technical vocabulary. A few of those technical words would fall in the category of low-frequency vocabulary (BNC/COCA). However since only 0.5% of the FIS corpus is made up of low-frequency words (K10-K25 words) it would not be worth the effort for students to study from lists of words beyond the 10,000 word level. Instead, important technical words crucial to understanding the specific topic being taught (that may be found in the mid to low-frequency bands (K3-25) of the BNC/COCA lists, or in the off-list category) is where (in terms of lower frequency words) students should focus their energies. These important technical words should not be too difficult to identify. They tend to be the highlighted words in text books, or the words in the glossary section at the end of textbook

chapters. In addition, content teachers tend to devote time to explaining these words, since due to their low frequency in general English usage, all students (ESL or L1) may be encountering them for the first time. Furthermore, since they are crucial to the topic at hand, they tend to be used frequently in class. Even though the meaning of these technical words would probably be addressed by subject teachers, ESL teachers would do well in paying attention to what they are, so that they can allow their ESL students to practice using those words during the ESL lessons.

4.4.2 What do the findings suggest about the nature of the high school sub corpora?

It was revealed in section 4.3.3 that the nature of the vocabulary spoken by English and History teachers is similar, but differs from vocabulary spoken by the Mathematics and Science teachers (who, in turn use similar vocabulary). English and History teachers used more vocabulary at the 1,000 word level than the Mathematics and Science teachers, while the Mathematics and Science teachers used more vocabulary from the 2,000, 3,000 and academic word levels. Both English and History are humanities or social science subjects, meaning that people and their relationships play an important role. This lends itself to a greater use of the everyday or more general vocabulary, found in the highest frequency band. Moreover, it can be argued that Science and Mathematics are more technical and specialised subjects, resulting in the greater use of words with lower frequencies than the 1,000 word level.

4.4.3 What do the findings suggest about the similarities and differences between the FIS corpus and corpora taken from universities?

From this chapter's findings one can conclude that high school teachers' speech is not necessarily hugely different from the speech found in academic situations at universities. However, it is more general, and less technical than that of university level written academic texts. From a pedagogical point of view, one can argue that if English language teachers wanted to prepare high school students for being able to understand the vocabulary spoken by their teachers, they should be wary of using material created exclusively from the NGSL. Although there is a high percentage of general vocabulary in the FIS corpus, there is not enough technical vocabulary contained therein to provide high school students with the necessary academic content. Similarly, it is clear from Table 4.7 that high school ESL teachers should avoid creating vocabulary learning material from university level academic

written texts, since, as evidenced by the BAWE list, it would include technical vocabulary of a level that is too advanced for ESL students.

On the other hand, with the similarity between the FIS corpus and the sample taken from MICASE, teachers could feel confident that pedagogical materials created from MICASE could be of value to high school students in helping them understand the vocabulary spoken by their high school teachers.

4.4.4 What do the findings of this study reveal about the vocabulary levels of the FIS students and vocabulary levels they should have in order to understand the words spoken by their teachers?

The low vocabulary levels of the FIS students

This study has shown that many of the high school students at FIS have low vocabulary levels which fall short of what is required to understand their Grade 9 and 10 teachers, particularly in the subjects of Mathematics and Science.

As revealed in Table 4.1, none of the ESL students who were tested in this study, were shown to be competent at the 3,000 word level. There was one student who had been in the ESL programme when the initial round of testing took place that did excel at the 3,000 word level, but by the time he showed mastery of the 3,000 word level he had exited the programme. All students who were still in the ESL programme during the time of the second round of testing in 2015, did not show adequate mastery of words at the 3,000 word level.

Why knowledge of 6,000 words should be the ultimate goal for the students at FIS

In a study conducted by Hu and Nation (2000 in Nation 2006, 61) it was found that ‘with a text coverage of 90%, a small minority (of participants) gained adequate comprehension (of a text). With text coverage of 95% ... a few more gained adequate comprehension’ and ‘it was calculated that 98% text coverage ... would be needed for most learners to gain adequate comprehension’. A student with a vocabulary of 3,000 might be able to comprehend their teachers, since some students in Hu and Nation’s (2000) study demonstrated comprehension with 95% knowledge of the words in a text, and Table 4.3 showed that 3,000 words covered 94.35% of the FIS corpus. If one adds the 1.7% off-list words that are a ‘minimal learning burden’ (Nation, 2006) to that percentage then one would gain 95% coverage. However, since most students gain comprehension of a text with 98% coverage (Nation, 2006), knowledge of 6,000 words should be the goal of high school students. Table 4.3 shows that 6,000 words

covers 96.37% of the FIS corpus but if one adds the 1.7 off-list words to that, then one will reach 98% coverage.

The results presented in this section reveal a large gap between the vocabulary levels of the FIS students, especially the ESL students, and the vocabulary level required for comprehension of their teachers' classroom talk. The implications of this gap is that a huge drive to improve the vocabulary levels of the FIS students needs to be put in place in order to bridge that gap. Ideally this drive would begin well before students reach Grade 9, so that by the time students enter Grade 9 they have been taught vocabulary up to the 6,000 word level. Of course if this programme were to begin in the elementary school, then such a goal can be attained; however, such is the nature of FIS that there is a strong chance students might arrive at the school for the first time in high school and have very low vocabulary levels. Therefore, if FIS were to be able to extend its ESL programme to Grade 11 (and possibly Grade 12) then those students with vocabulary levels below the 6,000 word level would have a better chance of closing the gap.

Support for extending the ESL programme into Grade 11 can also be found from the PVLТ results presented in Table 4.1. The ESL programme at FIS only runs from Grade 6 through Grade 10. Students in Grade 11 do not get ESL support. Table 4.1 shows that most of the Grade 9 ESL students (who took the first round of testing in 2014) improved their scores a year later when they were in Grade 10; however, most of the students who were in Grade 10 in 2014 did not improve on their scores when they were administered the PVLТs in 2015. Perhaps if the Grade 11 ESL students had continued to receive the more personalised attention and language learning focus provided by ESL teachers when they were in Grade 11 then they would have shown better vocabulary growth.

4.4.5 What factors in addition to vocabulary levels help or hinder students' comprehension of spoken discourse?

Of course there are factors, in addition to individual words, that help or hinder students' comprehension of spoken discourse. The pronunciation of speakers, the speakers' use of idiomatic phrases, lexical bundles, and the time given for processing information are some factors that may be involved. In addition, students' language plays a role, in that L1 students will find it easier to follow their teachers than ESL students, because their mother-tongue would be similar to that of the teachers. These factors, however, fall beyond the scope of this study and thus were not analysed. From a pedagogical stance, there is a lot that teachers can

do to aide students' comprehension. ESL teachers can expose their students to rich input, make many repetitions of new words, explicitly teach vocabulary and vocabulary learning strategies, have students do more reading and listening to boost vocabulary development and, finally, to have students practice using vocabulary and new sentence patterns in speaking and writing.

4.5 Conclusion

This chapter presented findings that addressed this study's aims which were first presented in Chapter 1 of this dissertation. Firstly the vocabulary levels of the participating Grade 9 and 10 students at the research school (FIS) were analysed followed by an analysis of the spoken discourse of the participating high school teachers. This chapter showed that most of the FIS students have vocabulary levels below the 5,000 word level, with the ESL students revealing particularly low levels of less than 3,000 words. This chapter also revealed that in order for most students to comprehend the teacher discourse in the classes they would need a vocabulary of 6,000 words. There is therefore a gap (and hence a challenging learning goal for students) between their current low vocabulary levels and that required in order to comprehend their teachers.

Another of this study's findings which was presented in this chapter was that up to 94% of teacher discourse is made up of high frequency/general service vocabulary, or the 3,000 most commonly spoken words in English and 2% of the teacher discourse was made up of academic words as found on the New Academic Word list. This chapter also revealed that when the sub-corpora of the FIS corpus were compared, the Science and Mathematics teachers made more use of academic vocabulary and lower frequency vocabulary than the English and History teachers. This chapter also compared the FIS corpora with samples of two other corpora – a written one (BAWE) and a spoken one (MICASE). As one could expect the FIS corpus made use of more high frequency vocabulary and less academic vocabulary than the other two corpora, with the written corpus making the most use of academic vocabulary. Despite these differences between the corpora, the differences are not so great as to suggest they are incompatible with each other. It is fair to say that the vocabulary profile of the high school teachers at FIS provides a sound preparatory foundation for the vocabulary students will encounter at university, particularly in its spoken form, as revealed by the comparison with the FIS corpus and MICASE. Finally this chapter analysed the nature of the academic words spoken by the FIS teachers. It was revealed that when the teachers did use

academic words they tended to be the higher frequency academic words in Coxhead's (2000) Academic Word List.

CHAPTER 5

CONCLUSION

5.1 Introduction

This chapter briefly summarises this study and its findings and discusses how it contributes to research in corpus linguistics, and specifically the contribution made to identifying and analysing the nature of the vocabulary spoken by high school teachers as they teach their subjects. This chapter also looks at the limitations and implications of the study and makes suggestions as to where future research in this field may lie.

5.2 Review

This study had two main aims. The first aim was to identify the vocabulary profile of high school teachers' discourse in the classroom of an international school where English is the language of learning and teaching. The second aim was to assess the vocabulary levels of the students who were being taught by the teachers whose vocabulary was analysed.

5.2.1 Summary of findings

I will discuss the main findings of the study by revisiting each research question and its subquestions one by one.

Research question 1

The first research question of this study asked: *What are the productive vocabulary levels of the Grade 9 and 10 high school students at an international high school?*

In order to identify the vocabulary levels of the Grade 9 and 10 students at FIS, all available students at these grades, had their vocabulary levels tested using Laufer and Nation's (1999) PVLT. These results were presented and analysed in section 4. 2. Table 4.1 presented the PVLT results of the ESL students and non-ESL students separately. The ESL students' vocabulary levels were shown to be particularly low, with their productive vocabulary levels below the 3,000 word level. The non-ESL students, however, showed competence at the 3,000 word level with a mean score of 93% in their second round of testing in 2015. Half of the non-ESL students showed mastery of the UWL level with scores of over 80% and a few (but not most) showed mastery at the 5,000 word level. These results suggest that most of the ESL students at FIS do not have sufficient vocabulary knowledge to comprehend what their

teachers are saying, and at least half of the non-ESL students do not have the vocabulary knowledge to fully comprehend a written academic text unassisted. As discussed in section 4.4.4, high school students should strive to have a vocabulary of 6,000 words because this study showed that 6,000 covered approximately 98% of the FIS corpus, and knowledge of 98% of the words in a text has been shown to be the desired coverage for comprehension of a text (Hu and Nation 2000). The fact that the majority of the ESL students at the research school do not have productive vocabulary levels of 2,000 words means that they will face challenges comprehending what their teachers are saying.

Research question 2

The second research question that this study addressed was: *What is the vocabulary profile of the spoken discourse of Grade 9 and 10 high school teachers at the research school?* This question was unpacked through the use of the following five subquestions:

Research question 2(a)

The first subquestion in identifying the vocabulary profile of high school teachers' spoken discourse was:

How much of the high school teachers' spoken discourse is made up of general and academic English?

This question was answered in terms of identifying what percentage of the teachers' discourse was made up of high frequency/general English vocabulary and what percentage was made up of academic vocabulary.

The results, as presented in section 4.3.2, show that high school teachers make use of a high percentage of high frequency general service vocabulary in their speech. In particular, 91% of their spoken discourse is made up of the most frequent 2,000 words as found in the BNC/COCA corpus. As presented in section 2.3.3, 91% is precisely the same percentage coverage of the high frequency 2,000 words contained in the GSL which make up the vocabulary contained in short novels (Hirsh 1992 in Waring and Nation 1997). This finding shows that much of the vocabulary contained in short novels is the same as that which is spoken by high school teachers.

Coxhead (2000) found that the GSL made up 76% of written academic texts. This percentage coverage is significantly lower than the 91% usage of GSL words by high school teachers. This, together with the fact that Coxhead's academic words (the AWL) covers 10% of

academic texts, while the NAWL covers only 2% of high school teacher discourse, shows that the profile of the research school teachers' spoken English is not nearly as academic in nature as the vocabulary used in academic written texts which are geared towards university students.

Research question 2(b)

How much of the high school spoken corpus is made up of high, mid and low-frequency vocabulary?

In addition to identifying how much high-frequency vocabulary is used by the high school teachers, this study also identified how much of the teachers' discourse was made up of mid and low frequency vocabulary. Table 4.3 revealed that 94.35% of teacher discourse is made up of the most common 3,000 words in English (3,000 words is what Schmitt and Schmitt (2014) claim should be considered high frequency vocabulary). Approximately 2.5% of the FIS corpus is mid-frequency vocabulary (words in the 3,000-9,000 range) and less than 2% is low-frequency vocabulary (words in the 9,000-25,000 word range).

Research question 2(c)

How do the high school spoken sub-corpora compare with one another in terms of their coverage of general and academic English?

In a comparison of the four sub-corpora – English, History, Mathematics and Science – it was found that English and History teachers made use of more words at the 1,000 word level than the Mathematics and Science teachers, and Mathematics and Science teachers made use of more words at the 2,000 and 3,000 word levels than the English and History teachers. The greater use of higher frequency vocabulary by the English and History teachers could well be because of the social nature of those subjects, which lends itself to a greater use of everyday more general vocabulary usage. It is not surprising then that the English and History teachers used almost half as much academic vocabulary as the Mathematics and Science teachers. The implication of this finding is that students with low vocabulary levels will be able to understand more of what their teachers are saying in English and History classes than in Mathematics and Science classes. It also points towards the difference in basic interpersonal communicative skills (BICS) and cognitive/academic language proficiency (CALP) required at school level, which was discussed in section 1.2. This study's findings suggest that vocabulary required for BICS plays a more central role in English and History, whereas

students tend to encounter more vocabulary required for CALP in the subjects of Mathematics and Science than they do in English and History.

Research question 2(d)

How does the nature of the words spoken by the high school teachers compare with the vocabulary profile found in other corpora?

To answer this question I analysed the vocabulary profile of the high school teachers by comparing the FIS corpus to two other academic corpora: the Michigan Corpus of Academic Spoken English (MICASE) and the British Academic Written English (BAWE) and the New General Service List (NGSL). Section 4.3.4 revealed that teachers use more general service vocabulary – particularly at the 1,000 word level – than MICASE and an even greater amount more than BAWE. This is unsurprising because both MICASE and BAWE are university-based corpora, as opposed to a high school one, and therefore one would expect high school teachers to make more use of lower frequency vocabulary or more academic vocabulary. Nevertheless with the FIS, MICASE and BAWE corpora all making use of a similar percentage of vocabulary at the 3,000 word level and academic vocabulary between 2 and 3.41%, one can argue that the FIS corpus provides a reasonable foundation for the vocabulary students can expect at university.

Research question 2(e)

What is the nature of the academic vocabulary spoken by the high school teachers?

It was not surprising to find that the nature of high school teachers' vocabulary was not as academic as found in academic corpora. This is because the academic word list (AWL) from which Coxhead (2000) analysed academic vocabulary was created from mostly written academic texts at the university level. It is interesting and encouraging to see that most of the academic vocabulary that the high school teachers did use was the academic vocabulary found in the most frequent academic word sublists of the AWL. Table 4.11 showed that for the most part with each AWL sublist containing more frequent academic vocabulary, the number of academic words that the FIS teachers used grew more. The pedagogical implications of this is that high school students can be confident that if they learn the words in the higher frequency academic vocabulary sublists of the AWL, they will be learning words that high school teachers use with relative frequency.

5.2.2 Contributions of the study

This study contributes to the field of corpus linguistics in that it created a corpus that identifies the vocabulary profile of a group of high school teachers' spoken discourse, a relatively under-researched area. It is easy to find and analyse written corpora, and MICASE is a good source of spoken corpora from a tertiary setting; however, very little research seems to have been conducted into the nature of the vocabulary spoken by high school teachers. As part of the analysis of teacher discourse in this study, it was possible to identify the vocabulary frequency levels of the participating teachers. Specifically it was found that 6,000 words, including off-list words, covered 98% of the teacher discourse (98% coverage of a text is considered to be the optimum coverage required for comprehension). This is a significant finding as there is little information in the current literature that informs us about how much vocabulary students require in order to comprehend the vocabulary used by high school teachers. Nation's (2006) study is helpful in identifying how much vocabulary is required to comprehend the vocabulary contained in newspapers, a variety of novels, graded readers, children's movies and general conversation. Nation (2006) found that 8,000-9,000 words are needed in order to comprehend a novel. With regards to general spoken discourse, Nation (2006) found that 6,000-7,000 words were required to understand talk-back radio and conversations between family and friends. However, Nation (2006) did not set out to identify the vocabulary required to comprehend teachers' spoken discourse at high school. By identifying the vocabulary required to comprehend teachers' spoken discourse, this study fills a gap in current research.

A complementary aspect of this study was the research conducted into the mismatch between ESL student and teacher vocabulary levels. Here too this study makes a valuable contribution. There is a growing presence of non-native speakers in schools around the world where English is the Language of learning and teaching. Many of these schools, like FIS, direct their curricula towards gearing students to widely recognized high school diplomas such as the International Baccalaureate and the Cambridge IGCSE and A levels qualifications. Because teachers aim to prepare their students for such examinations, it is reasonable to expect that they would use not only high-frequency vocabulary, but mid-frequency, academic and technical vocabulary that one might find in the lower frequency bands. This study, however, showed that the FIS students, especially the ESL students, do not have large enough vocabularies to comprehend their teachers, particularly when the teachers use mid-frequency vocabulary.

5.2.2.3 Pedagogical implications

The extent of the learning task

A contribution that this study makes is to reveal exactly what the vocabulary gap is between ESL students' vocabulary knowledge and the vocabulary levels used by teachers; and this gap is a large one. As more and more ESL students enter international schools, stakeholders need to be made aware, if the results from this study are anything to go by, that ESL students entering schools at the high school level could possibly possess as many as 4,000 words fewer than what they require to comprehend their teachers. Having said that, because this study has identified the precise vocabulary levels at which teachers speak across four subject areas, it is easier to establish exactly what vocabulary students need in order to gain more understanding from their lessons. Furthermore, the PVLTs used for this study are freely and easily available. Teachers could use them to test their students in order to identify the students' vocabulary levels and to then inform the students as to what their vocabulary levels are, so that each party knows the extent of the learning task.

Teaching approach to improving students' vocabulary levels

ESL students need to know that having a vocabulary of 6,000 words should be the goal and teachers need to put in considerable effort to helping their students attain these levels. Pikulski and Templeton (2004) provide a number of ideas that teachers can use in order to help students learn vocabulary: Firstly, they suggest that teachers and students read aloud in class and that teachers take the time to discuss the meanings of words in the texts. English teachers should also systematically teach the meaning of prefixes, suffixes and root words (morphological analysis). It is also suggested that teachers link spelling instruction to reading and vocabulary instruction, that word learning strategies, such as looking for contextual clues within the sentence, be encouraged, and that time is spent familiarising students with the use of dictionaries and thesauruses. Finally, Pikulski and Templeton (2004) emphasise the importance of wide independent reading for improving vocabulary.

Certainly, extensive reading can be a rich medium for vocabulary growth as is posited by Horst, Cobb and Meara (1998, 221) when they say that through extensive reading learners can 'enrich their knowledge of the words they already know, increase lexical access speeds, build network linkages between words, and a few words will be acquired'. While the majority of independent reading will be done outside the classroom, teachers can support students by

devoting a portion of each lesson, for example 10 minutes, on having each student read their own independent reading books silently in class. To monitor that students are reading at home and paying attention to vocabulary, teachers can encourage the use of reading journals where students are required to include a set number of journal entries that could also include glossaries of newly learnt words per week. Some studies suggest that Extensive reading in itself may not be a very effective method for learning new vocabulary incidentally (e.g. Waring & Takaki, 2003). This may be because often the focus from the reader is reading for pleasure or for text comprehension. However, there are also several researchers who argue strongly in favour of the value of extensive reading for building up vocabulary in the long run (Cunningham & Stanovich, 1998; Vidal, 2011). It is important for teachers to encourage students to incorporate vocabulary learning strategies, such as the aforementioned glossaries and morphological analysis, while they read. Students should also be motivated to explore multiple contexts, such as the internet – a multi-sensory source, in order to gain more exposure to vocabulary and increase the number of encounters with words.

There are many graded readers that extend up to the 6,000 word level. To take advantage of the various levels of graded readers, students' vocabulary levels should be tested, as previously suggested. Then students can systematically work through as many graded readers at the appropriate levels as possible. Provided they incorporate vocabulary learning strategies while they read, ESL students can build the number of words they know.

Materials development

ESL and English teachers could also create their own versions of the PVLТ. This would be simple to do. The various bands of 1,000 vocabulary levels are readily available from a number of sources on the internet. All teachers would need to do is access the vocabulary levels and then create modified cloze exercises using just the first 2 or 3 letters of the target word in a sentence that one might find in natural English usage.

It also seems imperative that ESL teachers try to improve the listening skills of their students. As was discussed in section 2.2.4, it is more difficult to learn new words through listening than it is through reading. However, much of students' school life is spent listening to teachers. Students' participation in class would certainly be greatly enhanced if they were better equipped to understand their teachers. One way to improve students' listening skills would be to create listening tasks from a spoken corpus, such as the one created for this study. There are any number of listening tasks one can assign, such as having students listen for

certain words or having them answer conventional comprehension style questions. ESL teachers could also help students cope with the speed of stream of speech by pausing and going through the text step by step to show how L1 speakers' words flow into one another. In this way the students might become better able to detect more individual words (and of course to identify collocations, and words that so often go together that they sound like one word). Another advantage of ESL teachers using a spoken corpus (and specifically a corpus of teachers' classroom discourse) is that students will have more opportunities to hear the same 'lecture', and by doing this they will have more repetitions of new words, making it more possible for them to increase their vocabulary, something that has tended to be difficult to do through a once off listening opportunity in a mainstream lesson.

Since there is a large gap between students' vocabulary levels and the levels that are required for comprehension of teachers, this study has shown that vocabulary learning should become a central role in the school curriculum. Particularly in schools with large numbers of ESL students, all teachers from all subject areas need to devote time and effort to building up the students' vocabulary levels. As students get closer to the 6,000 vocabulary level, they will find themselves comprehending their teachers more in class which will improve their motivation and ultimately their results. This will in turn please all stakeholders – not only the students and teachers, but parents and school administrators too.

5.2.3 Limitations of the study

Despite attempts to ensure rigor in the design and implementation of the study, the study has some limitations. One limitation is the small size of the FIS corpus. It consists of only 37,000 words. If more recordings and transcriptions had been made, then the results may be a better representation of high school teacher spoken discourse in general. Similarly, if more teachers had participated then it is possible that a greater range of vocabulary could have been included in the corpus. If the corpus were larger it is possible that different words could have shown themselves to be high-frequency words in the school context, and hence important for students to learn. If this study is to be replicated, I would also recommend that students' discourse be recorded and transcribed in addition to the teachers'. In this way a broader variety of linguistic analyses can be made, such as an analysis of turn-taking in the classroom.

Another limitation is in the administration of the PVLIT. Since a number of students failed to score well at the 2,000 word level, it may have been worthwhile to administer a 1,000 word level test. The reason why one was not administered was because Laufer and Nation (1999)

did not create a PVLТ at the 1,000 word level. However they did create a *receptive* test at the 1,000 word level. In situations where ESL students came into high school with limited language proficiency, an alternative approach would be to administer the 1,000 level receptive test and because students are for the most part listening to teachers when they speak, results on a receptive test could have reliably indicated to what extent the students comprehend the vocabulary in the FIS corpus.

The small number of participants who took the PVLТs is another limitation. Had more students been tested, perhaps from additional high school grade levels, or even different schools, the results could be more generalizable. Also this would have meant that the departure of some students from the school would not have had as much of an effect on the overall mean scores of the PVLТs taken. Ideally all participants would have completed all the PVLТs at the start of the study as well as towards the end. In this way the comprehensive results which such a longitudinal approach would have created could have given the teachers at the school, particularly the ESL teachers, more information on each student's vocabulary knowledge and growth, and in so doing they could have been in a better position to help the students improve their vocabulary. Of course students too would have benefited from seeing if and by how much their vocabulary levels had grown through the course of a year.

Nevertheless, fundamental to this study's aims, the way I administered the PVLТs did indicate that the participants' (especially those in the ESL programme) vocabulary levels were low and that for the ESL students it revealed what the gap is between their vocabulary levels and the vocabulary they require in order to comprehend their teachers' discourse.

5.3 Recommendations for further research

As stated in the previous section, there seems to be very little research into the vocabulary profile of high school teachers. For this study I created a small corpus from which I was able to identify what vocabulary teachers in Grade 9 and 10 were using. However, the corpus could be considered small with only six participating teachers and four subjects included as data sources. Future research could involve increasing and expanding this corpus by recording and transcribing more teachers, from more topics and subjects, from more high school grade levels and from more schools. The larger and the more comprehensive the corpus, the more opportunities ESL teachers will have to use the corpus to help teach useful and contextually relevant vocabulary and other aspects of language and subjects to their students.

A more comprehensive high school corpus will be particularly useful to ESL students who are studying at a school where English is the language of learning and teaching in a country where English is not widely used. This is because many ESL students in such circumstances have few opportunities to interact with English outside of their school environment. Like MICASE, a corpus from a high school context could include speech acts other than those used by teachers in the classroom, incorporating different types of school-based discourse, such as students talking at lunch, classroom discussions and notices read out at assemblies. Such naturally occurring speech acts often exclude students with low vocabulary levels which does not allow them to participate in school life as fully as they should. A larger corpus with a wider range of speech acts from a high school context could be studied by high school students with low vocabulary levels, giving them opportunities to learn the English they need to interact better with those around them and to improve their ability in their school work. Of course, simply having a corpus is not useful unless students and teachers know how to use it. The implication then would be for ESL teachers to find the most appropriate ways to use a high school corpus in order to aid the growth of their students' language skills. Teachers need to be the decision makers as they should be the ones who know best how to use corpora. Perhaps future research could also identify the most effective ways to use a corpus, like the FIS one, as a lesson tool. Could a transcription of teacher discourse be used like any other English comprehension text in the ESL classroom? Certainly, in using such a corpus, one would be sure of giving ESL students access to precisely the vocabulary used by their teachers – vocabulary (and other aspects of language) that they would be able to use in the specific subject areas as well as in general English usage.

5.4 Conclusion

This study has examined a relatively poorly studied area, namely the vocabulary profile of high school teachers and provided insight into the nature of the spoken English that high school teachers use in the classroom. This is an important area of study because, since students spend so much time in the classroom, it is critical that they understand what their teachers are saying.

The study also makes a contribution to ESL vocabulary development. By identifying the vocabulary levels of students, in particular ESL students, and the vocabulary levels of teachers as they speak in class, this study can give ESL teachers a clearer understanding of what and how much vocabulary they need to help their students learn in a particular subject.

Knowing what the gap is between learners' vocabulary and what is required for a particular subject will, in turn, help students to gain the most out of their lessons, and ultimately will help them succeed in high school.

REFERENCES

- Alejo, R., Adel, A., Kruis, J. & Swales, J. 2007. 'End up' in MICASE. <http://lw.lsa.umich.edu/eli/micase/Kibbitzer/endup.htm> (accessed 1 June 2013).
- Anderson, R. & Nagy, E. 1993. 'The vocabulary conundrum.' University of Illinois at Urbana-Champaign Library. *Technical Report No. 570*.
- Baker, M. 1988. Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, 4(2): 91-105.
- Bennett, G. 2010. Using corpora in the language learning classroom: Corpus linguistics for teachers. <http://www.press.umich.edu/titleDetailDesc.do?id=371534> Michigan ELT.
- Bereiter, C. & Scardamalia, M. 1989. Intentional learning as a goal of instruction. In *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Ed: L. B. Resnick, 361-392. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Biber, D. sept 2012. Corpus-based and corpus-driven analyses of language variation and use. *The Oxford Handbook of Linguistic Analysis* DOI: 10.1093/oxfordhb/9780199544004.013.0008
- Biber, D., Conrad, S. & Reppen, R. eds. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Brezina, V. & Gablasova, D. 2015. Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, 36 (1): 1-22.
- Brown, R., Waring, R. & Donkaewbua, S. 2008. Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20: 136-163.
- Browne, C., Culligan, B. & Phillips, J. 2013. *The new general service list*. <http://www.newgeneralservicelist.org> (accessed 20 December 2014).
- Calkins, G. 2005. *Applied Statistics – Lesson 5 Correlation Coefficients*. <http://www.andrews.edu/~calkins/math/edrm611/edrm05.htm>. (accessed 02 February 2016).
- Carder, M. 2011. ESL in International Schools in the IBMYP: the elephant under the table. *International Schools Journal*, 30(1): 50-58.

- Coady, J. 1993. Research on ESL/EFL vocabulary acquisition: Putting it in context. In *Second Language Reading and Vocabulary Learning*. Eds. T. Huckin, M. Haynes, and J. Coady, 3-23. Norwood, NJ: Ablex.
- Coady, J. 1997. L2 vocabulary acquisition: A synthesis of the research. In *Second Language Acquisition. A Rationale for Pedagogy*. Eds. J. Coady and T. Huckin, 273-290. Cambridge: Cambridge University Press.
- Cobb, T. 2014. *Web Vocabprofile*. <http://www.lex tutor.ca/vp/comp/> (accessed 8 January 2015).
- Cook, G. 2003. *Applied Linguistics*. Oxford: Oxford University Press.
- Conrad, S.M., & Biber, D. 2005. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, 20: 56-71.
- Corson, D. 1997. The learning and use of academic English words. *Language Learning*, 47(4): 671-718.
- Coxhead, A. 2000. A new academic word list. *TESOL Quarterly*, 34(2): 213-238.
- Crystal, D. 1995. Speaking of writing and writing of speaking. *Longman Language Review*. 1: 5-8
- Cummins, J. 2008. BICS and CALP: Empirical and theoretical status of the distinction, in: *Encyclopaedia of Language and Education, 2nd Edition, Volume 2: Literacy*. Eds. B. Street and N.H. Hornberger, 71-83. New York: Springer Science and Business Media LLC.
- Cunningham, A., & Stanovich, K. 1998. What reading does for the mind. *American Educator/ American Federation of Teachers*. Spring/Summer: 1-8.
- Davies, M. 2008. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at <http://corpus.byu.edu/coca/>
- Ebner, R., & Ehri, L. 2013. Vocabulary learning on the internet: using a structured think-aloud procedure. *Journal of Adolescent & Adult Literacy*, 56 (6): 480-489.
- Flowerdew, J. 1993. Concordancing as a tool in course design. *System*, 21(2): 231-244.
- Flowerdew, J. 2013. *Discourse in English language education*. London and New York: Routledge.
- Grabe, W. & Stoller, F.L. 2002. *Teaching and researching reading*. Harlow: Longman.

- Granger, S. & Paquot, M. 2010. *In search of a general academic vocabulary: A corpus-driven study*. http://www.uclouvain.be/cps/ucl/doc/adri/documents/In_search_of_a_general_academic_english.pdf. (accessed 1 March 2015).
- Horst, M. 2010. How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language*, 22(1): 161-180
- Horst, M., Cobb, T. & Meara, P. 1998. Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11, 207–223.
- Hu, M. & Nation, I.S.P. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1): 403-430.
- Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Keck, C.M. 2004. Corpus linguistics and language teaching research: Bridging the gap. *Language Teaching Research*, 8(1): 83-109.
- Kennedy, G. 1998. *An introduction to corpus linguistics*. London: Addison-Wesley Longman.
- Laufer, B. 1998. The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, 19(2): 255-271.
- Laufer, B. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1): 15-30.
- Laufer, B. & Hulstijn, J. 2001. Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1): 1-26.
- Laufer, B. & Nation, P. 1999. A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1): 33-51.
- Lee, D. 2001. Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradient of variation from the British National Corpus. *Journal of English Linguistics*, 29 (3): 250-278.
- Leech, G., Rayson, P. & Wilson, A. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.
- Lesaux, N., Kieffer, M., Faller, S. & Kelley, J. 2010. The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly*, 45(2): 196-228.

- Lewis, M. 1997. *Implementing the lexical approach: Putting theory into practice*. Hove: Language Teaching Publications.
- McLaughlin, M. & Rasinski, T. 2015. *Struggling Readers: Engaging and Teaching in Grades 3-8*. International Literacy Association.
- Nation, P. 1990. *Teaching and Learning Vocabulary* New York: Newbury House.
- Nation, P. 2001. Using small corpora to investigate learner needs. In *Small Corpus Studies and ELT Theory and practice*, eds. M. Ghadessy, A. Henry and R. Roseberry, 31-45. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Nation, P. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63(1): 59-82.
- Nesselhauf, N. 2005. *Corpus linguistics: A practical introduction*. <http://www.as.uni-heidelberg.de/personen/Nesselhauf>. (Accessed 16 July, 2014).
- Nunan, D. 1992. *Research Methods in Language Learning*. Cambridge: Cambridge University Press.
- Oxford, R. & Scarcella, R. 1994. Second language vocabulary learning among adults: state of the art in vocabulary instructions. *System*, 22(2): 231-243.
- Pica, T. 1994. Questions from the language classroom: Research perspectives. *TESOL Quarterly* 28(1): 49-79.
- Pikulski, J.J. & Templeton, S. 2004. *Teaching and developing vocabulary: Key to long-term reading success*. Boston, MA: Houghton & Mifflin.
- Römer, U. 2008. Corpora and language teaching. In *Corpus Linguistics: An International Handbook*, eds. A. Lüdeling and M. Kytö, 112-131. Berlin: Walter de Gruyter.
- 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, 589-619.
- Schmitt, N. 2010. *Researching Vocabulary*. DOI: 10.1057/9780230293977: Palgrave Macmillan.
- Schmitt, N. & Schmitt, D. 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4): 484-503.
- Scott, M. 2012. WordSmith Tools version 6. Liverpool: Lexical Analysis Software.

- Shin, D. and Nation, I.S.P. 2008. Beyond single words: the most frequent collocations in spoken English. *ELT Journal*. 62 (4): 339-348.
- Simpson, R.C., Briggs, S.L., Ovens, J. & Swales, J.M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan. <http://micase.elicorpora.info/about-micase> (accessed 25 May 2013).
- Sokmen, A. 1997. Current trends in teaching second language vocabulary. In *Vocabulary: Description, Acquisition and Pedagogy*, eds. N. Schmitt and M. McCarthy, 237-257. New York, NY: Cambridge University Press.
- Sutarsyah, C., Nation, I.S.P. & Kennedy, G. 1994. How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, 25 (2): 34-50.
- Staeher, L. 2009. Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31: 1-31.
- Stevenson, A. 2015. *Oxford dictionary of English*. Oxford university press. <http://www.oxforddictionaries.com/definition/english/vocabulary>. (accessed 12 May 2016).
- Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.
- Van Zeeland, H. & Schmitt, N. 2013. Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41: 609-624)
- Vidal, K. 2011. A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61(1): 219-258.
- Waring, R. & Nation, I.S.P. 1997. Vocabulary size, text coverage, and word lists. In *Vocabulary: Description, Acquisition and Pedagogy*, eds. N. Schmitt and M. McCarthy, 6-19. Cambridge: Cambridge University Press.
- Waring, R. & Takaki, M. 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a foreign language*, 15(2): 130-163).
- Wei, L. & Huo, Y. 2011. On the role of formulaic sequences in second language acquisition. *US-China Foreign Language*, 9(11), 708-713.
- West, M. 1953. *A General Service List of English Words*. London: Longman, Green and Co.

- Xue, G. & Nation, I.S.P. 1984. A university word list. *Language Learning and Communication*, 3(2): 215-229.
- Yali, G. 2010. L2 Vocabulary acquisition through reading – Incidental learning and intentional learning. *Chinese Journal of Applied Linguistics (Bimonthly)*, 33(1): 74-93.
- Zahar, R., Cobb, T. & Spada, N. 2001. Conditions of vocabulary acquisition. *The Canadian Modern Language Review* 57, 541-572.
- Zhang, M. 2013. A corpus-based comparative study of semi-technical and technical vocabulary. *The Asian ESP Journal*, 9(2): 148-172.
- Zhou, S. 2010. Comparing receptive and productive academic vocabulary knowledge of Chinese EFL learners. *Asian Social Science*, 6(10):14-19.

APPENDIX A

Productive vocabulary levels test versions A and B

Productive Vocabulary Levels Test Version A

Complete the underlined words as in the following example. He was riding a bi_____.

He was riding a bicycle.

THE 2,000 WORD LEVEL (version A)

- 1 They will restore the house to its orig_____ state.
- 2 My favourite spo_____ is football.
- 3 Each room has its own priv_____ bath and WC.
- 4 The tot_____ number of students at the university is 12,347.
- 5 They met to ele_____ a president.
- 6 Many companies were manufac_____ computers.
- 7 In AD 636 an Arab army won a famous vict_____ over another army.
- 8 The lakes become ice-free and the snow mel_____.
- 9 They managed to steal and hi_____ some knives.
- 10 I asked the group to inv_____ her to the party.
- 11 She shouted at him for spoi_____ her lovely evening.
- 12 You must spend less until your deb_____ are paid.
- 13 His mother looked at him with love and pri_____.
- 14 The wind roa_____ through the forest.
- 15 There was fle_____ and blood everywhere.
- 16 She earns a high sal_____ as a lawyer.
- 17 The sick child had a very high tempe_____.
- 18 The bir_____ of her first child was a difficult time for her.

THE 3,000 WORD LEVEL (version A)

- 1 They need to spend less on adminis and more on production.
- 2 He saw an ang from Heaven.
- 3 The entire he of goats was killed.
- 4 Two old men were sitting on a park ben and talking.
- 5 She always showed char towards those who needed help.
- 6 He had a big house in the Cape Prov.
- 7 Oh Harold darl, I am sorry. I did not mean to upset you.
- 8 Judy found herself listening to the last ec of her shoes on the hard floor.
- 9 He cut three large sli of bread.
- 10 He sat in the shade beneath the pa trees.
- 11 He had a crazy sch for perfecting the world.
- 12 They get a big thr out of car-racing.
- 13 At the beginning of their journey they encoun an English couple.
- 14 Nothing illus his selfishness more clearly than his behaviour to his wife.
- 15 He took the bag and tos it into the bushes.
- 16 Every year she looked forward to her ann holiday.
- 17 There is a defi date for the wedding.
- 18 His voice was loud and sav, and shocked them all to silence.

THE 5,000-WORD LEVEL (version A)

- 1 Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap strings.
- 2 After finishing his degree, he entered upon a new ph in his career.
- 3 The workmen cleaned up the me before they left.
- 4 On Sunday, in his last se in Church, the priest spoke against child abuse.
- 5 I saw them sitting on st at the bar drinking beer.
- 6 Her favorite musical instrument was a tru.
- 7 The building is heated by a modern heating appa.
- 8 He received many com on his dancing skill.
- 9 The government raised extra rev through tax..
- 10 At the bottom of a blackboard there is a le for chalk.
- 11 After falling off his bicycle, the boy was covered with bru.
- 12 The child was holding a doll in her arms and hu it.
- 13 We'll have to be inventive and de a scheme for earning more money.
- 14 The picture looks nice; the colours bl really well.
- 15 Nuts and vegetables are considered who food.
- 16 The garden was full of fra flowers.
- 17 Many people feel depressed and gl about the future of the mankind.
- 18 She ski happily down the path.

THE UNIVERSITY WORD LIST LEVEL (version A)

- 1 The afflu_____ of the western world contrasts with the poverty in other parts.
- 2 The book covers a series of isolated epis_____ from history.
- 3 Farmers are introducing innova_____ that increase the productivity per worker.
- 4 They are suffering from a vitamin defic_____.
- 5 There is a short term oscill_____ of the share index.
- 6 They had other means of acquiring wealth, pres_____, and power.
- 7 The parts were arranged in an arrow-head configu_____.
- 8 The learners were studying a long piece of written disco_____.
- 9 People have proposed all kinds of hypot_____ about what these things are.
- 10 The giver prefers to remain anony_____.
- 11 The elephant is indig_____ to India.
- 12 You'll need a mini_____ deposit of \$20,000.
- 13 Most towns have taken some eleme_____ civil defence precautions.
- 14 The presentation was a series of sta_____ images.
- 15 This action was necessary for the uli_____ success of the revolution.
- 16 He had been expe_____ from school for stealing.
- 17 The lack of money depressed and frust_____ him.
- 18 The money from fruit-picking was a suppl_____ to their regular income.

THE 10,000-WORD LEVEL (version A)

- 1 He wasn't serious about art. He just da in it.
- 2 Her parents will never acq to such an unsuitable marriage.
- 3 Pack the dresses so that they won't cre.
- 4 Traditionally, men were expected to nu women and children.
- 5 Religious people would never bl against God.
- 6 The car sk on the wet road.
- 7 The politician delivered an arrogant and pom speech.
- 8 The Romans used to hire au troops to help them in their battles.
- 9 At the funeral, the family felt depressed and mo.
- 10 His pu little arms and legs looked pathetic.
- 11 A vol person will change moods easily.
- 12 The debate was so long and tedious that it seemed int.
- 13 Drink it all and leave only the dre.
- 14 A hungry dog will sa at the smell of food.
- 15 The girl's clothes and shoes were piled up in a ju on the floor.
- 16 Some monks live apart from society in total sec.
- 17 The enemy suffered heavy cas in the battle.
- 18 When the Xmas celebrations and rev ended, there were plenty of drunk people everywhere.

Productive Vocabulary Levels Test Version B

THE 2,000-WORD LEVEL (version B)

Complete the underlined words as in the following example. He was riding a bi_____.

He was riding a bicycle.

1. It is the de_____ that counts, not the thought.
2. Plants receive water from the soil through their ro_____.
3. The nu_____ was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wa_____.
5. This year long sk_____ are fashionable again.
6. Laws are based upon the principle of jus_____.
7. He is walking on the ti_____ of his toes.
8. The mechanic had to replace the mo_____ of the car.
9. There is a co_____ of the original report in the file.
10. They had to cl_____ a steep mountain to reach the cabin.
11. The doctor ex_____ the patient thoroughly.
12. The house was su_____ by a big garden.
13. The railway con_____ London with its suburbs.
14. She wan_____ aimlessly in the street.
15. The organisers li_____ the number of participants to fifty.
16. This work is not up to your usu_____ standard.
17. They sat down to eat even though they were not hu_____.
18. You must have been very br_____ to participate in such a dangerous operation.

THE 3,000-WORD LEVEL (version B)

1. Despite the humiliation he conducted himself with dig_____.
2. Having defeated the competitors, the winner became the new tennis cha_____.
3. A new exhibition opened at the mu_____.
4. I have no id_____ what you mean.
5. He hid the coin in the pa_____ of his hand.
6. Boxers are not allowed to hit below the be_____.
7. The family spent the hol_____ in the mountains.
8. One of her mer_____ as a teacher is her ability to explain well.
9. On cold nights we cover ourselves with heavy bla_____.
10. The child pee----- at the stranger curiously.
11. He qu_____ his job when he found a better one.
12. She scr_____ all through the horror film.
13. Some people choose to dw_____ in exotic places.
14. She aba_____ her family and went to live with another man.
15. The police pur_____ the driver of the stolen car for two hours.
16. His behaviour seemd a little o_____.
17. The view from the top of the mountain was mag_____.
18. We could barely see each other in the di_____ light.

THE 5,000-WORD LEVEL (version B)

1. Henry's hobby is chess. His ze_____ for it is quite remarkable.
2. They took a second mort_____ on the house.
3. The road is gr_____ but they'll pave it soon.
4. His writing was childish with large lo_____ on the letters.
5. After a long climb we reached the su_____.
6. The Christian e_____ is counted from the birth of Christ.
7. She filled the pa_____ with water.
8. Tired of the typewriter they appreciated the nov_____ of the computer.
9. The colorful bal_____ flew away when the child let go of it.
10. If you prov_____ the dog too much it will bite you.
11. He is so depressed that he is cont_____ suicide.
12. She was unconscious and the doctors could not rev_____ her.
13. The economic problems mani_____ themselves in inflation and unemployment.
14. Astrologers try to pred_____ the future from the position of the stars.
15. Slavery in America was abo_____ long time ago.
16. The old lady is too ill and fra_____ to live alone.
17. All the residents of the city must pay muni_____ tax.
18. The response to our appeal exceeded all our expectations. It was incr_____.

THE UNIVERSITY WORD LIST LEVEL (version B)

1. The Far East is one of the most populated reg_____ of the world.
2. She spent her childhood in Europe and most of her ad_____ life in Asia.
3. Many people get bored with the endless cyc_____ of getting up, going to work and coming home, day after day.
4. Even though I don't usually side with you, in this ins_____ I must admit that you're right.
5. The parents gave their con_____ to their daughter's marriage.
6. I like mathematics in general and geo_____ in particular.
7. The resignation of the president upset the country's economic equil_____.
8. You usually pay less if you buy in bu_____.
9. He backed up his assertions by quoting the latest research and stat_____.
10. The plaster on the wall was removed to exp_____ the original bricks underneath.
11. His new book will be pub_____ at the end of this month by a famous University Press.
12. It is not easy to abs_____ all this information in such a short time.
13. The students were inhi_____ by the presence of the Dean of the Faculty.
14. Judging by the political changes in the world, our economic scheme will have to be rev_____.
15. If I were you I would con_____ a good lawyer before taking action.
16. There is still no val_____ data that supports your theory.
17. A civ_____ centre was built for community activities.
18. She didn't openly attack the plan, but her opposition was imp_____ in her attitude.

THE 10,000-WORD LEVEL (version B)

1. The beer is kept in small ke_____ in the cellar.
2. There are alab_____ statues in front of the palace.
3. The carpenter used a ra_____ to smooth the edges.
4. The rock-star's stage ant_____ included smashing guitars.
5. The children ate all the cookies. I'll make a new bat_____.
6. He is an art conn_____. People ask his advice before buying paintings.
7. He claimed to have seen a strange and frightening appa_____ last night.
8. Bo_____ is his hobby. He has always liked trees and flowers.
9. The rain left many small pu_____ in the street.
10. He drew attention to himself by swa_____ arrogantly into the room.
11. He blu_____ out the secret and regretted it instantly.
12. The murderer stra_____ his victim mercilessly.
13. His handwriting is horrible. Nobody can understand what he scra_____.
14. The woman's voice qui_____ with emotion.
15. She twir_____ from one dancing partner to another.
16. The teacher punished the student for his imper_____ behaviour.
17. We cycled through prime_____ forests.
18. A mot_____ collection of costumes dazzled everyone at the carnival.

APPENDIX B

PVLT memoranda

PVLT Version A memorandum

THE 2,000 WORD LEVEL (version A memo)

- 1 They will restore the house to its original state.
- 2 My favourite sport is football.
- 3 Each room has its own private bath and WC.
- 4 The total number of students at the university is 12,347.
- 5 They met to elect a president. (not accepted: *elective*)
- 6 Many companies were manufacturing computers. (accepted: *manufacture*)
- 7 In AD 636 an Arab army won a famous victory over another army.
- 8 The lakes become ice-free and the snow melts.
- 9 They managed to steal and hide some knives.
- 10 I asked the group to invite her to the party.
- 11 She shouted at him for spoiling her lovely evening. (accepted: *spoil*)
- 12 You must spend less until your debts are paid.
- 13 His mother looked at him with love and pride.
- 14 The wind roared through the forest. (not accepted: *road*)
- 15 There was flesh and blood everywhere.
- 16 She earns a high salary as a lawyer.
- 17 The sick child had a very high temperature.
- 18 The birth of her first child was a difficult time for her.

THE 3,000 WORD LEVEL (version A memo)

- 1 They need to spend less on administration and more on production.
- 2 He saw an angel from Heaven.
- 3 The entire herd of goats was killed.
- 4 Two old men were sitting on a park bench and talking.
- 5 She always showed charity towards those who needed help.
- 6 He had a big house in the Cape Province.
- 7 Oh Harold darling, I am sorry. I did not mean to upset you.
- 8 Judy found herself listening to the last echoes of her shoes on the hard floor. (accepted: *ecos*)
- 9 He cut three large slices of bread.
- 10 He sat in the shade beneath the palm trees.
- 11 He had a crazy scheme for perfecting the world.
- 12 They get a big thrill out of car-racing.
- 13 At the beginning of their journey they encountered an English couple.
- 14 Nothing illustrates(d) his selfishness more clearly than his behaviour to his wife.
- 15 He took the bag and tossed it into the bushes.
- 16 Every year she looked forward to her annual holiday. (not accepted: *annevesairy*)
- 17 There is a definite date for the wedding.
- 18 His voice was loud and savage, and shocked them all to silence.

THE 5,000-WORD LEVEL (version A memo)

- 1 Some people find it difficult to become independent. Instead they prefer to be tied to their mother's apron strings.
- 2 After finishing his degree, he entered upon a new phase in his career.
- 3 The workmen cleaned up the mess before they left.
- 4 On Sunday, in his last sermon in Church, the priest spoke against child abuse.
- 5 I saw them sitting on stools at the bar drinking beer.
- 6 Her favorite musical instrument was a trumpet.
- 7 The building is heated by a modern heating apparatus.
- 8 He received many compliments on his dancing skill.
- 9 The government raised extra revenue through tax..
- 10 At the bottom of a blackboard there is a ledge for chalk.
- 11 After falling off his bicycle, the boy was covered with bruises.
- 12 The child was holding a doll in her arms and hugging it.
- 13 We'll have to be inventive and devise a scheme for earning more money.
- 14 The picture looks nice; the colours blend really well.
- 15 Nuts and vegetables are considered wholesome food.
- 16 The garden was full of fragrant flowers. (Accepted: fragile (BNC K-10))
- 17 Many people feel depressed and gloomy about the future of the mankind.
- 18 She skipped happily down the path.

THE UNIVERSITY WORD LIST LEVEL (version A memo)

- 1 The affluance of the western world contrasts with the poverty in other parts.
- 2 The book covers a series of isolated episodes from history.
- 3 Farmers are introducing innovations that increase the productivity per worker.
- 4 They are suffering from a vitamin deficiency.
- 5 There is a short term oscillation of the share index.
- 6 They had other means of acquiring wealth, prestige, and power.
- 7 The parts were arranged in an arrow-head configuration.
- 8 The learners were studying a long piece of written discourse.
- 9 People have proposed all kinds of hypotheses about what these things are.
- 10 The giver prefers to remain anonymous.
- 11 The elephant is indiginous to India.
- 12 You'll need a minimum deposit of \$20,000.
- 13 Most towns have taken some elemental civil defense precautions.
- 14 The presentation was a series of static images.
- 15 This action was necessary for the ultimate success of the revolution.
- 16 He had been expelled from school for stealing.
- 17 The lack of money depressed and frustrated him.
- 18 The money from fruit-picking was a supplement to their regular income.

PVLT Version B memorandum

THE 2,000-WORD LEVEL (version B memo)

1. It is the delivery_____ that counts, not the thought.
2. Plants receive water from the soil through their roots_____.
3. The nurse_____ was helping the doctor in the operation room.
4. Since he is unskilled, he earns low wages_____.
5. This year long skirts_____ are fashionable again.
6. Laws are based upon the principle of justice_____.
7. He is walking on the tips_____ of his toes.
8. The mechanic had to replace the motor_____ of the car.
9. There is a copy_____ of the original report in the file.
10. They had to climb_____ a steep mountain to reach the cabin.
11. The doctor examined_____ the patient thoroughly.
12. The house was surrounded_____ by a big garden.
13. The railway connects_____ London with its suburbs.
14. She wandered_____ aimlessly in the street.
15. The organisers limitted_____ the number of participants to fifty.
16. This work is not up to your usual_____ standard.
17. They sat down to eat even though they were not hungry_____.
18. You must have been very brave_____ to participate in such a dangerous operation.

THE 3,000-WORD LEVEL (version B memo)

1. Despite the humiliation he conducted himself with dig_____.
2. Having defeated the competitors, the winner became the new tennis champion_____.
3. A new exhibition opened at the museum_____.
4. I have no idea_____ what you mean.
5. He hid the coin in the palm_____ of his hand.
6. Boxers are not allowed to hit below the belt_____.
7. The family spent the holiday_____ in the mountains.
8. One of her merits_____ as a teacher is her ability to explain well.
9. On cold nights we cover ourselves with heavy blankets_____.
10. The child peeped_____ at the stranger curiously.
11. He quit_____ his job when he found a better one.
12. She screamed_____ all through the horror film.
13. Some people choose to dwell_____ in exotic places.
14. She abandoned_____ her family and went to live with another man.
15. The police pursued_____ the driver of the stolen car for two hours.
16. His behaviour seemed a little obscure_____. (accepted: *off, odd, obstinate* not accepted: *obvious*)
17. The view from the top of the mountain was magic_____. (accepted: *magnificent* (BNC K-4))
18. We could barely see each other in the disappearing_____ light. (accepted: *dim* (BNC K-4) and *dismal* (BNC K-6))

THE 5,000-WORD LEVEL (version B memo)

1. Henry's hobby is chess. His zeal_____ for it is quite remarkable.
2. They took a second mortgage_____ on the house.
3. The road is gravel_____ but they'll pave it soon.
4. His writing was childish with large loops_____ on the letters.
5. After a long climb we reached the summit_____.
6. The Christian era_____ is counted from the birth of Christ.
7. She filled the pan_____ with water.
8. Tired of the typewriter they appreciated the novelty_____ of the computer.
9. The colorful balloon_____ flew away when the child let go of it.
10. If you provoke_____ the dog too much it will bite you.
11. He is so depressed that he is contemplating_____ suicide.
12. She was unconscious and the doctors could not revive_____ her.
13. The economic problems manifested_____ themselves in inflation and unemployment.
14. Astrologers try to predict_____ the future from the position of the stars.
15. Slavery in America was abolished_____ long time ago.
16. The old lady is too ill and fragile_____ to live alone. (accepted: *frail* (BNC K-9))
17. All the residents of the city must pay municipal_____ tax.
18. The response to our appeal exceeded all our expectations. It was incredible_____.

THE UNIVERSITY WORD LIST LEVEL (version B memo)

1. The Far East is one of the most populated regions of the world.
2. She spent her childhood in Europe and most of her adult life in Asia.
3. Many people get bored with the endless cycle of getting up, going to work and coming home, day after day.
4. Even though I don't usually side with you, in this instance I must admit that you're right.
5. The parents gave their consent to their daughter's marriage.
6. I like mathematics in general and geometry in particular.
7. The resignation of the president upset the country's economic equilibrium.
8. You usually pay less if you buy in bulk.
9. He backed up his assertions by quoting the latest research and statistics.
10. The plaster on the wall was removed to expose the original bricks underneath.
11. His new book will be published at the end of this month by a famous University Press.
12. It is not easy to absorb all this information in such a short time.
13. The students were inhibited by the presence of the Dean of the Faculty.
14. Judging by the political changes in the world, our economic scheme will have to be revised.
15. If I were you I would contact a good lawyer before taking action.
16. There is still no valid data that supports your theory.
17. A civic centre was built for community activities.
18. She didn't openly attack the plan, but her opposition was implied in her attitude.

APPENDIX C

Samples of transcriptions of all four subjects recorded for the study

Sample of Science transcription:

Any of these things not present you will have fire. You know you have a big. You know what it contains. Carbon Dioxide is very cold. Yes, when you put that on fire okay it takes up the heat and also it covers it and so it does two things okay. It removes the heat and. So this is the basic principal of the fire extinguisher. Yes inside the cylinder you can make. But when it comes out. You know have you ever heard the process of sublimation. Have you heard of melting, boiling? What is melting? Solid to liquid. And what is boiling? To gas. Solid to gas directly. Sublimation. Yes, the dry ice thing. We call it sublimation. Okay. So it is a process of changes made of from solid to gas directly. Okay. So if it is liquid to solid it would be freezing. Good. So the main idea I guess you got it. We are going to go through it quickly.

Heat and light are given off. Many substances can act as fuel. You know that. Diesel oil. Kerosene. And you have many others. Okay. Combustion is a reaction of one substance with oxygen around it. So once this reaction occurs. So there are three necessary components of combustion which is oxygen fuel and heat so it's important to know about this triangle and you will have questions later on when you will be asked to draw the fire triangle. Yes. You need to heat it up to a certain temperature. Yes but when you rub something like friction you produce heat once it's started it initiates the whole process because once it starts burning it's going to produce heat and this heat is going to be used for further burning. Okay. So you need this initial process which is actually friction you need. And you produce fire. Got it.

Sample of Mathematics transcription:

So basically today we're going to continue with the quadratic equation chapter. There are several types of quadratic equations. We're going to start with the easiest one of them all. Last lesson I reminded you of how to factorise quadratic expressions but right now we don't need to use that technique. We just going to solve equations by re-arranging just like we did for the formulae chapter. For example if I tell you X squared is equal to forty nine. That's an equation. Why is it an equation because it has what an equal sign. This is not a linear equation because X is squared so it's a quadratic equation because that's the maximum power of the polynomial. How do we solve that? Exactly. You square root the forty now or basically you square root both sides but we're over that. Any situations here? Is that it? is that the solution to X squared equals forty nine. What's the problem? So you're saying there should be two answers positive and negative seven. Why? Anyone have an idea why this equation had two answers and not just one. Correct. Because of the rule that positive times positive is positive and negative times negative is positive. If we transfer this mathematical equation in English as a question, they would be asking which number when you multiply it by itself give forty nine and because of student X there said we have that the number seven when you square it you get forty nine but the number minus seven when you square it you also get forty nine. Now that's the English answer to why we have the plus and minus. There is also a nice

mathematical explanation which has to do with the difference of squares but we not going to do that yet. We're just going to remember now when you have an equation that's X squared equal to a number you just have to square root the other side but remember there are two answers the positive and the negative so that gave two answers. If you don't like writing plus and minus seven you can write X equals seven or X equals minus seven. Please don't put and like a lot of people do. X cannot be seven and minus seven at the same time. All right we not talking about deities here. It can either be one number or the other. Do we agree. How do we solve X squared equals one. X is equal to what. thank you you are paying attention to what I said ten seconds ago. Plus minus the square root of one and what is the square root of one? One. Two answers. Do not forget the plus or minus. There's actually two answers to that equation. Positive and negative one. Good. Done. It gave the same as the original well that's the good thing about one. It's its own square root. How about this one. But to get rid of the one word answer that I twenty seconds ago that I said I did not like we would say plus minus the square root of zero which is kind of dumb because the square root of zero is zero and plus or minus zero is zero so yes student Y over there is correct. This weird question only has one answer. Zero. How about this one? Z equals minus four. Sorry I messed that up Z squared equals minus four. How do we solve that? No why are you being so creative. C equals plus or minus what? the square root of negative four. Do we have a problem? We cannot square root a negative. It is impossible in the real numbers and until you get to eleventh grade higher level mathematics the real numbers is all you're going to see. So what would be the solution to Z squared minus four? We're in the real numbers don't get ahead of yourself. Anyone. We're trying to go forward with the whole idea that we cannot square root minus four. It is impossible for us to square root minus four. Hence and therefore the solution to that equation. Thank you. It has no solution. Why would we randomly make up a number if we just said that it is impossible to get the procedure we need to use so it doesn't have a you might have equations that do not have a solution in the real numbers. Yes thank you. Okay. Anyhow. Forget about that we are not doing complex numbers. Okay so any questions so far on how to solve quadratic equations of the form X squared equals a number. Okay now let's make it a little more complicated it's still going to be the same type of end result where you end up square rooting plus or minus but it's not going to be as easy. For example two X squared equals fifty. Oh my god what do we do now. So first we divide by two just like said now we get that is correct twenty five and then right there and we just do the final technique plus or minus square root of twenty five giving us. Good. So it's not the end of the world. It can get ever so slightly more complicated like so. Three X squared plus seven equals eighty eight. Messed it up. So it's normal re-arranging of functions and the final technique is the same.

Sample of History transcription:

That Japan will give support give aid to Americans that are shipwrecked. What else? Small US legation. What's that? Not a port but it's kind of like a special trade embassy but it's not as big as an embassy right. you know how embassies in other countries and that's basically how countries relate to one another through the embassy. well it's just for the purposes of trade and it's just a small group of Americans that are given a presence in Japan. Was there anything else about that treaty? Was that it? Just America. So after Kanagawa who else signs

treaties. How are the Americans considered to the other? What does that mean - most favoured nation status. So if Japan decides to give benefits to other countries that it hasn't given America, America automatically gets those benefits. Now just as a little aside. This week Australia signed a free trade agreement with Japan and in the articles I've been reading, certainly the Australian press says this is the best deal that any country has received from Japan. the deal that Australia gets to sell beef and other agricultural products into Japan. And I read in one of the articles that Australia has most favoured trading nation status so if Japan signs another agreement with say America and give something else then Australia will automatically get that deal. that was just this week. Anyway back in 1854 when the treaty of Kanagawa is signed America is the first nation to ever have most favoured nation status with Japan. What else did we talk about on Monday? Right it was called and all the other treaties they were called unequal treaties why were they unequal. why were they unequal. one side is giving a lot not receiving much. who is that? which side is giving up lots but not getting much in return. Japan. What did the Americans promise not to do? Because Japan signed this treaty America made promises that they will not do certain things. basically they not going to sell opium into Japan like the British had done in China. That's pretty much what Japan gets out of this. And what else? sorry. right and basically it's a way of avoiding war with the Americans. is there anything else we did before the lesson ended? Right tell me about him. Famous teacher about what? he had been influenced by renraku. what is renraku? Dutch studies. Okay so he was interested in seeing how Japan could be modernised. In fact what's he called or revered as. one of the fathers of modern Japan. during this period of the bakumatsu he wants to put Japan on a modern course. what else was he influenced by? Just Dutch studies. what's kokugaku? yes and what was that about. so it was like reinterpreting Japanese ancient literature and history as being unique special to Japan. it didn't come from china it wasn't influenced by anyone else it was specially Japanese and this school called kokugaku also was really the movement that began the feeling of Japanese nationalism pride in their own country that they are somehow special unique. what happened next? what's sonno joi. What is it actually? animal vegetable or mineral? no that's just a joke. it's a movement right. it's a movement people that believe in something attracting support and it grows and it grows during this period of the bakumatsu. what is the movement for? expel the foreigners, worship the emperor. okay. sonno joi. I don't know how you get such a long message from such a short word. but apparently you do. let's continue. remember I also introduced a word to you after two hundred and fifty years being told all sorts of things about foreigners it was no surprise really that Japan people were xenophobic. which means fear of other races. doesn't surprise you. the same can be said of Europeans. even though they did have experience with other races. usually this comes about where you have no contact. right. you start to believe all sorts of myths all sorts of stories about people that are different to you. that's why it's very very important for societies to make sure that different cultures and different social classes are able to mix able to deal with each other. can we remember what extraterritoriality was? because this is another issue that really causes problems in this time for the shogun. okay so this is any foreigners coming into Japan they are not subject to Japanese law they are subject to the law of their country. today this no longer exists so for example if you are an Australian tourist and you are found with drugs which is a very famous tourist destination in Indonesia, you will be put in prison and you will face the death sentence

because you are under Indonesian law not under Australian law. in the embassy? that's different because the land of the embassy around the world is considered the sovereign territory of that country. okay. so and diplomats have what they call diplomatic immunity. so the Korean embassy the one that's opposite the dome inside of that embassy is Korean law okay those people are subject to Korean law not Japanese law if there is a difference they would be tried under Korean law.

Sample of English transcription:

Alright so your exam question is going to be on Lord of the Flies and your exam is on Wednesday is that right? so that gives you a week to get your head around Lord of the Flies. so we got today to work on it we got Monday and then on Tuesday with Ms so that's all going to be revision. you'll have a choice of questions they are going to be essay questions. Did Ms talk a little bit about it? so expect to get a question on conflict. start that off with characters because conflict is between characters. expect to get a choice and it might be connected you're not sure how they going to be connected but theme as well. now maybe it's not separate questions maybe you'll get a question like how do certain characters reveal a particular theme. so in that way characters and theme might be connected. and of course even if it is just about a particular theme you would need to give examples and the examples might come from the characters actions, right. I think those are the two main areas that you need to focus on. so you need to have a good understanding of the characters, conflict which is relationship so which characters have a relationship of conflict. What about theme? What would be probably the main theme of Lord of the Flies. what's the main idea of this story? why did Golding this novel? what was he trying to what message was he trying to get out? right. correct. so evil is. let me give you a word here. innate in mankind humankind. innate? do you know the word. it's inside of us. it's natural. we are born with it. it's not something we learn from other people. that was the whole idea behind the experiment. that's why Golding put these children on an island. if he'd put adults on the island we could maybe say but they learned the evil way while they were growing up in civilization. so he wanted to have these innocent boys on an island with no adults to steer them in the wrong direction so he wanted to show that innocent children will have this potential to do evil. it's something they are born with. just because evil is innate in people does that mean everybody has to become evil. no. obviously there's also good in people as well. what's that? Simon, exactly so that's a good example so we've got this character Simon who never does anything wrong. there's Ralph who is definitely on the good side but he slips into doing bad things but Simon is a character that shows us that yes even though you might be born with some evil inside you it's possible to keep it away for much of your life. so there's a little bit of positive in the book so you don't need to leave the book being completed depressed that you are going to do evil things. okay. this is kind of connected with savagery versus civilisation and the evil in this book is connected with the savagery aspect.

APPENDIX D

Signed consent forms

Request for permission to conduct research at Fikaoka International School

To: Head of School at Fikaoka International School

I am a registered MA student at the University of South Africa (student no: 30059615) and I would like your consent for me to conduct classroom-based research in your school as part of my Master's degree.

The title of this research is: *A corpus linguistics analysis of high school teachers' spoken discourse in the classroom and learners' lexical profiles.*

The reason why I am conducting this research is to study the nature of high school teachers' content subject discourse in the classroom as well as the lexical profile of learners. In order to better understand the teachers' discourse I will be collecting data by recording teachers' spoken use of English while they are teaching. At the same time I will be analysing the lexical profile of learners in order to assess to what extent the learners' knowledge of English is similar or different to the discourse spoken by teachers.

The research is non-invasive and does not put participants at risk. The research involves teachers' voices recorded during classroom lessons at various times during the school year. Recordings will be taken from four subjects and a total of eight teachers will be recorded. In addition to teachers being recorded, I will also assess the learners' knowledge of words at different frequency levels.

I hope that these arrangements will meet with your approval. If you have any further queries please feel free to contact my supervisors Prof Lilli Pretorius (pretobj@unisa.ac.za) and Dr Nanda Klapwijk (klapwn1@unisa.ac.za).

Thank you for your consideration.

Yours faithfully

Graham Creighton

By signing below, I give consent for the above-referenced research to be conducted at Fikaoka International School

Name of Head of School: _____ Date: _____

Signature of Head of School: _____

Please tick this box if you would like to receive a summary of the results by e-mail:

Participant Consent Form
BACKGROUND INFORMATION

Researcher: My name is Graham Creighton and I am a registered MA student from the University of South Africa, Department of Applied Linguistics.

Title of research: The title of this research is: *A corpus linguistics analysis of high school teachers' spoken discourse in the classroom and learners' lexical profiles.*

Reason for the research: I am studying the nature of high school teachers' content subject discourse in the classroom as well as the lexical profile of learners. In order to better understand the teachers' discourse I will be collecting data by recording teachers' spoken use of English while they are teaching. At the same time I will be analysing the lexical profile of learners in order to assess to what extent the learners' knowledge of English is similar or different to the discourse spoken by teachers.

Details of participation: The research is non-invasive and does not put participants at risk. The research involves teachers' voices recorded during classroom lessons at various times during the school year. Recordings will be taken from four subjects and a total of eight teachers will be recorded. In addition to teachers being recorded, I will also assess the learners' knowledge of words at different frequency levels. Please feel free to ask questions now if you have any.

CONSENT STATEMENT

1. I understand that my participation is voluntary and that I may withdraw from the research at any time without giving any reason.
2. I am aware of what my participation will involve.
3. I understand that there are no risks involved in the participation of this study.
4. I understand that all data will be kept confidential and anonymous.
5. All questions that I have about the research have been satisfactorily answered.

In light of the above conditions, I agree to participate.

Participant's signature: _____ Date: _____

Participant's name (print): _____

Participant's E-mail address: _____

Researcher's signature: _____ Date: _____

Researcher's name (print): _____

Researcher's E-mail address: _____

Tick this box if you would like to receive a summary of the results by e-mail:

Student Assent Form

Fukuoka International School

I, _____, have been asked to take part in a research study about vocabulary levels. The study was explained to me by Graham Creighton, an MA student at the University of South Africa and a teacher at Fukuoka International School.

I understand that I will complete a vocabulary test that will take me about fifteen minutes to complete. I will face no known risks by taking part in the study. My participation is voluntary, and I can change my mind at any time without any penalty.

About twenty students will take part in this study.

The person conducting the study will not reveal my name to anyone, and my name will not appear in any reports on this study.

I was also informed that if I have questions about the research and my rights, I can contact Prof Lilli Pretorius (pretoej@unisa.ac.za) or Dr Nanda Klapwijk (klapwn1@unisa.ac.za).

I am willing to take part in this study: _____ (Student signature)

I am willing to have my vocabulary level tested: _____ (Student signature)

Date: _____

Parental Letter of Consent for Minors

Dear Parents

My name is Graham Creighton and I am a registered MA student at the University of South Africa. In order to conduct my research I am requesting your consent for your child's participation in the study.

I am conducting research into the nature of teachers' classroom discourse as well as the vocabulary levels of learners. In order to study what teachers are saying in the classroom I will be recording their voices. If your child says something during a recording, his or her words will not be used because learners' discourse is not a part of this study. Instead, your child's participation in this study involves his or her vocabulary level tested a maximum of two times during the study. Each test will take approximately fifteen minutes and will take place at a convenient time at Fukuoka International School. [REDACTED]

Your child's participation is voluntary and he or she may withdraw at any time during the research with no penalty. This research is non-invasive and will not put your child at risk. Results of the tests will in no way affect your child's school grades. The research study may be published but your child's name will not be used.

This study will have potential benefits to your child. By identifying his or her vocabulary level, teachers may be better able to provide material that could aide your child's vocabulary growth.

If you have any questions regarding your child's participation in this study please call Fukuoka International School at [REDACTED] or contact me by email at: gcreighton@fukuoka-international-school.com

Sincerely,

Graham Creighton

By signing below, I give consent for my child to participate in the above-referenced study.

Parent's Name: _____ Child's Name: _____

Parent's Signature: _____ Date: _____

APPENDIX E

Accuracy of Transcription

I, Julia Metzger, voluntarily agreed to check the accuracy of Graham Creighton's transcriptions related to his Master's dissertation in the field of Applied Linguistics. I am a university graduate and teaching colleague of Graham Creighton. In checking for accuracy, I listened to randomly selected samples of the recordings he made as part of his study, while reading along with the transcriptions he made of the aforementioned recordings. I found the recordings to which I listened to be accurately transcribed. If you have any questions I can be contacted at +2673952244 or by email at: julia.metzger@maruapula.org.

Signed: Julia W Metzger

Full name: Julia Winslow Metzger